Evaluer et valoriser les apports effectifs des tweets géolocalisés émis en réponse aux catastrophes naturelles. Application aux phénomènes hydrométéorologiques extrêmes du Texas

Camille Cavalière

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé, environnement (MBS)**

Arrêté ministériel : 25 mai 2016

Présentée par

Camille CAVALIERE

Thèse dirigée par **Paule-Annick DAVOINE, Professeur, Université Grenoble Alpes**,
codirigée par **Céline LUTOFF, Maître de conférences HDR, Université Grenoble Alpes**, et
coencadrée par **Etienne DUBLE, Ingénieur de recherche, CNRS**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Evaluer et valoriser les apports effectifs des tweets géolocalisés émis en réponse aux catastrophes naturelles.

Application aux phénomènes hydrométéorologiques extrêmes du Texas.

Thèse soutenue publiquement le **10 février 2020**,
devant le jury composé de :

Mme. Sandrine ANQUETIN
Directrice de recherche CNRS, UMR 5001 IGE, Examinatrice
M. Johnny DOUVINET
Maître de conférences HDR, Université d'Avignon, Rapporteur
M. Thierry JOLIVEAU
Professeur, Université Jean Monnet de Saint-Etienne, Examinateur, Président du jury
Mme. Lena SANDERS
Directrice de recherche CNRS, UMR 8504 Géographie-Cités, Rapporteure
M. Etienne DUBLE
Ingénieur de recherche CNRS, Laboratoire d'Informatique de Grenoble, Coencadrant, Invité

« *Quelle que soit la manière dont je mène ma vie… j'ai toujours été un voyageur en quête d'une terre d'accomplissement de soi. Et je le resterai tant que je verrai la lumière.* »

Gustav Meyrink, *L'Ange à la fenêtre d'Occident*.

Remerciements

J'adresse mes premiers remerciements à mon co-encadrant, Etienne Dublé, non seulement pour son travail préalable qui a permis à cette recherche doctorale d'exister, mais encore pour ses précieux conseils qui ont facilité mes premiers pas dans l'informatique constituée de lettres blanches sur fond d'écran noir.

J'exprime ma plus profonde gratitude envers tous les membres de mon jury de soutenance : Mme. Lena Sanders et M. Johnny Douvinet comme rapporteurs, ainsi que Mme. Sandrine Anquetin et M. Thierry Joliveau comme examinateurs ; je suis très honorée que vous ayez accepté, en cette période agitée, de consacrer de votre temps à l'évaluation et à la discussion autour de mes travaux.

Merci à ma directrice de thèse, Paule-Annick Davoine, de m'avoir accordé sa confiance et dont l'implication à mes côtés m'a permis de décrocher cette bourse de recherche doctorale m'ayant finalement donné la chance de vouer mon énergie à la discipline géographique qui capture toute mon affection : la cartographie. Merci pour ses conseils et pour son soutien dans les moments de doute. Merci encore à Céline Lutoff pour son engagement à co-diriger cette thèse et pour ses suggestions dans l'orientation thématique de la recherche. Je remercie le Laboratoire d'Informatique de Grenoble ainsi que l'équipe STeamer et sa responsable, Marlène Villanova-Oliver, de m'avoir accueillie, en tant que géographe, pendant quatre ans, ainsi que les enseignants chercheurs avec lesquels j'ai pu collaborer pendant mon monitorat et mon année d'ATER à l'Institut d'Urbanisme et de Géographie Alpine de Grenoble : Charles Ambrosino, Elise Beck, Nathalie Dubus, Sébastien Leroux, Pascal Mao, Sandra Rome et Kamila Tabaka.

J'adresse mes *remerciements spéciaux* à Christiane Plumere et à Vincent Blanc pour les dépannages récurrents de mon ordinateur, à Nadine Mandran pour la clarté de ses cours de méthodologie de la recherche, à Le Van Tuan dont les travaux réalisés en stage ont assuré une partie de l'automatisation de tâches proposées dans cette recherche, ainsi qu'à Anne-Laure Bernardin pour son ultime soutien à l'impression de la première version de mon manuscrit !

Et enfin, un grand merci à celles et ceux qui m'ont apporté sympathie, soutien matériel ou réconfort moral jusqu'à la fin de cette aventure : mon père pour son renfort culinaire, aide inestimable pendant les derniers mois de rédaction, ma mère pour sa relecture minutieuse du manuscrit, la petite Hannah pour son sourire, ses yeux pétillants et ses éclats de joie, Marc, Ginou et Gabrielle, les chats de la maison, les occupants du bureau 371, Cécile, Fifamè, les innombrables inconnus du Web qui alimentent blogs et forums sur R et PostgreSQL, la boulangerie pâtisserie Bourbon, Muriel et Yoann, ainsi que tous les compositeurs et artistes écoutés pendant les longues heures de relecture et de correction du manuscrit.

Avant-propos

Il y a tout juste douze ans, j'avais ouvert ma propre chaîne YouTube, je diffusais mes propres vidéos de musique, j'avais une cinquantaine d'*amis* sur quatre continents. Même si je ne correspondais régulièrement qu'avec une dizaine d'entre eux, nous partagions la même musique, nous jouions la même musique. A la demande de mes *amis* américains, j'avais même créé un compte MySpace (même si je n'en voyais pas l'utilité, étant déjà inscrite et active sur YouTube) ; ce devait être alors la *mode* du moment, chez les adolescents d'outre-Atlantique.

Puis, la face du Web a changé : retirée de YouTube pendant quelques années, je m'aperçois un peu tardivement que la plateforme a été rachetée par le n°1 du Web. Il est trop tard pour tenter de récupérer ma chaîne, le nouveau concept me plaît moins, alors tant pis ; je ne serai plus qu'un *usager passif* de YouTube, comme il s'en connecte des millions chaque jour. Quant à MySpace, il n'est plus qu'un fantôme du Web. Cela étant, j'ai tout de même créé un compte Twitter mais uniquement pour les besoins de cette recherche, et donc à l'activité limitée. Mais avec le temps, j'ai finalement trouvé un intérêt à l'utilisation de mon smartphone dans mes pratiques spatiales. Randonneuse depuis de nombreuses années, amatrice de botanique et attentive aux paysages, la carte mémoire de mon téléphone intelligent s'avère désormais chargée de photographies de plantes, animaux, insectes, ou divers éléments du paysage montagnard. J'ai même fini par *accepter* d'activer la géolocalisation car dans tous les cas, à chaque voyage en TER, mon opérateur m'envoie un SMS de "*Bienvenue en Suisse !*", alors que le train n'arrive qu'en gare d'Annemasse. J'accumule ainsi les *clichés géolocalisés* (et donc cartographiables) témoins de la présence d'une orchidée rare, de la réapparition du lys martagon, des "reposoirs" à bouquetins, mais aussi des phénomènes qui affectent le paysage montagnard (laves torrentielles, glissements de terrain, avalanches, etc.). Seul petit bémol : ces clichés sont *en partie* mis en ligne via Flickr, mais sur un *compte privé*…

Au-delà de son aspect anecdotique, cette histoire personnelle révèle deux enjeux majeurs liés aux usages du numérique et de son évolution future : les pratiques numériques d'un individu connecté sont étroitement liées à son comportement et à son intérêt (ceux-ci pouvant parfois s'avérer collectifs). Combien d'individus choisissent de s'inscrire sur les réseaux sociaux populaires, combien préfèrent leur tourner le dos et quelles raisons motivent ces choix ? Parmi ces inscrits, combien participent activement à la production de contenus publics ? Par ailleurs, tout propriétaire d'un smartphone concède sur le fait d'être localisable à tout moment, alors pourquoi ne pas activer son GPS ? Le second enjeu mis en exergue correspond à la durée de vie d'un réseau social : nous ne sommes qu'à l'aube de la société numérique. Qui nous dit que d'ici à 10 ans, Twitter, dont le nombre d'utilisateurs dans le monde stagne ces dernières années, n'aura pas connu le sort de MySpace ? Ainsi, les conclusions de cette recherche restent éphémères, face à des pratiques numériques individuelles ou collectives qui évoluent chaque année.

Evaluer et valoriser les apports effectifs des tweets géolocalisés émis en réponse aux catastrophes naturelles

Application aux phénomènes hydrométéorologiques extrêmes du Texas

Les traces numériques ont envahi notre quotidien : capturées en temps réel par divers outils connectés fixes ou nomades, elles présentent l'avantage d'être fréquemment géolocalisées. Ces traces constituent ainsi des marqueurs virtuels attestant de la présence physique d'un individu dans un espace précis à un moment connu. Depuis une dizaine d'années, elles font ainsi l'objet de publications régulières dans de nombreuses disciplines, dont la géographie, et sont considérées comme le pilier de la construction de nouvelles connaissances des phénomènes sociaux, selon une approche verticale ascendante (*bottom-up).* La recherche doctorale se focalise sur une trace numérique particulière, le tweet géolocalisé : en raison des nombreux phénomènes violents survenus au début des années 2010, la gestion des risques et catastrophes naturels a fait partie des premiers thèmes d'exploration du potentiel des tweets géolocalisés comme nouvelle source d'information de terrain.

Pour autant, l'étude géographique des traces numériques géolocalisées rencontre des difficultés qui restent marginales dans l'approche des *data analysts* : devant la variabilité d'accès et d'utilisation des nouvelles technologies mobiles, quelle est la représentativité sociale et spatiale de ces traces ? Comment positionner les outils de l'analyse spatiale et de la cartographie face à ces nouveaux types de données hétérogènes et acquises en dehors de toute norme conventionnelle ? Peut-on valoriser les traces numériques géolocalisées en information géographique ? Dans cette recherche, nous explorons ces questions à partir des phénomènes extrêmes d'origine hydrométéorologique, survenus au Texas au printemps 2016 et en août 2017.

En premier lieu, nous formalisons des méthodes d'extraction sémantique et spatiale destinées à améliorer l'étape de constitution d'un corpus de tweets relatifs aux phénomènes étudiés (nommés tweets de crise). L'analyse des tweets de crise est ensuite fondée sur deux axes : d'une part, l'exploration des comportements spatiaux et statistiques des lieux de l'activité virtuelle ; d'autre part, la question de la pertinence, ainsi que de la valorisation, du tweet de crise géolocalisé comme marqueur des différents paramètres du phénomène (localisation à une échelle fine, intensité, *etc*.) et de la vulnérabilité des territoires, en croisant les tweets avec des données externes officielles de réalité-terrain.

Diagnosis and valorization of the actual benefits of geolocated tweets issued in response to naturel disasters

Application to extreme hydrometeorological phenomena in Texas

Digital footprints have overwhelmed our daily lives: they are captured in real-time by mobile web-linked devices and they are often geolocated. Such footprints are considered as virtual markers witnessing the physical presence of any connected individual, in a given space, at a given time. For a decade, they have been integrated as a new research field in many disciplines, including geography. Furthermore, they are often considered as an opportunity to build a new individual-based knowledge about social phenomena, based on a bottom-up approach. This research focuses on geolocated tweets: because of the many violent phenomena that occurred in the early 2010s, natural disasters management quickly became a major field of research to evaluate the potential of these particular footprints as a new field and real-time information.

However, geographical studies based on geolocated digital footprints are now facing some underlying difficulties that data analysts do not encounter: as the use of mobile technologies is heterogeneous considering populations and places, what is the social and spatial representativeness of such footprints ? Are the traditional tools of cartography and spatial analysis adapted to those unstructured data that elude any professional standard? Can we transform geolocated tweets into geographical information? In this research, we explore these questions from the extreme hydrometeorological phenomena that occurred in Texas in spring 2016 and in August 2017.

First, we formalize tweets semantic and spatial retrieving methods to improve the step of building a crisis tweets dataset. The analyse of this crisis tweets dataset is based on two approaches: on one hand, we study statistic and spatial behaviors of virtual activity hotspots. On the other hand, we explore the questions of crisis tweets relevance and cartographic valorization: we evaluate crisis tweets as an indicator of the field and phenomenon properties (local-scale intensity, vulnerability of populations, level of damages), by cross-checking tweets with official ground-truth data.

Table des matières

Table des figures

Table des tableaux

Liste des acronymes

ACS *American Community Survey*

CBD *Central Business District*

CDC *Centers for Disease Control and Prevention*

FEMA *Federal Emergency Management Agency*

NOAA *National Oceanic and Atmospheric Administration*

NWS *National Weather Service*

SEDB *Storm Events Database*

SVI *Social Vulnerability Index*

TIC Technologies de l'Information et de la Communication

USGS *United States Geological Survey*

Introduction générale

Introduction générale

Numérique et réseaux sociaux sont désormais omniprésents et ubiquistes : la *révolution numérique*[1], amorcée à partir des années 1980, connaît depuis une décennie une démocratisation technologique qui entraîne dans ses rouages une série de mutations sociétales et territoriales (Lebreton, 2013). L'individu, quel que soit le lieu dans lequel il se trouve, devient alors un utilisateur-consommateur de nouvelles technologies de l'information et de la communication (les *TIC*). Ainsi, en 2016, le temps moyen quotidien passé par un adulte français devant un écran connecté au Web était estimé à 3h46, dont 1h41 de connexion par usage du smartphone[2]. Comment les pratiques virtuelles valorisent-elles ou transforment-elles les rapports des usagers à l'espace ?

Contexte général de la recherche

De nouvelles logiques de production de l'information.

Si le Web s'est imposé comme nouvel espace virtuel multi-activités englobant aussi bien vie professionnelle que loisirs, l'une de ses caractéristiques principales reste indéniablement sa capacité à bouleverser les cadres et contraintes qui régulaient le fonctionnement des sociétés occidentales avant l'introduction du numérique (Lesteven et Godillon, 2017) :

- le système traditionnel de verticalité et de hiérarchie par lequel on diffusait et produisait l'information se trouve déséquilibré par l'introduction d'une nouvelle logique horizontale (Lebreton, 2013) : la connaissance est accessible à tous et tout internaute peut devenir acteur de la production des savoirs (les exemples les plus connus de cette nouvelle logique restent l'encyclopédie collaborative Wikipédia et la cartographie libre OpenStreetMap). De même, tout utilisateur connecté et réactif a la capacité de témoigner virtuellement d'un événement localisé réel, et ce avant qu'un journaliste des médias traditionnels ne soit dépêché sur le lieu de survenue de l'événement en question ;
- l'espace virtuel s'affranchit de l'espace réel et de ses acteurs traditionnels : les nouveaux espaces de rencontre et de communication que constituent les réseaux sociaux sont a-territoriaux ; les transactions commerciales se libèrent des magasins et des horaires : elles ont lieu à la maison, dans les transports, dans les lieux publics, tous les jours, à toute heure. Pour son déjeuner, l'employé pressé n'aura qu'à passer commande, par une application

[1] Bien que l'expression reste toujours débattue, elle fait référence aux mutations engendrées dans les sociétés par l'adoption générale des outils de l'informatique et de l'Internet ; celles-ci se manifestent principalement par de nouvelles formes de communication et d'échanges qui favorisent l'émergence d'une intelligence collective ainsi qu'une redistribution horizontale des réseaux hiérarchiques traditionnels.

[2] Source : http://www.lefigaro.fr/secteur/high-tech/2016/11/22/32001-20161122ARTFIG00100-en-2017-les-francais-passeront-4-heures-par-jour-sur-leurs-smartphones-et-leurs-pc.php (Consulté pour la dernière fois le 14/01/2019)

smartphone, auprès d'une enseigne de restauration qui la fera livrer au pied du bureau, par un coursier cycliste, en un minimum de temps.

Bien que dans un premier temps, on a pensé que l'intégration des *TIC* à la société allait sédentariser les individus (Aguiléra et Belton-Chevalier, 2017), le numérique n'annonce pas pour autant la fin des interactions entre l'individu utilisateur des TIC et le territoire. Si les réseaux sociaux deviennent des espaces virtuels privilégiés en termes de communication, ils peuvent favoriser les phénomènes collectifs dans l'espace réel, de par leur pouvoir mobilisant (Klein et Huang, 2012). Ainsi, en 2011, pendant les *Printemps Arabes*, *TIC* et plateformes de diffusion de médias créés par des usagers de l'Internet mobile se sont révélées comme de puissants vecteurs en termes de mobilisation et de contestation civile[3] (Wilson et Corey, 2012). Récemment, c'est le mouvement des *Gilets Jaunes*, en France, qui a été initié sur les principaux réseaux sociaux par lesquels les appels à manifester ont été relayés[4].

Perception de l'espace et nouvelles pratiques territoriales à l'heure de la société numérique.

Les usages des TIC participent ainsi au remodelage des pratiques territoriales chez les populations : celui-ci est manifeste dans le domaine des mobilités. Les nouvelles formes de déplacements, engendrées par le numérique, s'illustrent notamment par l'essor des mobilités partagées : dès les premiers jours de la grève des cheminots au printemps 2018, la plateforme de covoiturage BlaBlaCar enregistrait un pic de demandes de réservation ainsi qu'une explosion de nouveaux inscrits et de trajets covoiturés entre domicile et travail[5]. Ainsi, c'est le smartphone, par lequel on accède aux différentes plateformes de réservation de transports partagés ou à la demande, qui est devenu l'outil central de la réorganisation des flux spatio-temporels liés aux mobilités des individus (Aguiléra et Belton-Chevalier, 2017).

Si l'on considère les activités virtuelles récréatives, un loisir numérique particulier tend à s'intégrer comme nouvelle source d'informations sur les pratiques spatiales des individus : les jeux en réalité augmentée, basés sur la localisation de l'individu et joués par l'usage du smartphone. D'après (Colley *et al*., 2017), la participation régulière à ces loisirs fondés sur les interactions entre espace virtuel et espace réel a la capacité de modifier les pratiques territoriales des individus : les joueurs réguliers peuvent ainsi adopter de nouveaux trajets et fréquenter de nouveaux espaces urbains dans lesquels ils sont enclins à consommer.

[3] Les événements des *Printemps arabes* illustrent le nouveau pouvoir des *TIC* et réseaux sociaux : dans une logique verticale, les gouvernements tunisien et égyptien ont bloqué l'accès Internet dans leur pays respectif. D'un côté, cette décision ne fait qu'aggraver la contestation dans les rues ; de l'autre côté, il y a des individus qualifiés issus de la société civile, qui peuvent contourner le blocage et reprendre leurs activités sur les réseaux sociaux. Par les *TIC*, l'individu a donc la capacité de devenir acteur et de s'affranchir de la logique verticale descendante de la décision politique.

[4] Source : https://www.bfmtv.com/tech/comment-facebook-a-contribue-a-l-eclosion-des-gilets-jaunes-1572771.html (Consulté pour la dernière fois le 14/01/2019)

[5] Source : https://blog.blablacar.fr/newsroom/news/greves-trains-6-plus-de-demandes-de-covoiturage (Consulté pour la dernière fois le 17/01/2019)

Enfin, à l'échelle de l'individu, le numérique, et plus particulièrement les réseaux sociaux permettent d'appréhender le *sensible* (qu'on appréhendera ici comme tout objet ou situation de l'environnement suggérant un sens particulier pour l'individu utilisateur des TIC). L'Internet nomade et le smartphone offrent en effet un avantage considérable : ils permettent de photographier le territoire ou un élément du territoire et de le diffuser sur un réseau dans l'immédiat. En d'autres termes, ils assurent la sauvegarde, à travers le regard de l'usager, du lieu ou de l'environnement à l'instant *t*. Ce mode de capture du territoire reste subjectif et sélectif mais nous indique quels objets ou quels événements font sens pour un individu. Le tableau 0.1 ci-après propose quelques exemples de contenus photographiques, associés à des phénomènes météorologiques ou hydrométéorologiques, capturant des objets de territoires, choisis par les usagers, et émis via la plateforme Twitter :

Tableau 0.1 : Quelques exemples de témoins de lieux et objets de territoires, photographiés par smartphone et postés sur Twitter

Photographie	Objet	Interprétation
	Zouave du Pont de l'Alma, crue de la Seine à Paris, janvier 2018	Témoin visuel relatif des habitants de la capitale pour "mesurer" la crue du fleuve. L'objet reste inscrit dans les représentations collectives comme indicateur de l'intensité d'une crue
	Panneaux d'itinéraires de randonnée - 31/12/2018 (haut) et 09/02/2019 (bas)	Point de vue identique pour des dates différentes : identifier un repère visuel de l'espace pour "mesurer" la variabilité de l'enneigement pendant la saison hivernale
	Aiguille du Goûter – 27/07/2019 et 09/08/2019	Point de vue et objet identiques pour des dates différentes : en utilisant un repère de haute montagne, vulnérable face aux aléas climatiques, visualiser la fonte des neiges et des glaces après les périodes de canicule de l'été 2019

En conséquence, la spatialité n'est pas détruite par la première impression que suggère l'individu isolé rivé sur son écran. Les trois exemples cités dans le tableau 0.1 témoignent de la capacité de l'individu connecté à devenir un instrument d'observation et de suivi d'objets

(ou de contextes environnementaux), susceptible de présenter un intérêt scientifique double : comment les utilisateurs des TIC appréhendent-ils les éléments de leur environnement et comment peuvent-ils aider des chercheurs à assurer le suivi de l'évolution de cet environnement à travers leurs productions numériques ? Pour autant, en dépit de cette perspective favorable, une question se pose alors : la production de contenus numériques nécessitant l'acquisition d'appareils particuliers ainsi qu'un accès à une connexion nomade, ces nouvelles pratiques permettent-elles d'envisager une prise en compte équitable des territoires et de leurs populations ?

Une géographie du numérique marquée par les inégalités socio-spatiales.

L'accès à un réseau numérique de qualité, en particulier au très haut débit, est en effet généralement considéré comme un enjeu fondamental d'attractivité et de compétitivité des territoires à l'échelle locale (Bakis, 2012 ; Lebreton, 2013). Les défauts actuels qui persistent dans la qualité des réseaux sont perçus comme un frein au développement économique des territoires situés en marge des infrastructures de très haut débit. En effet, les opérateurs ont tendance à privilégier, dans la modernisation des réseaux de télécommunication, les espaces où le marché est rentable, c'est-à-dire les métropoles (Bakis, 2012). L'intégration des *TIC* constitue donc un facteur d'aggravation des disparités territoriales pré-existantes : l'inégale répartition des équipements creuse les écarts entre métropoles qui concentrent les activités technologiques de pointe (Bakis et Schon, 2012) et villes d'un rang inférieur dans la hiérarchie urbaine (Bakis, 2012). Ainsi, en France, il suffit de consulter les données de modernisation des réseaux pour constater ces inégalités territoriales, perceptibles à toutes les échelles spatiales : les grandes agglomérations disposent de services performants alors que les espaces ruraux ou montagnards n'ont qu'un accès minimal sur des réseaux anciens voire obsolètes : par exemple, en 2017, si 92,5% des foyers de Grenoble sont éligibles au raccordement par câble, ils ne sont plus que 75% à Saint-Martin-d'Hères et ce pourcentage est quasi nul dès qu'on sort de l'agglomération (Voiron, Voreppe). Le constat est identique dans les territoires de montagne : 98,2% des foyers annéciens sont éligibles à ce raccordement alors qu'à Chamonix, il n'y en a aucun, malgré le statut international de la ville[6].

On peut alors appréhender le numérique comme une forme de sélection et de hiérarchisation des territoires et des populations : les inégalités territoriales et sociales d'accès aux outils numériques ont ainsi donné naissance, dès le début des années 2000, au concept de la fracture du numérique comme le *"fossé séparant ceux qui bénéficient de l'accès à l'information numérique et ceux qui demeurent privés des contenus et des services que ces technologies peuvent rendre"* (Morin-Desailly, 2018). Cette fracture se révèle comme un phénomène multidimensionnel : elle est d'abord matérielle car elle traduit les déficits territoriaux en matière de disponibilité et d'accessibilité à des équipements de qualité. Elle est également intellectuelle et sociale car elle révèle les disparités en termes de maîtrise des

[6] Source des données : ARCEP sur https://www.data.gouv.fr/fr/datasets/niveau-des-debits-sur-les-reseaux-dacces-a-internet-adsl-cable-fibre-ftth-t2-2015-t2-2017/ (Consulté pour la dernière fois le 26/11/2019)

compétences fondamentales en fonction des niveaux de vie et d'études : si les réseaux sociaux semblent être utilisés par toutes les strates sociales de la population, les activités de recherche d'informations et l'utilisation des plateformes de services administratifs, bancaires ou commerciaux restent l'apanage des plus diplômés (Morin-Desailly, 2018). Par ailleurs, pour reprendre les exemples cités dans la section précédente, le jeu en réalité augmentée est un phénomène essentiellement urbain. Il en est de même pour le recours aux plateformes numériques collaboratives destinées à organiser la mobilité à l'échelle individuelle : dans une enquête menée sur les usages de véhicules électriques partagés dont la réservation et l'emprunt s'appuient exclusivement sur des plateformes accessibles depuis un smartphone, (Eskenazi *et al*., 2017) ont montré que les usagers témoignaient d'une maîtrise avancée des *TIC* propres aux smartphones.

On l'aura compris : le numérique n'est pas uniforme mais il intègre une multiplicité de supports physiques ou nomades associés à une diversité d'usages et de besoins. En outre, par sa capacité à fédérer les individus autour d'intérêts communs, à capitaliser et à diffuser rapidement de l'information sur le Web, le numérique s'insinue et rebat les cartes de processus hérités, selon une nouvelle logique horizontale. En effet, le cœur du fonctionnement des plateformes numériques de l'Internet repose sur l'usager : celui-ci n'est pas un individu se contentant de rechercher et de recevoir de l'information dans une posture passive mais devient acteur de la construction du contenu des plateformes. L'une des situations où cette logique hiérarchisée a également été bouleversée pendant ces dernières années, est celle de la gestion des risques et catastrophes naturels (Hecker, 2014).

Quand le numérique s'invite dans la gestion des risques et catastrophes naturels.

Dans le champ de la gestion des risques naturels, l'adoption du numérique ne s'est pas réduite à la modernisation des méthodes assurant la construction de cartes d'aléas, d'enjeux ou encore de vulnérabilités des populations, soit l'ensemble des paramètres traditionnellement analysés afin de construire une connaissance des risques sur un territoire (Davoine, 2014). L'intégration du numérique s'illustre désormais par de nouvelles pratiques qui s'exercent dans les temporalités réelles de la crise.

La diffusion d'une alerte mêle ainsi médias conventionnels (sirènes, radio, journaux, télévision) et médias numériques. Ainsi, on retrouve l'ensemble des acteurs institutionnels sur le Web voire même sur les réseaux sociaux : Météo France actualise ses cartes de vigilance deux fois par jour sur son site et les diffuse via son compte Twitter ; les radios régionales et organismes d'Etat disposent également de comptes sur les réseaux sociaux via lesquels ils publient de l'information auprès des usagers du numérique pendant la phase d'alerte. Ce phénomène est particulièrement palpable aux Etats-Unis : mis en cause dans la gestion de l'ouragan Katrina en 2005, institutions fédérales et pouvoirs publics locaux ont développé une véritable stratégie de communication fondée sur la diffusion d'informations via des

plateformes de réseaux sociaux de large audience, depuis leurs comptes officiels (Hecker, 2014).

Aussi, lorsqu'on évoque le numérique dans la gestion des risques et catastrophes naturels, a-t-on coutume de se focaliser sur l'accès à l'information, perçu comme nouveau pilier de la résilience[7] (Hecker, 2014). Cet accès à l'information s'inscrit également dans l'organisation des réponses à la survenue d'une crise. Dans cette logique, des outils et interfaces cartographiques destinés au public se déploient à profusion et intègrent de nouveaux acteurs. Ces applications ont un double objectif : elles servent d'abord à collecter des informations relatives aux conditions réelles du terrain et, en retour, elles diffusent des consignes de prudence et permettent de localiser les lieux de ravitaillement, de soin, *etc.*. Depuis le début des années 2010, après la survenue de séries de catastrophes naturelles (dont Fukushima en 2011), les services publics de différents Etats ont intégré la veille des communications émises via les réseaux sociaux numériques à leur politique de gestion de crise locale (McDougall, 2012 ; Hecker, 2014). C'est notamment le cas aux Etats-Unis et en Indonésie. Ces services des Etats peuvent être épaulés, voire même substitués dans les pays les plus pauvres, par des associations internationales comme le *Visov* ou les *Digital Humanitarians*[8] qui se chargent des activités de veille, de collecte, de cartographie et de diffusion d'informations via les réseaux sociaux numériques (Hecker, 2014 ; Meier, 2015 ; Douvinet *et al*., 2017).

Du fait de l'ubiquité du numérique, on assiste à un renouvellement des matériaux traditionnels sur lesquels s'appuyait la gestion des catastrophes naturelles. D'une part, ce renouvellement correspond à l'intégration de nouvelles formes de données géolocalisées produites par les usagers de plateformes numériques, à la connaissance de terrain et à la cartographie de l'espace en crise. D'autre part, l'utilisation de ces matériaux semble s'imposer comme nouveau paradigme chez les divers acteurs de la gestion de crise, dans la mesure où ces formes de données géolocalisées et acquises en temps réel peuvent produire des informations qu'on ne pourrait capturer aussi facilement par les outils et méthodes traditionnels : *"The whole point of using additionnal data sources from social media is to complement existing data sources that may not be available as quickly or in real-time"* (Meier, 2015).

On aurait ainsi tendance à construire une nouvelle forme de gestion de crise, centrée sur le numérique mais sans certitude réelle sur un enjeu primordial : qui, quoi, quand, et quels territoires le numérique représente-t-il ? Et c'est ce paradoxe que soulignait le chercheur américain en géomatique Anthony Robinson, après le passage de l'ouragan Harvey sur les côtes du Texas en août 2017 : *"The tragedy in Texas will magnify the fact that the most*

[7] La notion de résilience, mobilisée dans de nombreux domaines, est entendue ici comme l'ensemble des capacités des populations - communautés affectées à retourner dans les plus brefs délais à un état d'équilibre lorsque survient un phénomène naturel dommageable (Hecker, 2014 ?)

[8] http://www.visov.org/ et http://www.digital-humanitarians.com/

vulnerable ppl[9] *are not going to tweet about it. Response can't be driven by tweets"* (message émis sur la plateforme Twitter le 27/08/2017). Et c'est bien là l'enjeu de ce travail de thèse : peut-on construire une connaissance objective des populations et des territoires en situation de crise dans les conditions actuelles d'utilisation des plateformes numériques ?

Problématique de la recherche

Positionnement de la recherche.

L'enjeu de la présente recherche ne s'inscrit pas dans l'optique d'un usage du numérique orienté sur la gestion de crise en temps réel, tel que décrit ci-dessus. Par conséquent, notre objectif ne consiste pas à identifier les lignes directrices pour orienter la conception d'une application destinée à améliorer la résilience d'une population face à un risque, ni d'évaluer le rôle des applications existantes dans les réponses à la crise, dans une perspective de retour d'expérience.

Notre démarche se positionne en parallèle de cette logique ; plus précisément, elle s'inscrit dans la réflexion géographique qui devrait accompagner l'usage des données numériques produites par les usagers des nouvelles plateformes du Web, en ce qui concerne les questions impliquant une variabilité spatiale, temporelle et sociale, comme dans le cas de la gestion des risques naturels. D'un côté, les paragraphes précédents ont montré que les usages et la distribution spatiale actuelle de la couverture numérique tendent à façonner des espaces et des sociétés à plusieurs vitesses (et nous avons d'ores-et-déjà identifié un certain nombre de facteurs spatiaux discriminant individus, territoires et objets du territoire). D'un autre côté, le volume de l'information géolocalisée créée quotidiennement par les usagers des plateformes du Web représente une base potentielle de connaissances inédites et incommensurables, dont nous avons également cité quelques exemples. Dans cette perspective, il serait logique de considérer le numérique comme une nouvelle opportunité de constituer une base de connaissances pour comprendre les effets des crises d'origine naturelle au sein des territoires : en effet, l'enjeu consiste à identifier les apports réels, ainsi que les limites, des données de terrain produites par les individus connectés en temps réel, dans un contexte perturbé où la donnée expertisée fait souvent défaut.

Question de recherche.

Maintenant que ces données font l'objet d'études dans le champ des risques naturels et dans la géographie depuis une dizaine d'années, nous nous attachons à caractériser la géographie de l'usage d'un réseau de communication virtuelle mais localisable dans un territoire réel traversant une crise d'origine naturelle. Cette géographie se décline selon trois volets principaux :

[9] Abréviation du mot *people*

- l'identification des espaces de réactivité et des populations participatives (individus diffusant de l'information géolocalisée sur les réseaux de communication numériques) pendant les différentes phases d'une crise d'origine naturelle ;

- l'identification des dynamiques spatio-temporelles de l'information géolocalisée créée et diffusée sur un réseau en réaction à un phénomène naturel qui se déroule sur le territoire réel;

- l'identification et la mesure des biais sociaux ou/et spatiaux susceptibles d'influencer les phénomènes mis en évidence sur le réseau.

Questionnement de recherche

Dans la première partie de cette introduction, nous avons souligné les difficultés à appréhender les rapports entre l'espace virtuel qu'est le réseau numérique et l'espace réel : le numérique est-il créateur ou destructeur de pratiques spatiales ? Le numérique peut-il restituer intégralement pratiques et perceptions spatiales de ses utilisateurs ? Le numérique permet-il d'appréhender les dynamiques d'émergence de la crise au sein d'un territoire de manière objective ?

Le questionnement central de la recherche s'inscrit dans cette logique : incontestablement, les territoires de demain sont en train d'être façonnés, en partie, par les réseaux numériques. Le problème actuel reste que ces réseaux virtuels sont relativement récents et qu'on ne connaît pas encore leur pertinence pour la compréhension des processus à l'œuvre en cas de survenue d'un phénomène naturel générateur de dangers.

Quels sont alors les apports géographiques effectifs des données géolocalisées issues d'une plateforme numérique, pour la connaissance des dynamiques d'un territoire en crise ? Peut-on s'appuyer sans risques sur ces nouvelles sources d'information pour construire une connaissance objective des liens entre territoires et populations en crise, dans un contexte particulier où la donnée de terrain produite en temps réel est rare ?

Hypothèses et objectifs de la recherche

Hypothèses de recherche.

L'émission de tweets géolocalisés, tout comme le jeu en réalité augmentée, constitue un exemple d'articulation entre, d'une part, le territoire réel et vécu et, d'autre part, l'espace virtuel du réseau. Les manifestations physiques d'un phénomène d'origine naturelle s'inscrivent dans l'espace et dans le temps : une perturbation atmosphérique se déplace et n'atteint pas simultanément tout point d'un territoire ; les pluies consécutives à cette

perturbation n'ont pas la même intensité dans l'espace et dans le temps ; de la même manière, l'inondation s'inscrit dans une dynamique spatio-temporelle et son intensité varie en fonction d'une dynamique amont-aval d'un bassin versant. Or, les travaux effectués dans la thématique des données numériques géolocalisées produites par des usagers ont d'ores-et-déjà montré la disposition des usagers à réagir dans l'instant, sur les réseaux sociaux, en cas de survenue subite d'une perturbation de leur environnement local, quelle que soit sa nature (Lin *et al.*, 2013 ; Kounadi *et al.*, 2015 ; Lucchini *et al.*, 2016). Cette capacité à réagir s'illustre dans le champ des risques : les manifestations d'un phénomène physique (pluies intenses, crues, inondations, vents violents, etc.) engendrent des perturbations environnementales et sociales sur un territoire donné. Afin de distinguer ces phénomènes physiques de leurs conséquences réelles et virtuelles, nous qualifions les perturbations qu'ils engendrent par l'expression événements du monde réel (fermetures préventives de lieux ou d'infrastructures, maisons inondées, routes coupées, ouverture de refuges, *etc.*). L'usager équipé de son smartphone a la capacité de retranscrire, sur le réseau virtuel, non seulement la survenue du phénomène physique mais également des événements réels sur le réseau virtuel (figure 0.1).

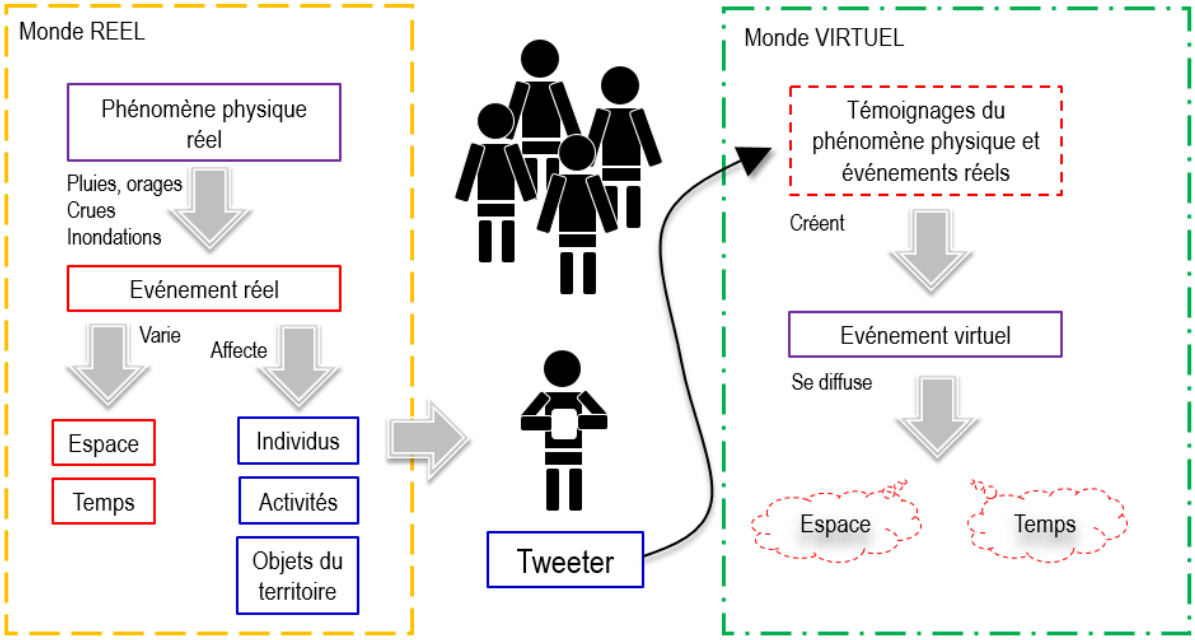


Figure 0.1 : Articulation entre phénomènes et événements du monde réel et transcription dans un événement virtuel

On peut alors espérer identifier, grâce à la géolocalisation et aux métadonnées temporelles ajoutées aux tweets, la diffusion et l'évolution d'un *événement virtuel* dans le temps et dans l'espace[10], qui se construit en parallèle des phénomènes et événements réels, d'où la formulation de l'hypothèse n°1 :

[10] A la condition essentielle qu'on trouve, sur le territoire affecté, des utilisateurs du numérique assistant au phénomène physique réel ou acteurs des événements réels.

Hypothèse n°1 - L'articulation de l'espace réel et du réseau virtuel

Le tweet géolocalisé peut être considéré comme un marqueur spatio-temporel qui détecte et enregistre toute anomalie survenant sur un territoire. L'information étant capturée par smartphone, outil suivant l'usager dans ses activités et déplacements du quotidien, on peut la considérer comme locale et ciblée sur un élément du territoire parcouru ou fréquenté. En conséquence, on peut considérer que la dynamique spatio-temporelle de l'événement virtuel créé sur le réseau reflète la dynamique spatio-temporelle des phénomènes et événements du monde réel. Les tweets géolocalisés peuvent ainsi être mobilisés dans une perspective de base de connaissances de terrain.

Dans un deuxième temps, il convient de poser un cadre épistémologique spécifique guidant l'exploration de ces données. Informatique, sciences de la communication et sciences sociales se sont rapidement emparés des données numériques géolocalisées produites par les usagers et ont annoncé un renouveau certain des théories encadrant les sciences humaines et sociales (Anderson, 2008 ; Kitchin, 2013), sous l'impulsion des caractéristiques inédites de ces données : capturées massivement, en temps réel et de disponibilité immédiate. Cette assertion mérite d'être discutée : certes, l'observation des impacts d'un événement et des sociétés se trouve désormais facilitée, étant donné l'usage quasi quotidien des plateformes numériques. En revanche, le chercheur ne peut pas prévoir la richesse d'un jeu de données numériques collecté, ou au contraire, sa pauvreté : étant donné l'absence de contrainte exercée sur les conditions de production des données numériques géolocalisées, tout usager peut évoquer un sujet dans n'importe quels termes, et selon ses propres perceptions.

En 2008, Chris Anderson publiait, dans la revue Wired, un article intitulé "*The end of theory: the data deluge makes the scientific method obsolete*" dans lequel il soulignait alors l'inadéquation entre d'une part, méthode scientifique traditionnelle ancrée dans la démarche hypothético-déductive et la recherche de modèles prédictifs, et d'autre part, nouvelles données numériques capturées en masse. La réponse proposée visait ainsi à ancrer la recherche dans l'observation directe des données et l'hypothèse selon laquelle elles ont la capacité à parler d'elles-mêmes "*Let the data speak from themselves*". Quelle approche peut-on alors utiliser pour *faire parler les données*, et mettre en évidence les questions auxquelles elles ont la capacité de fournir des réponses, ou le cas contraire (Anderson, 2008) ?

Hypothèse n°2 - L'adoption d'une posture spécifique pour déployer le plein potentiel des données

Formuler des hypothèses *a priori* et construire un modèle de description et de prédiction d'un comportement ou d'un phénomène n'est plus une démarche appropriée face aux données numériques géolocalisées produites par des usagers connectés. En effet, ce type de données n'est pas scientifique, dans la mesure où leur production n'est pas régulée par des protocoles d'acquisition rigoureux et contrôlés par les chercheurs. Construire une connaissance des données numériques ancrée dans le paradigme scientifique hypothético-déductif, c'est prendre le risque de masquer ou de biaiser le potentiel de ces données, dans la mesure où l'on risque de limiter les analyses aux questions préalables qu'on se pose.

En réponse à la nature nouvelle de ces données, il faut alors reconsidérer le positionnement scientifique par lequel on construit la connaissance : dans l'hypothèse précédente, nous supposons que les tweets enregistrent la dynamique spatio-temporelle du phénomène naturel et des événements réels. Mais d'autres facteurs, des objets du territoire ou des lieux particuliers peuvent influencer les émissions de tweets dans l'espace et dans le temps, s'ils ont un sens particulier pour les populations locales affectées. Or, ce paramètre s'avère difficile à prévoir car il relève d'une expérience propre à chaque individu, à laquelle le chercheur ne peut pas se substituer. Questionnement et pistes d'analyse doivent alors être orientés progressivement en fonction des observations : la théorie est déduite par la juxtaposition de ces observations, dans une logique inductive. En conséquence, d'après (Anderson, 2008), il faut changer d'approche pour démontrer une connaissance tout en cherchant à identifier ce qu'on peut démontrer ; le problème reste qu'on ne sait pas encore si ce paradigme est efficace dans ce contexte de réactivité aux crises d'origine naturelle.

Enfin, dans la présente recherche, inscrite en géographie et non dans une approche de *data analyst*, nous nous démarquons, dans une certaine mesure, de ce *credo* du "*Let the data speak from themselves*" : dans un premier temps, l'argumentation de (Anderson, 2008) s'appuie en effet sur une approche quantitative des données massives traitées par des algorithmes[11]. Dans une approche géographique, on ne sait pas si les quantités de données sont systématiquement significatives d'un événement particulier (Quesnot, 2016). Dans un second temps, la distribution spatiale des données numériques capturées par les plateformes de l'Internet mobile à l'échelle de l'individu est un phénomène complexe à saisir : dans l'absolu, la répartition des données numériques géolocalisées suit la distribution spatiale des densités de population. Cet effet des foyers de peuplement, correspondant à des territoires

[11] Le vocabulaire employé illustre bien cette approche : "*Petabytes of data*" , "*massive amounts of data*".

urbains ou inclus dans les couronnes des grandes métropoles, qui bénéficient de réseaux d'infrastructures numériques performants, doit être atténué par le recours à des jeux de données décrivant les caractéristiques démographiques de la population. Pour autant, toute donnée démographique publiée par un institut national est captée à l'échelle fixe du domicile alors que le propre de la donnée numérique est d'être capturée dans une situation de mobilité quotidienne individuelle ou collective : ce phénomène n'est pas sans conséquences sur l'étude des dynamiques spatio-temporelles capturées par les données numériques à l'échelle locale (rien ne garantit que la carte des densités de population, construite à partir de données de recensement, soit le miroir de la carte des densités de tweets géolocalisés).

Par ailleurs, les quantités des données numériques émises en un lieu précis dépendent des comportements des utilisateurs des plateformes numériques : si la participation d'une majorité d'usagers des plateformes numériques à la production de données reste ponctuelle, certains individus sont considérablement actifs et peuvent rapidement conduire à la détection d'un phénomène non significatif. La concentration de données numériques localisées en un lieu précis du territoire peut ainsi résulter de la super-activité d'un unique individu (Cavalière *et al*., 2015). De même, une grande quantité de données numériques produites par des usagers de plateformes et localisées sur un objet précis du territoire n'est pas un indicateur fiable pour interpréter le sens d'un lieu ; rien ne garantit que le lieu qui rassemble des adeptes du numérique ait une popularité identique dans l'ensemble de la population (Quesnot, 2016). L'hypothèse n°3 est alors formulée en réponse à ces considérations :

Hypothèse n°3 - La non-autosuffisance des données numériques

Les tweets géolocalisés, en tant que données produites par des usagers de plateformes numériques, ne sont pas autosuffisants : ils ont en effet la caractéristique d'être a-contextuels, c'est-à-dire que le chercheur, qui ne fait que collecter les tweets *a posteriori*, n'a aucune idée sur le contexte et les motivations qui ont incité un utilisateur à créer le message sur la plateforme en question.

Pour révéler le potentiel de ces données numériques, il faut tenter de reconstituer ce contexte environnemental et social d'émission en croisant ces données produites par les usagers avec des jeux de données externes, d'autres natures.

Objectifs de la recherche.

Les travaux entrepris s'articulent alors autour de l'objectif suivant : explorer et qualifier le potentiel du tweet géolocalisé, en identifiant les apports et limites de l'utilisation de cette nouvelle forme de donnée numérique localisable face à une problématique géographique, qu'on peut décliner sous deux angles :

- les concepts théorisant les risques naturels, comme la vulnérabilité, impliquent une certaine variabilité spatiale, temporelle et sociale, à l'échelle de l'individu comme à l'échelle du collectif.

- à cette variabilité des enjeux s'ajoutent les disparités d'accès aux outils et d'usage des plateformes numériques, ainsi que la potentielle création de nouvelles pratiques territoriales spécifiques au territoire en crise. En effet, quelle que soit l'échelle géographique considérée, tous les territoires ne bénéficient probablement pas de la même visibilité sur les réseaux numériques comme Twitter. Comment considérer et réagir face à un territoire qui subirait une crise grave et dont les populations ne seraient guère consommatrices des réseaux sociaux ? Et *a contrario*, que penser d'un espace où se déclencherait une crise moins intense mais qui concentrerait davantage d'usagers s'empressant de photographier leur environnement perturbé et d'émettre un message sur le réseau ?

C'est là l'enjeu de cette recherche qui se focalise alors sur trois points méthodologiques et thématiques :

- tout d'abord, il s'agit de mettre en place une chaîne de traitements destinés à faciliter l'extraction automatisée des tweets géolocalisés en lien avec la problématique spatiale des risques, cette étape représentant généralement une tâche longue et fastidieuse.

- Dans un deuxième temps, la recherche s'attache à l'exploration et au test de méthodes combinant des approches statistiques et cartographiques pour l'exploration des tweets géolocalisés, combinés à d'autres jeux de données. L'enjeu consistera alors à diagnostiquer les apports effectifs de ces données particulières, par rapport au potentiel annoncé dans la littérature, c'est-à-dire la connaissance des territoires en crise et de leurs dynamiques, qu'on peut construire à partir d'un réseau virtuel. Il conviendra en outre d'expliciter la significativité, à différentes échelles géographiques, des composantes de ces données et des facteurs influençant l'hétérogénéité de leur répartition spatiale. En d'autres termes, il s'agit de tester la solidité du matériau tweet géolocalisé dans le cadre d'une utilisation géographique.

- Enfin, nous devons interroger l'adéquation entre les méthodes, attentes et outils traditionnels de la géographie et de la géomatique d'une part, et les caractéristiques de ce nouveau matériau d'autre part. Il s'agit de mettre en exergue les écueils inhérents aux conditions d'utilisation actuelles qui entravent le plein déploiement des tweets géolocalisés comme source de données permettant la construction et la représentation d'une connaissance géographique fiable et globale.

Terrain d'étude : réseau social numérique retenu et territoire cible

Critères de détermination du terrain d'étude.

La sélection d'un terrain d'étude ancré dans l'espace réel s'avère d'ores-et-déjà contrainte par deux facteurs : d'une part, la problématique du numérique et de son ancrage territorial et social, et d'autre part, la sélection d'un territoire à risques. Nous cherchons donc

un territoire subissant la récurrence de phénomènes naturels générateurs de dangers et suffisamment peuplé pour bénéficier d'une large adhésion aux plateformes permettant aux utilisateurs de produire de l'information numérique géolocalisée. En raison de ces contraintes, notre choix s'est porté sur les Etats-Unis :

- dans le pays d'origine des *Géants du Web*, les réseaux d'infrastructures de l'Internet mobile et très haut débit sont bien mieux développés que dans les pays européens : en 2017, aux Etats-Unis, 62% des foyers disposent d'un accès à Internet en très haut débit et 75% des foyers sont équipés d'au moins un smartphone[12]. La même année, en Europe, 63% des foyers disposaient d'un smartphone et l'utilisaient comme outil d'accès à Internet[13].

- L'adhésion aux plateformes et aux réseaux sociaux numériques via lesquelles les usagers diffusent de l'information est plus forte que dans les pays européens : en janvier 2017, le taux de pénétration des réseaux sociaux aux Etats-Unis était de 60% (49% en Europe). Notons néanmoins qu'entre 2017 et 2019, l'accroissement de l'utilisation des réseaux et médias sociaux s'est nettement plus marqué dans les pays du golfe persique (Qatar, Arabie Saoudite, Emirats Arabes Unis), en Asie du Sud et de l'Est (Malaisie, Inde, Indonésie, Phillippines), en Amérique du Sud (Brésil, Argentine) ainsi qu'en Turquie[14].

- Lorsque les médias européens évoquent la survenue de phénomènes naturels dommageables aux Etats-Unis, ils ont coutume de se focaliser sur les phénomènes extrêmes comme les incendies, les ouragans et les tempêtes de blizzard. Or, en dehors de ces phénomènes d'intensité exceptionnelle, des phénomènes locaux d'origine hydrométéorologique (tornades, inondations et crues torrentielles) sont récurrents dans certaines régions et peuvent provoquer des dégâts considérables[15].

- Enfin, aux Etats-Unis, une grande diversité d'organismes locaux ou fédéraux diffusent gratuitement, via des portails officiels, des jeux de données démographiques ou environnementaux : le bureau de recensement de la population (*Census Bureau*), la *NOAA*, le *NWS*, l'*USGS*, la *FEMA*, les portails *Open Data* des métropoles, *etc.*.

Parmi les territoires américains fréquemment frappés par des phénomènes naturels d'origine hydrométéorologique, on trouve les Etats du centre et du sud, dont le Texas : à l'instar des territoires bordant le Golfe du Mexique, le Texas subit des phénomènes météorologiques majeurs et réguliers qui entraînent pluies intenses, crues éclair, inondations ou encore tornades, auxquels s'est ajouté en août 2017 l'ouragan Harvey. Mais dans cet Etat, seuls les comtés proches du littoral sont exposés aux aléas cycloniques ; les comtés de l'ouest sont au contraire exposés à un aléa sécheresse élevé alors que tous les territoires de l'est et

[12] Chiffres issus du portail de données du bureau de recensement de la population *United States Census Bureau* : https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

[13] Chiffres issus du portail de données européen Eurostat : https://ec.europa.eu/eurostat/fr/data/database

[14] Sources : https://wearesocial.com/fr/blog/2017/01/digital-social-mobile-les-chiffres-2017 et https://wearesocial.com/fr/blog/2019/01/global-digital-report-2019 (Consulté pour la dernière fois le 10/06/2019)

[15] Source : https://www.texastribune.org/2018/07/05/houston-sees-first-major-flooding-harvey/ (Consulté pour la dernière fois le 24/01/2019)

du sud sont soumis à un aléa inondation élevé. Les populations les plus vulnérables face aux différents risques se concentrent quant à elles dans l'ouest et le sud, ainsi que dans la métropole de Houston (Emrich et Cutter, 2011) qui reste soumise à tous les types d'aléas d'origine hydrologique et sensibles au changement climatique (pluies intenses, cyclones, élévation du niveau des eaux, inondations).

En dépit de son exposition aux risques hydrométéorologiques, le Texas est un acteur économique majeur du pays : deuxième état le plus riche après la Californie, il compte, d'après les estimations du *Census Bureau*, 27 419 612 habitants en 2017 dont 66% se répartissent au sein de quatre aires métropolitaines : Dallas - Forth Worth (6ème aire du pays en termes de population), Houston (7ème aire), San Antonio (26ème aire) et Austin (37ème aire). Les activités économiques de l'Etat reposent sur trois piliers : l'élevage, l'industrie pétrolière et l'industrie des hautes technologies. Que dire de la répartition de la population et de la connectivité des territoires et des individus[16] ?

Caractéristiques démographiques et numériques de l'Etat du Texas.

Les foyers de peuplement se concentrent principalement dans l'est de l'Etat, au sein des grandes aires métropolitaines et sur le littoral du Golfe du Mexique ; quelques foyers se distinguent également dans le nord de l'Etat et le long de la frontière avec le Mexique (figure 0.2 en page suivante). Le *vide* constaté à l'ouest correspond à la région du Trans-Pecos, le Far West texan, une région désertée au climat aride. Mais qu'il s'agisse d'une aire métropolitaine comme Dallas ou d'un pôle urbain en milieu rural comme Amarillo au nord, ou encore El Paso à la frontière mexicaine à l'ouest, la transition entre territoires de banlieue résidentielle américaine et l'espace rural quasiment vide est rapide.

[16] On décrira ici la connectivité par l'accès au réseau très haut débit, qui constitue le réseau le plus performant, et par l'usage du smartphone qui représente le principal outil nomade de capture de données numériques géolocalisées créées par les usagers.

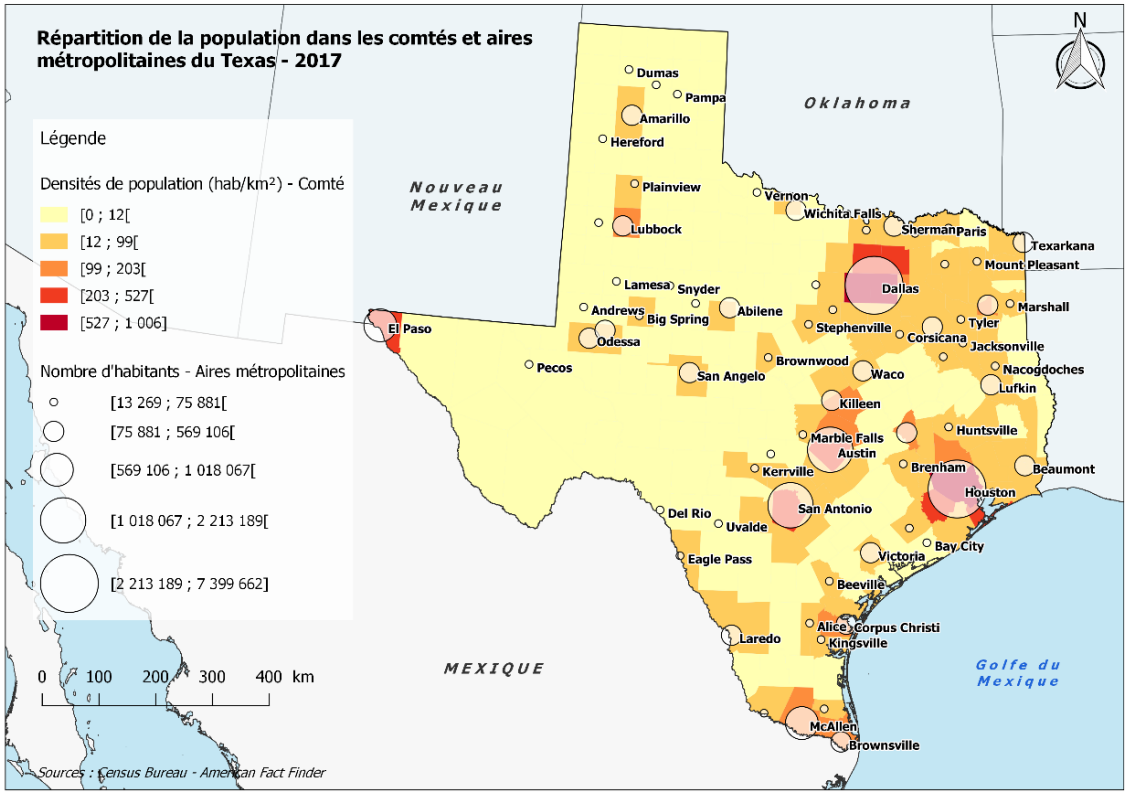


Figure 0.2 : Répartition de la population dans l'Etat du Texas, 2017 (C.Cavalière)

En dépit du poids économique de ses aires métropolitaines, le Texas ne fait pas partie des Etats dans lesquels les réseaux fixes de communication numérique sont les plus modernisés : en fait, les Etats du sud et du centre-est du pays accusent d'un léger retard dans la modernisation des réseaux et dans le passage au très haut débit : en 2017, le *Census Bureau* estime que 62% des foyers texans disposent d'un abonnement à l'Internet en très haut débit contre 72% pour des foyers californiens et 77% dans le Massachusetts. En revanche, l'usage des dispositifs mobiles comme le smartphone reste relativement homogène entre ces différents Etats : la même année, 75% des foyers texans disposent d'au moins un smartphone (76% en Californie).

A l'échelle des comtés texans, on peut constater une certaine hétérogénéité des territoires face à l'intégration du numérique (figure 0.3). Les comtés des grandes métropoles et pôles urbains constituent, sans surprise, des foyers d'usage du numérique. En dehors de ces territoires, aucune tendance ne semble transparaître : au moins 51% des foyers des comtés situés en marge des principales aires urbaines et métropolitaines disposent d'un abonnement au très haut débit (au nord-est d'Amarillo, les deux comtés de Lipscomb et Roberts affichent même des taux de foyers abonnés supérieurs à 78%). A l'inverse, les taux d'abonnés les plus faibles (tout au plus 50% des foyers) se distinguent dans les comtés de l'ouest et du sud de l'Etat, malgré la présence d'aires urbaines comme McAllen, Laredo et Brownsville.

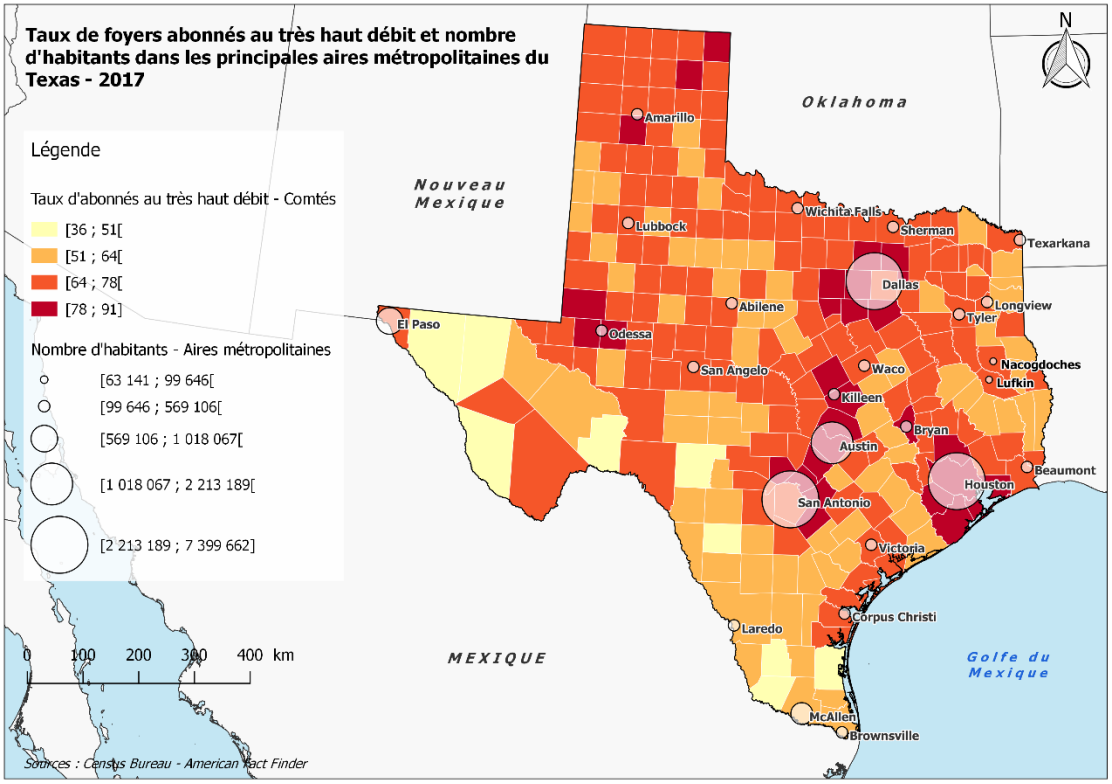


Figure 0.3 : Connexion au très haut débit et nombre d'habitants dans les principales aires urbaines et métropolitaines, 2017 (C.Cavalière)

A l'échelle d'une aire métropolitaine, les tendances se révèlent davantage perceptibles : la figure 0.4 représente le taux de foyers abonnés au très haut débit des *census tracts*[17] inclus dans l'aire métropolitaine de Houston, ainsi que le taux de foyers ne disposant d'aucun abonnement à un réseau internet quelconque. On constate une certaine homogénéité des territoires intra-urbains, quelle que soit leur distance au centre : dans 62% des *census tracts*, au moins 75% des foyers disposent d'un accès au très haut débit. A l'inverse, les *census tracts* dont la connectivité est plus faible se répartissent selon deux logiques : on les trouve dans les territoires en marge de l'aire métropolitaine et sur une couronne nord-est-sud autour du centre des affaires (le *Central Business District*, représenté par une étoile). Cette même logique se dessine sur la carte des taux de foyers ne disposant d'aucun abonnement à l'Internet : dans 36% des *census tracts*, au moins un quart des foyers sont exclus du très haut débit.

[17]Aux Etats-Unis, le *Census Tract* correspond à une unité de recensement d'échelle fine, qui contient une population statistiquement homogène en termes de profil démographique, économique et social.

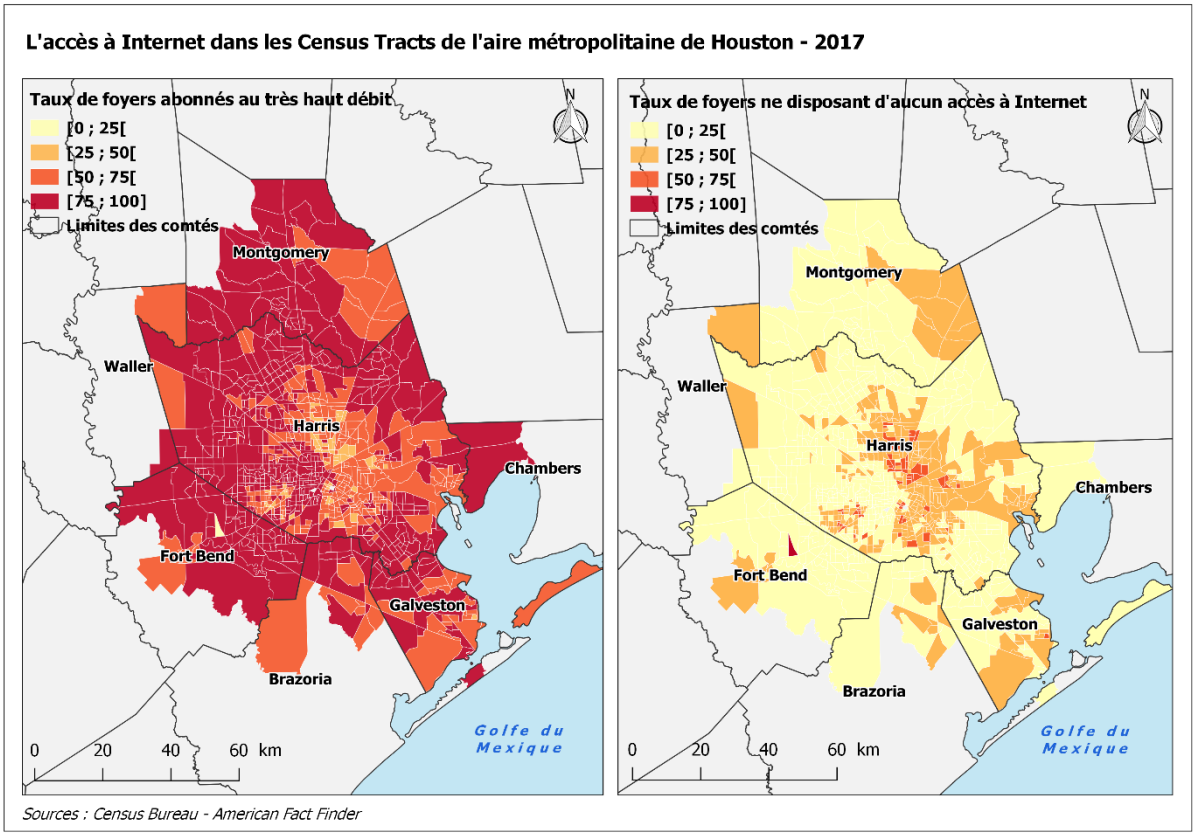


Figure 0.4 : Taux d'accès à l'Internet très haut débit des foyers de l'aire métropolitaine de Houston, 2017 (C.Cavalière)

Des structures similaires sont identifiables lorsqu'on s'intéresse à l'usage du smartphone dans ces mêmes unités statistiques de recensement (figure 0.5) : dans 57% des *census tracts*, 75% des foyers disposent d'au moins un smartphone. Ils sont en revanche 5,4% dans lesquels moins de 50% des foyers disposent d'au moins un smartphone. Une nouvelle fois, ces territoires dans lesquels l'empreinte du périphérique nomade semble la plus timide se concentrent sur une couronne nord-est-sud par rapport au *CBD*, et dans les marges de l'aire métropolitaine.

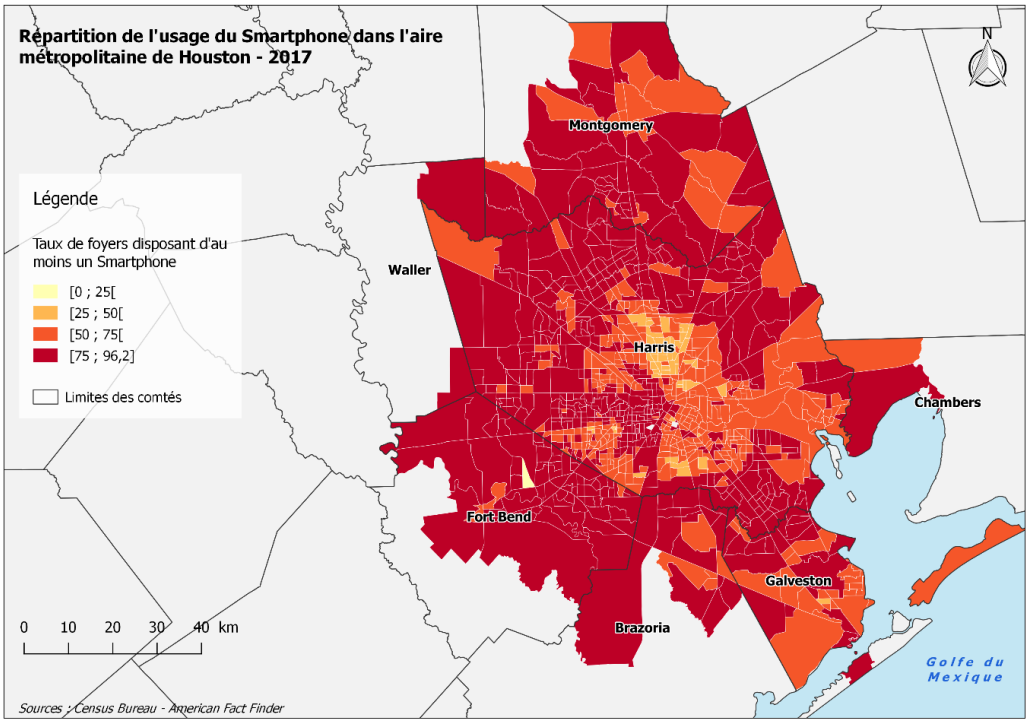


Figure 0.5 : Taux de foyers disposant d'au moins un Smartphone dans l'aire métropolitaine de Houston, 2017 (C.Cavalière)

Peut-on alors vraiment évoquer l'existence d'une fracture du numérique dans les territoires du Texas ? En dehors des grandes métropoles qui constituent les territoires de prédilection du numérique, l'accessibilité au réseau du très haut débit est marquée par une variabilité spatiale certaine mais difficile à décrire : certains comtés, bien que situés en marge des aires métropolitaines, bénéficient du déploiement du réseau très haut débit, alors que d'autres, qui concentrent pourtant le même nombre d'habitants, restent en marge du réseau. En revanche, la répartition des smartphones dans l'aire métropolitaine ne semble pas indiquer l'existence d'un fossé entre populations adeptes et populations exclues[18]. Contrairement à la répartition de la population qui contrastait entre *vides* et *pleins*, le numérique, dans ces *pleins*, ne marque pas de fracture tangible qui se concrétiserait par des territoires hyperconnectés et des territoires invisibles. On constate plutôt une hétérogénéité des territoires, se manifestant par des niveaux de connectivité variables.

Critères de détermination et présentation du réseau virtuel.

De la même manière, la détermination d'un réseau virtuel retenu est orientée autour de critères précis : en premier lieu, nous cherchons des données numériques dont la géolocalisation est précise, c'est-à-dire directement réalisée par l'intermédiaire du GPS inclus dans le smartphone (et non ajoutée *a posteriori* par l'usager de la plateforme numérique).

[18] En revanche, on ne dispose pas de données relatives à la téléphonie nomade, ni aux pratiques d'utilisation des réseaux Internet, qui pourraient nous aider à qualifier les fréquences d'utilisation des plateformes du Web.

Dans un second temps, et pour faciliter l'acquisition des données, celles-ci doivent pouvoir être collectées rapidement et en continu. Notre choix s'est alors porté sur la plateforme de médias sociaux *Twitter.* En plus de répondre à ces deux conditions préalables, ce média social offre d'autres avantages :

- son principe de communication est différent des autres réseaux sociaux : alors que Facebook s'apparente à un réseau *total*, qui englobe des fonctionnalités de partage de médias et différents modes de communication, Twitter est fondé sur un principe de communication unique, rapide et instantané, pouvant inclure des contenus variés, et dont le destinataire peut être soit un ou plusieurs individus désignés, soit le réseau dans sa globalité.

- Un grand nombre d'institutions disposent d'un compte Twitter et diffusent de l'information par ce réseau, en particulier lorsqu'il s'agit de messages d'alerte destinés à la population civile. C'est le cas notamment du *National Weather Service* de Houston (*NWS*, centre de prévision météorologique à l'échelle du comté), comme l'illustre la figure 0.6 : l'alerte mentionne le territoire concerné de manière globale (elle est ici associée à la rivière *Trinity* qui traverse le comté de Liberty), les temporalités de validité de l'alerte, le type de risques et événements associés (ici, un risque d'inondation d'axes de communication en milieu urbain) ainsi qu'une recommandation comportementale ("*turn around, don't drown*", en référence à la consigne de ne pas s'engager sur une route inondée).

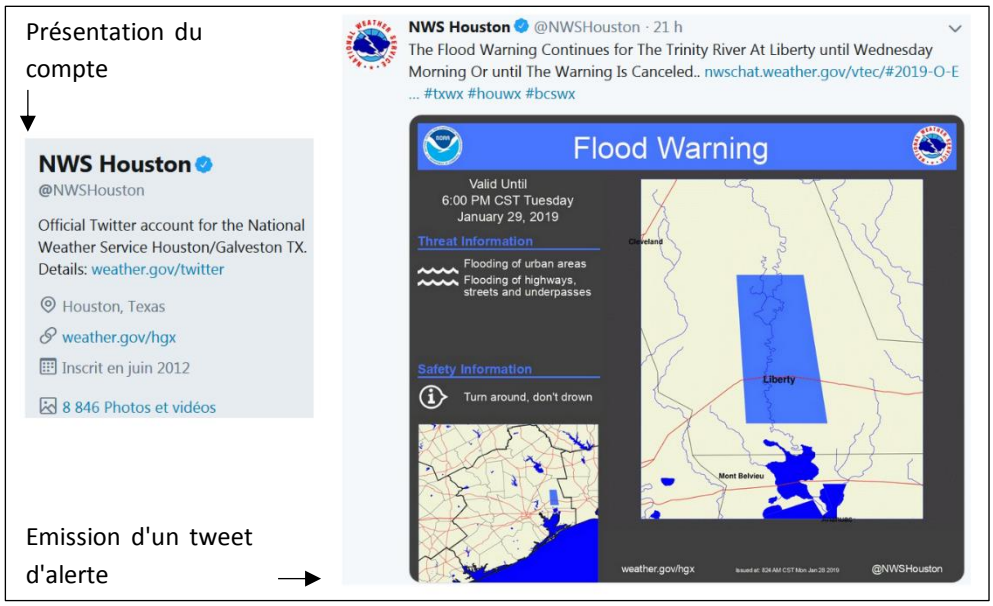


Figure 0.6 : Tweet diffusant une alerte inondation, émis le 28/01/2019 par le centre de prévisions météorologiques de Houston

- Twitter illustre bien l'*effet déluge* (Lin *et al*., 2013) des réseaux et médias sociaux : le réseau constitue l'espace virtuel via lequel les usagers s'expriment en masse et dans l'immédiat quand survient un événement inhabituel. La figure 0.7 représente les quantités de tweets contenant le hashtag *#Présidentielles2017* émis le 7 mai 2017 en France (soit le jour du second tour des élections présidentielles). Le *déluge de tweets* survient en réaction

immédiate à l'annonce officielle des résultats par les médias, soit dans les deux premières minutes après 20h.

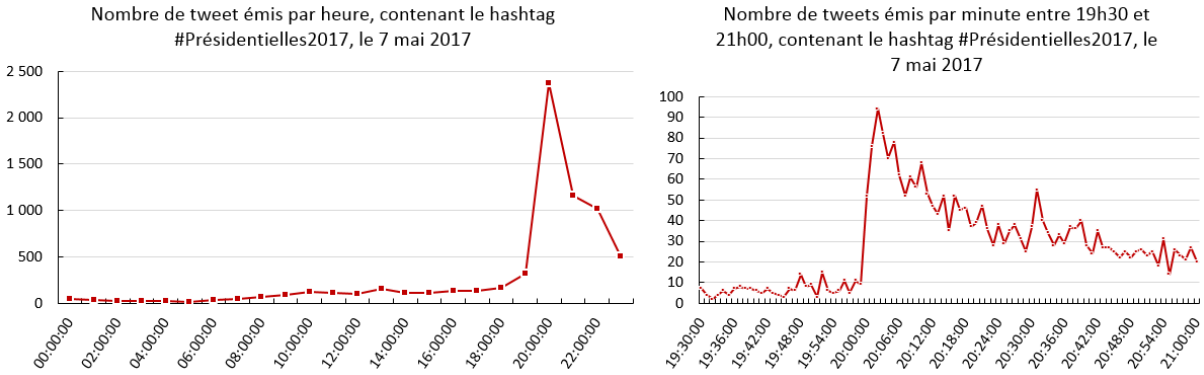


Figure 0.7 : Effet déluge des médias sociaux émis en réaction à un événement - Les résultats du second tour de l'élection présidentielle en France en 2017

- Quand bien même Facebook reste la plateforme numérique de réseau social la plus utilisée dans le monde, l'audience de Twitter, aux Etats-Unis, n'est pas négligeable : le média social enregistrait, en septembre 2016, 82,2 millions d'utilisateurs actifs sur smartphone, soit 25,6% des 320 millions d'utilisateurs mensuels actifs enregistrés dans le monde entier. En France, à la même période, ils étaient 6 millions, soit 1,8%.

Pour récapituler cette partie, le tableau 0.2 ci-après inventorie les phénomènes et échelles des territoires qui seront à l'étude dans ce manuscrit. Les premières analyses entreprises se focalisent sur des phénomènes récurrents de crues rapides et d'inondations liées à des épisodes pluvieux intenses saisonniers qui ont eu lieu en début d'année 2016. Le terrain d'étude ayant ensuite été frappé par un autre phénomène météorologique extrême en 2017, l'Ouragan Harvey, nous avons alors jugé nécessaire d'intégrer cet autre type de perturbation dans la recherche, dans une perspective de comparaison entre les résultats observés lors des phénomènes récurrents et ceux qui concernent cet autre phénomène extrême, moins fréquent sur ce territoire.

Tableau 0.2 : Caractéristiques des phénomènes météorologiques à l'étude

Phénomènes extrêmes récurrents de crues rapides et d'inondations	Phénomènes pluvieux et inondations résultantes - printemps 2016	Etat du Texas et aire métropolitaine de Houston
Phénomène extrême cyclonique moins fréquent	Ouragan Harvey (catégorie 4) - 23-31 août 2017	Ensemble des territoires affectés (sud-est du Texas et ouest de la Louisiane) Aire métropolitaine de Houston

Plan du manuscrit

Le manuscrit se subdivise en deux parties de trois chapitres. Dans la première partie, qui constitue l'état de l'art, nous nous focalisons sur le contexte de la transformation récente des acteurs et des formes des données géographiques, ainsi que sur les caractéristiques de ces nouvelles données et les problèmes que soulèvent leur exploitation. Alors que les jeux de données traditionnellement utilisés par les sciences géographiques reposent sur des méthodes rigoureuses, mises en œuvre par des organismes de professionnels, la donnée produite par l'usager des plateformes numériques, dont le tweet géolocalisé fait partie, se décline sous plusieurs formes et sa vocation première n'est pas nécessairement la cartographie et l'analyse spatiale. Dans un premier chapitre, nous rappelons alors ce que les sciences de l'information géographique considèrent comme données et information géographiques afin de positionner les nouvelles formes de données géolocalisées produites au quotidien sur le Web. Dans un deuxième chapitre, nous présentons les caractéristiques du matériau précisément à l'étude de cette recherche, le tweet géolocalisé par GPS : il s'agira notamment de souligner les écueils inhérents aux conditions actuelles d'utilisation de la plateforme, qui entravent le plein déploiement des tweets géolocalisés comme source de données permettant la construction d'une connaissance géographique fiable et globale. Enfin, le troisième chapitre soumet alors un bilan des horizons explorés dans les diverses problématiques spatiales et/ou comportementales, les méthodes d'analyse couramment employées ainsi qu'un état de la question des tweets géolocalisés, spécifique aux risques et catastrophes naturels.

Quels apports et limites peut-on identifier dans les tweets géolocalisés à la géographie des populations et des territoires en crise ? Peut-on prévoir les effets d'un phénomène naturel sur le réseau virtuel géolocalisé ? Peut-on envisager de produire, par l'exploitation des tweets géolocalisés, de l'information géographique au même titre qu'en ayant recours aux données traditionnelles ? Dans cette perspectives, la seconde partie du manuscrit présente les apports méthodologiques et thématiques de la recherche : le quatrième chapitre décrit la démarche méthodologique adoptée dans cette recherche, et définie par rapport à l'existant ; le cinquième chapitre aborde la question de l'étude spatiale, sémantique et quantitative des phénomènes et événements réels retranscrits sur le réseau virtuel à l'échelle globale du phénomène physique (il constitue la "primo-connaissance" qui a orienté les analyses dont les résultats sont présentés dans le dernier chapitre). Pour terminer, le sixième chapitre explore les caractéristiques du tweet géolocalisé comme indicateur potentiel des dynamiques et effets locaux des crises d'origine naturelle, à l'échelle de la métropole. L'organisation globale des chapitres du manuscrit est résumée dans la figure 0.8, en page suivante :

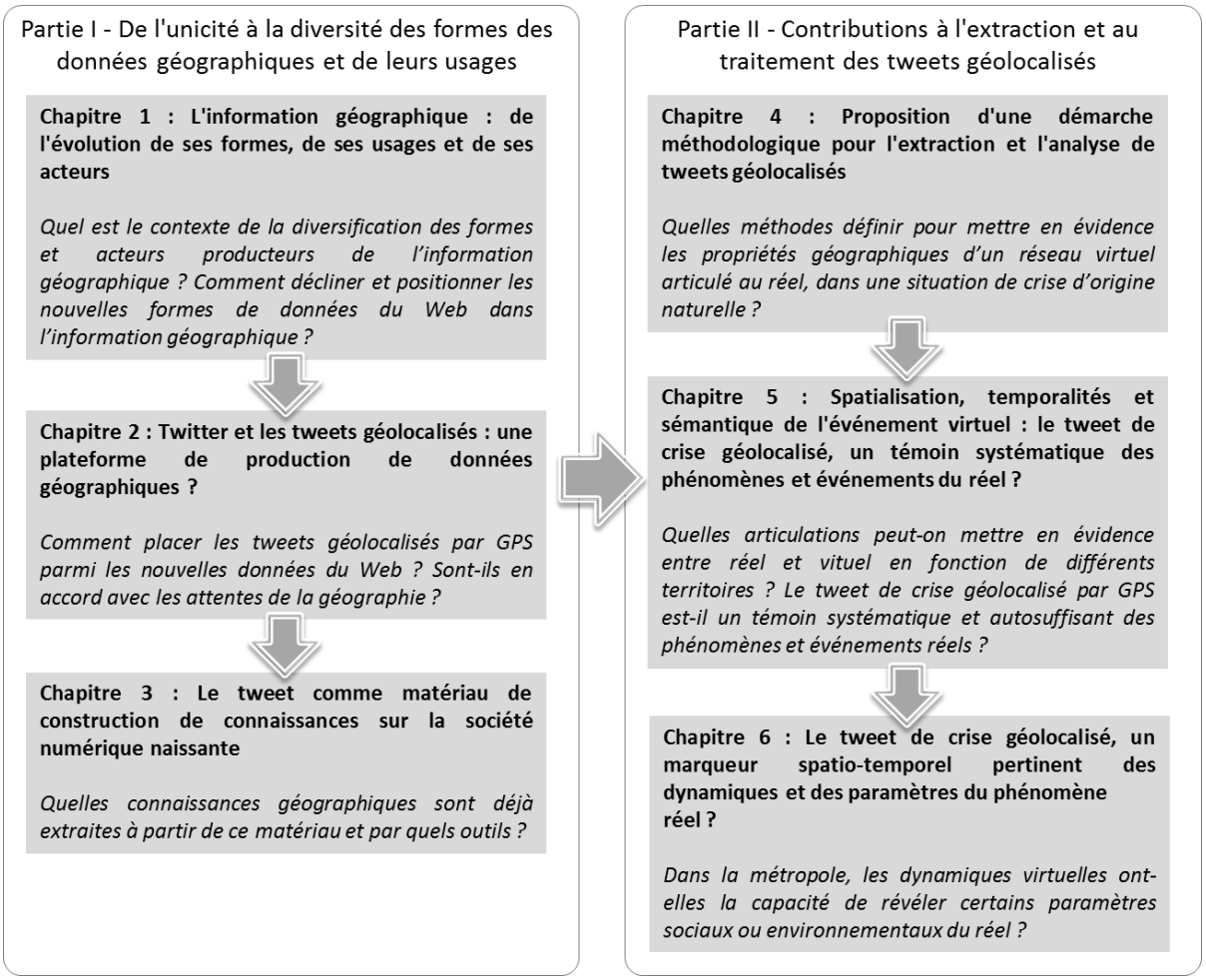


Figure 0.8 : Organisation des chapitres du manuscrit

Partie I

De l'unicité à la diversité des formes des données géographiques et de leurs usages

Partie I : De l'unicité à la diversité des formes des données géographiques et de leurs usages

La première partie de ce manuscrit est articulée autour de deux objectifs : en premier lieu, elle expose l'état présent de l'ensemble des matériaux géolocalisés, créés quotidiennement par des individus connectés, qui se dissimulent derrière l'appellation désormais trop générale de donnée ou d'information géographique. Le second objectif est articulé autour du matériau précis à l'étude de cette recherche, le tweet géolocalisé : il s'agit de le situer parmi les nouvelles formes de données géolocalisées mais également de souligner les enjeux associés à ses conditions de production ainsi qu'à ses possibilités de traitement dans une perspective géographique.

Cette partie est ainsi structurée en trois chapitres : le premier chapitre rappelle les éléments de définition traditionnels et académiques de l'information géographique et de sa cartographie, puis présente les évolutions des technologies et des usages ayant favorisé l'ubiquité et la diversité des formes des données géolocalisées. Le deuxième chapitre décrit les caractéristiques précises du matériau tweet géolocalisé afin de le positionner au sein des nouvelles formes de données géolocalisées, dans la continuité du premier chapitre. Enfin, le troisième chapitre dresse un état des méthodes et thèmes abordés dans une décennie de travaux scientifiques effectués à partir du matériau tweet.

1. L'information géographique : de l'évolution de ses formes, de ses usages et de ses acteurs

Ce premier chapitre spécifie le cadre référentiel de ce que le géographe professionnel qualifie d'*information géographique*, soit le matériau indispensable et préalable à toute représentation cartographique des phénomènes ancrés sur l'espace terrestre (Lambert & Zanin, 2016) et dont l'utilité publique n'est plus à démontrer (Béguin & Pumain, 2017). Il s'applique non seulement à en donner les définitions communément acceptées et qui restent traditionnellement enseignées aux étudiants de premier cycle universitaire, mais encore à en présenter les propriétés et l'évolution de son exploitation à travers son histoire. Pour autant, notons d'emblée que ce chapitre n'a pas vocation à être exhaustif quant aux nombreux concepts qui gravitent autour de l'information géographique et qui soulèvent encore des débats dans la recherche. Il met en exergue les ambigüités consécutives à la définition de l'information géographique, alors que ses origines, formes et usages sont désormais multiples et hétérogènes.

Restée en effet pendant des décennies aux mains des personnes dites *expertes* (McDougall, 2012), l'information géographique s'est, depuis une dizaine d'années, amplement démocratisée et s'invite aujourd'hui dans le quotidien des internautes. Qui n'a jamais recherché la localisation d'une rue, ou un itinéraire précis en interrogeant des outils de cartographie collaborative ? Ubiquiste, elle est acquise par une grande diversité de capteurs - de la station *ATMO* localisée en un point précis d'une commune, qui mesure le niveau de pollution aux particules fines dans l'air ambiant, au *GPS* embarqué dans un véhicule particulier, qui suit ses déplacements - et produite par une multiplicité d'acteurs, qu'ils soient professionnels ou amateurs. En conséquence, ces nouvelles sources d'information géographique, créées et diffusées quotidiennement sur le Web, et accessibles pour tout un chacun, s'avèrent parfois très différentes de l'information géographique dite *traditionnelle* et *expertisée*, en termes de forme, mais également de contenu et de richesse (Kitchin, 2013).

L'objectif de ce premier chapitre est ainsi triple : dans un premier temps, il définit et problématise le concept d'information géographique ; ensuite, il retrace l'évolution des pratiques liées à la production de l'information géographique et à son exploitation dans une perspective historique ; enfin, il propose un tour d'horizon des propriétés de ces formes géographiques nouvelles ainsi qu'une réflexion quant à leur positionnement dans la sphère de l'information géographique.

1.1. Un concept historiquement articulé autour d'acteurs et d'enjeux particuliers

Considérée comme support d'acquisition de connaissances et d'interrogation de phénomènes à dimension spatiale (Baud *et al*., 2009), l'information géographique a fait l'objet d'un spectre de définitions multi-niveaux, plus ou moins approfondies, aux critères déterminés par des communautés d'utilisateurs *professionnels*. Ainsi, les géographes considèrent, d'une manière générale, que toute information accompagnée d'une propriété de localisation sur ou à proximité de la surface de la Terre peut être qualifiée d'information géographique (Goodchild & Glennon, 2010 ; Lambert & Zanin, 2016). Si cette définition souligne le rôle clé que constitue la capacité à localiser un objet dans l'espace, elle reste néanmoins peu précise quant à la forme et au contenu attendus de ce qu'un professionnel qualifiera d'information géographique. Les paragraphes qui suivent soumettent alors une revue des caractères primordiaux requis pour considérer une information comme géographique.

1.1.1. Tour d'horizon des caractéristiques de l'information géographique des professionnels

1.1.1.1. La dimension spatiale

La capacité à localiser un objet de l'espace, c'est-à-dire à mesurer et à restituer son emplacement sur la surface de la Terre constitue le pilier fondamental de l'information géographique (Béguin & Pumain, 2017). Cette fonction représente le premier pas vers la représentation et donc vers la connaissance du territoire, depuis l'Antiquité : considérée comme la première carte d'un réseau routier, la Table de Peutinger est très vraisemblablement née d'un besoin pragmatique de localiser et de nommer les objets de l'*œcoumène* afin d'organiser des itinéraires rationnels et de maîtriser le territoire. Elle représente en effet les villes de l'Empire romain, hiérarchisées selon leur importance, ainsi que les voies romaines empruntées par le *cursus publicus* (le service de poste impérial). Dans l'histoire de la cartographie, ce besoin de cartes représentant des réseaux de communication et itinéraires n'a fait que s'intensifier et se moderniser, depuis les portulans jusqu'au moteur de recherche actuel, qui calcule automatiquement des itinéraires en fonction de contraintes (figure 1.1).

Figure 1.1 : Tracer son itinéraire de l'Antiquité (Table de Peutinger, Wikipédia) à l'heure du numérique (Via Michelin)

Depuis l'avènement des *Systèmes d'Information Géographique (SIG)*[1] et le recours généralisé au *GPS*[2], la définition de la composante spatiale de l'information géographique s'est complexifiée : elle englobe désormais les propriétés de localisation d'un élément de la surface terrestre, exprimées dans un système à double coordonnées géographiques (X pour la latitude et Y pour la longitude), mais également la forme géométrique qui décrit de type de l'élément (Lambert & Zanin, 2016) : ce type peut être de nature ponctuelle, linéaire ou zonale. Dans cette géométrie, on peut encore inclure la topologie des éléments figurant sur une carte, qui caractérise des logiques de relation spatiale (inclusion, éloignement, contiguïté, etc.) entre les éléments cartographiés (Lambert & Zanin, 2016).

La localisation d'un objet à la surface de la Terre constitue ainsi l'un des fondements de l'information géographique mais elle ne peut s'y réduire. Une définition plus complète de l'Association Française pour l'Information Géographique (AFIGEO) propose de considérer l'information géographique non seulement comme la description de la position géographique d'un objet mais encore comme l'ensemble des informations qui caractérisent les propriétés de cet objet[3].

[1] "*Système informatique de matériels, de logiciels, et de processus conçus pour permettre la collecte, la gestion, la manipulation, l'analyse, la modélisation et l'affichage de données à référence spatiale afin de résoudre des problèmes complexes d'aménagement et de gestion*".
Source : http://seig.ensg.ign.fr/fichchap.php?NOFICHE=FP15 (Consulté pour la dernière fois le 21/02/2019)

[2] Propriété des Etats-Unis mais libre dans le domaine civil depuis les années 2000, il s'agit du système de géolocalisation basé sur la trilatération satellitaire.

[3] Source : http://www.afigeo.asso.fr/information-geographique.html (Consulté pour la dernière fois le 07/03/2018)

1.1.1.2. La dimension attributaire

La composante attributaire ou sémantique est généralement définie comme les caractéristiques d'un objet géographique[4] ; elle regroupe alors l'ensemble des propriétés qui décrivent chaque objet localisé sur le territoire. Dans les SIG, cette composante prend la forme d'un tableau de données, nommé table attributaire, dont les lignes représentent les éléments géographiques du territoire et les colonnes, les attributs (ou variables) qui décrivent ces éléments. Les attributs peuvent être spatiaux (la ville est incluse dans tel département), quantitatifs (la ville recense un certain nombre d'habitants) ou qualitatifs (la ville peut avoir un statut particulier : chef-lieu de canton, sous-préfecture, préfecture, etc.).

La composante attributaire est à l'origine de la cartographie thématique. En géomatique, la cartographie thématique est définie comme *la répartition spatiale des données relatives à un ou plusieurs thèmes particuliers des secteurs géographiques choisis*[5]. Une carte thématique n'est donc pas réduite au rôle de localiser les objets ; elle permet de visualiser la distribution spatiale des modalités ou valeurs d'un ou de plusieurs attributs contenus dans la table attributaire et constitue ainsi un outil d'aide à la réflexion et à la décision[6].

De la même manière que la capacité à localiser des objets et à tracer des itinéraires constitue un héritage, la cartographie thématique n'est pas l'apanage des outils numériques : cette question de la représentation des données associées à des éléments de fond cartographique émerge en effet dans la cartographie du XIX^{ème} siècle. Ainsi, l'invention du procédé de constitution de la carte choroplèthe, une carte thématique qui associe une étape de discrétisation de données quantitatives continues et la communication visuelle de la partition résultante des valeurs, est attribuée à l'ingénieur polytechnicien Charles Dupin[7]. En 1826, celui-ci publie la *Carte figurative de l'instruction populaire* (figure 1.2) qui représente le taux de scolarisation par département français (alors que l'instruction primaire n'est pas encore obligatoire), en échelle de gris. Plus la tonalité est sombre, plus le taux de scolarisation est faible et réciproquement ; les règles de sémiologie graphique n'ayant été standardisées qu'au siècle suivant, les tonalités employées traduisent la métaphore de l'instruction (l'éducation apporte la lumière alors que l'ignorance condamne à l'obscurité).

[4] Définition donnée par l'entreprise ESRI, propriétaire du logiciel SIG ArcGIS.
[5] Source : https://www.statcan.gc.ca/pub/92-195-x/2011001/other-autre/theme/def-fra.htm (Consulté pour la dernière fois le 11/04/2018)
[6] Source : http://www.hypergeo.eu/spip.php?article377 (Consulté pour la dernière fois le 11/04/2018)
[7] Source : http://www.hypergeo.eu/spip.php?article274 (Consulté pour la dernière fois le 10/07/2018)

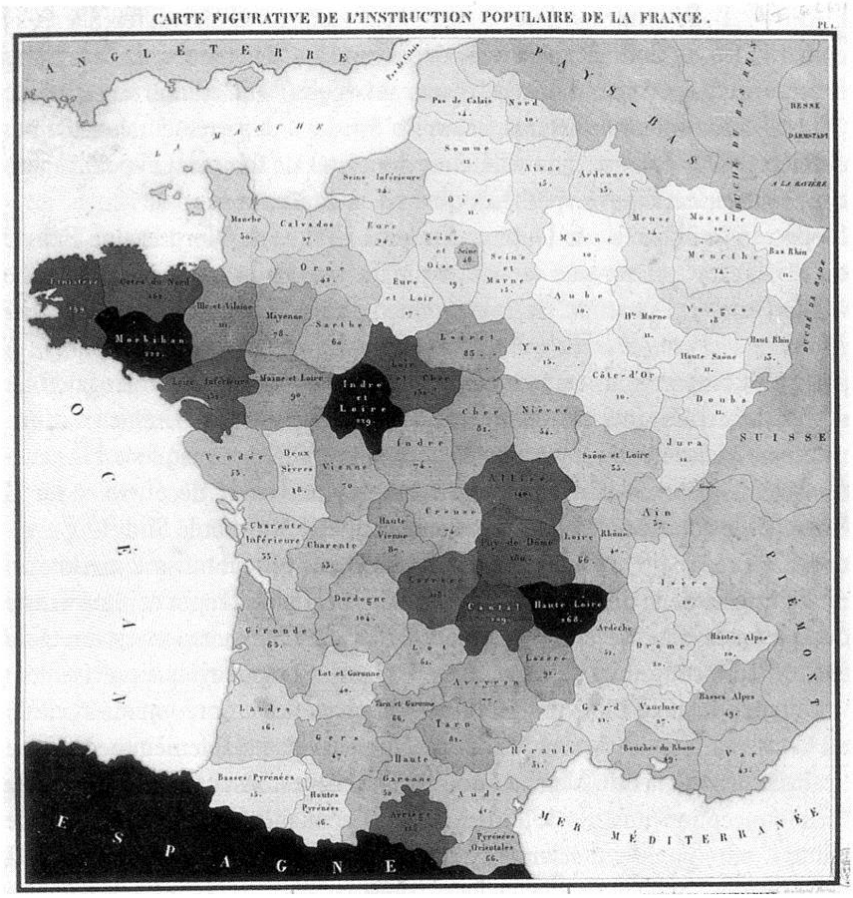


Figure 1.2 : Carte figurative de l'instruction populaire, réalisée par Charles Dupin en 1826 (BNF, Gallica)

L'intégration du numérique à la cartographie a favorisé l'accroissement rapide de la production de cartes thématiques, dans la mesure où les outils informatiques facilitent les étapes de collecte, de stockage, de diffusion et de représentation automatisée de l'information géographique (Béguin & Pumain, 2017). Et en effet, cartes et plans abondent sur le Web : la presse en ligne se positionne comme productrice régulière, notamment lorsqu'il s'agit de localiser des événements ou encore de montrer l'ampleur d'un phénomène à ses lecteurs. En outre, dans la lignée du mouvement de l'*Open Data*[8], il est désormais tâche facile et rapide pour un individu possédant des compétences élémentaires en géomatique, de télécharger des jeux de données et de produire ses cartes.

[8] L'*Open Data* fait l'objet d'une définition précise dans la deuxième partie de ce chapitre. Dans le cas de la France, une grande diversité de jeux de données sont disponibles sur le portail gouvernemental Data.gouv, qui affiche par ailleurs les "Réutilisations", c'est-à-dire les productions cartographiques réalisées à patir de jeux de données téléchargées via le portail et mises en ligne.

1.1.1.3. Les types d'information géographique

L'information de référence.

Les géographes peuvent évoquer l'information géographique de manière générale, mais celle-ci se subdivise en plusieurs catégories. Les cartes héritées de l'Antiquité ayant une finalité de localisation des objets sur le territoire, de repérage, de délimitation administrative de parcelles, etc. correspondraient aux données classifiées, de nos jours, comme information géographique de référence : ce qualificatif est ainsi associé à des données *généralistes* (Lambert & Zanin, 2016), c'est-à-dire des fonds de carte de type maillage administratif, réseau routier, etc. sur lesquels peuvent être superposés d'autres types de données (Pornon, 2015). Parmi ces informations de référence, on retrouve les gammes de produits SCAN®, ADMIN EXPRESS® et RGE® de l'Institut Géographique National (IGN) qui fournissent respectivement une bibliothèque d'images localisant les éléments de connaissance de base du territoire (villes, routes, etc.) français, ainsi que des couches d'information vectorielles ou matricielles permettant de cartographier les divers objets et formes des territoires (bâtiments, réseaux, parcelles, altitudes) ou encore les divers échelons du maillage territorial administratif français. Enfin, le portail de consultation de l'information géographique de l'IGN, Géoportail, propose également diverses couches d'information de référence, sous l'appellation "Fonds de carte" : cartes topographiques, photographies aériennes, parcelles cadastrales, etc.

L'information thématique.

Aux côtés de cette information de référence s'ajoute l'information dite thématique, relative à un domaine particulier et permettant d'enrichir la simple description de l'espace fournie par l'information de référence[9]. C'est ce type d'information qui va permettre de localiser et de quantifier, non plus des objets du territoire, mais des phénomènes spatiaux dont la nature et/ou l'intensité peut varier dans l'espace : dans la figure 1.2 (cf.page précédente), on peut assimiler l'information de référence au fond de carte constitué des départements français, et des Etats frontaliers ; l'information thématique est ici une variable quantitative continue, le taux d'enfants scolarisés dans l'enseignement primaire, représentée par l'échelle de tonalités grises.

L'information thématique peut ainsi concerner de nombreux domaines (environnement, risques, santé, éducation, démographie, conditions de vie, culture, criminalité, économie, etc.) et se trouve désormais diffusée, depuis l'avènement de l'*Open Data*, via de nombreux portails institutionnels, comme le *data.gouv* français ou *Eurostats*[10] à

[9] Source : http://www.afigeo.asso.fr/information-geographique.html (Consulté pour la dernière fois le 07/03/2018)

[10] Notons que les données diffusées par ces portails ne sont pas restreintes aux formats des couches d'information directement exploitable dans les SIG ; le plupart des jeux de données thématiques téléchargeables sont en effet des jeux de données statistiques (donc des tableaux) que l'utilisateur doit associer à son fond de carte de référence par une opération de jointure.

l'échelle de l'Union Européenne (Lambert & Zanin, 2016). La figure 1.3 ci-dessous affiche une carte représentant le zonage territorial établi en fonction des niveaux d'aléas naturels, par le Plan de Prévention des Risques Naturels (PPRN) de la commune de Sallanches. Les données de référence sont constituées d'un fond de carte *OpenStreetMap* qui localise l'ensemble des éléments permettant à l'utilisateur de se repérer sur le territoire. Les données thématiques, téléchargées sur le portail *data.gouv*, correspondent au zonage établi par la direction départementale des territoires de Haute-Savoie en fonction de trois couleurs : le rouge correspond à un niveau d'aléa fort (toute nouvelle construction est généralement interdite) ; les bleus correspondent à un aléa moyen : toute construction est envisageable mais reste soumise à prescriptions.

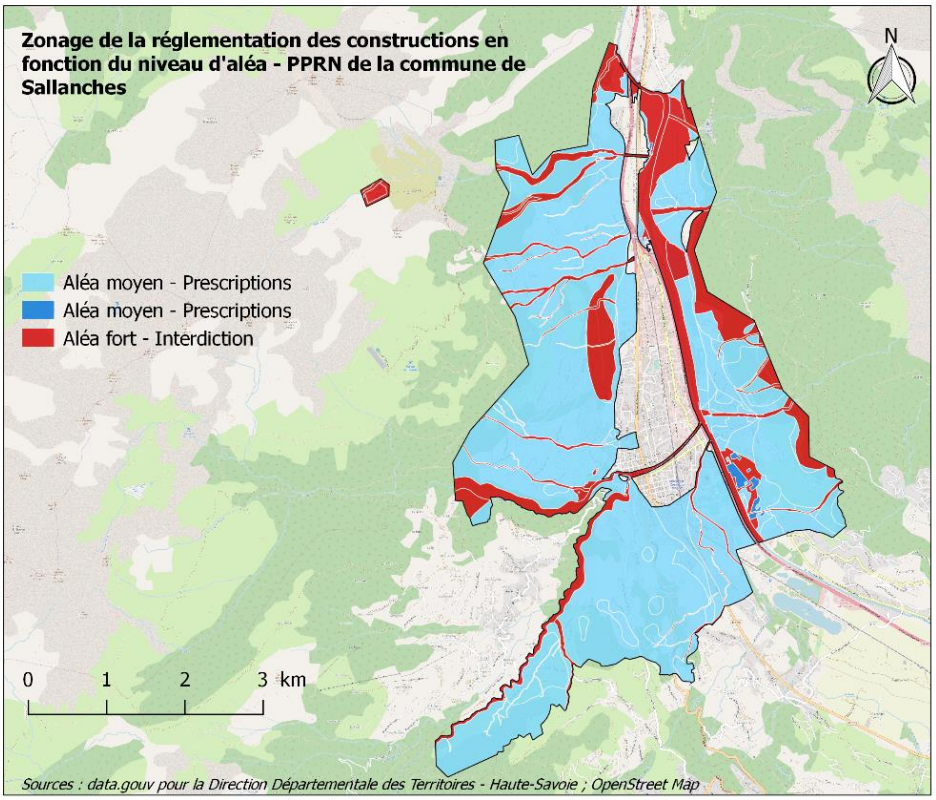


Figure 1.3 : Information de référence et information thématique : élément du finage communal et zonage des réglementations de construction en fonction des niveaux d'aléas (data.gouv , OpenStreetMap), C.Cavalière

L'information temporelle.

Enfin, les géographes sont fréquemment amenés à étudier des phénomènes et objets localisés dont l'évolution est mesurée dans le temps : le recensement d'une même population entre deux dates permet de calculer un taux d'évolution ; une station météorologique reste figée en un point précis du territoire, mais enregistre des séries de données (température, pluviométrie, vitesse du vent, humidité, etc.) dont les valeurs varient en fonction de différentes échelles de temps (heure, saison). On peut ainsi définir un dernier type d'information géographique : l'information temporelle. Les données relatives aux risques

naturels n'échappent pas à cette dimension : un aléa peut en effet se caractériser par sa fréquence d'apparition, par sa durée, ou encore par sa logique et sa vitesse de diffusion dans l'espace (Davoine, 2014). Comment les géographes caractérisent-ils alors les données temporelles ?

En premier lieu, le temps, à la manière de l'espace, se caractérise par un référentiel propre, le calendrier (Saint-Marc, 2017). Les données géographiques et temporelles, que les géographes qualifient alors de spatio-temporelles, disposent de deux localisations : la géolocalisation dans un repère spatial par des coordonnées géographiques et la "localisation" temporelle qui positionne l'objet, la mesure ou le phénomène dans le calendrier par une date. Cette date peut se décliner à différents niveaux de granularité (Saint-Marc, 2017) : un recensement de population s'effectue sur un pas de temps large (par exemple, tous les dix ans aux Etats-Unis) et est disponible pour une année donnée (1990, 2000, 2010, le prochain étant prévu pour 2020). La figure 1.4 ci-dessous représente la courbe des mesures associées à la station de jaugeage *Whiteoak Bayou*, pendant la survenue de l'ouragan *Harvey* en 2017. Cette station est implantée sur le fleuve *Buffalo Bayou*, dans le *CBD* de la ville de Houston, et traverse l'aire métropolitaine d'ouest en est avant de se jeter dans la baie de Galveston. La station est localisée et repérée sur le territoire par ses coordonnées géographiques et mesure la hauteur du niveau de l'eau selon un pas de temps de quinze minutes.

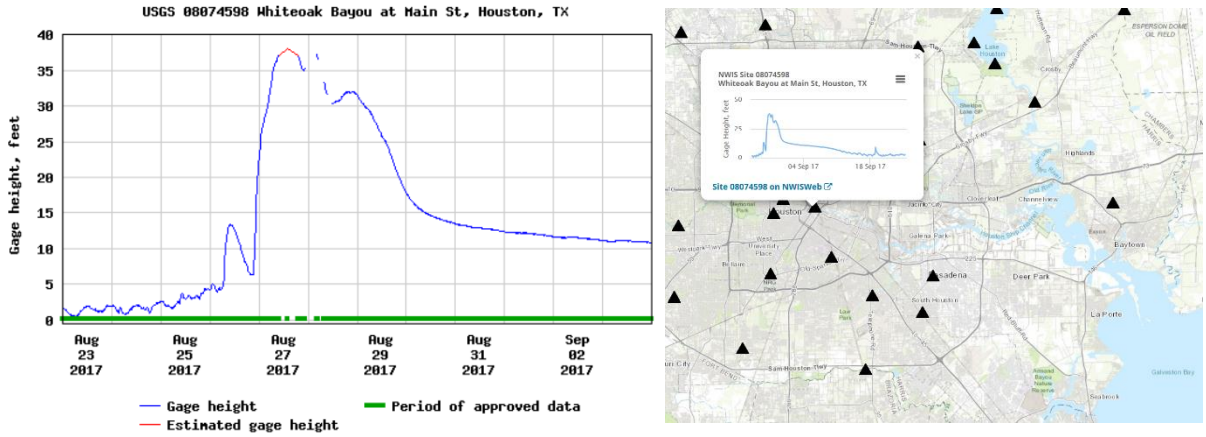


Figure 1.4 : La station de jaugeage : une information temporelle de granularité fine et géolocalisée (USGS Water Resources)

Une série de stations de jaugeage disposées de l'amont vers l'aval d'un cours d'eau constitue alors un réseau d'observation, certes fixe, mais dont la capacité à capturer des données dans un intervalle temporel précis, permet de suivre les pics de crue, en d'autres termes, la propagation d'un phénomène naturel.

Si l'information thématique permet de représenter la distribution et la variabilité d'un phénomène dans l'espace, l'information spatio-temporelle a donc l'avantage d'en saisir les dynamiques : quels territoires ont perdu ou attiré des habitants en l'espace de dix ans ? Observe-t-on sur plusieurs années, grâce aux stations de jaugeage implantées sur les rives

d'un cours d'eau, des modifications des normales saisonnières ou une intensification des phénomènes extrêmes ? Les dynamiques mises en exergue par les données spatio-temporelles sont alors généralement catégorisées en fonction de deux types (Saint-Marc, 2017) :

- le *mouvement* qui peut concerner des trajectoires d'objets mobiles, c'est-à-dire des déplacements d'individus ou d'objets dans l'espace, ou des flux quantitatifs. La figure 1.5 représente la propagation des pluies liées à l'ouragan Harvey entre le 26 et le 27 août 2017, sur les côtes texanes ;

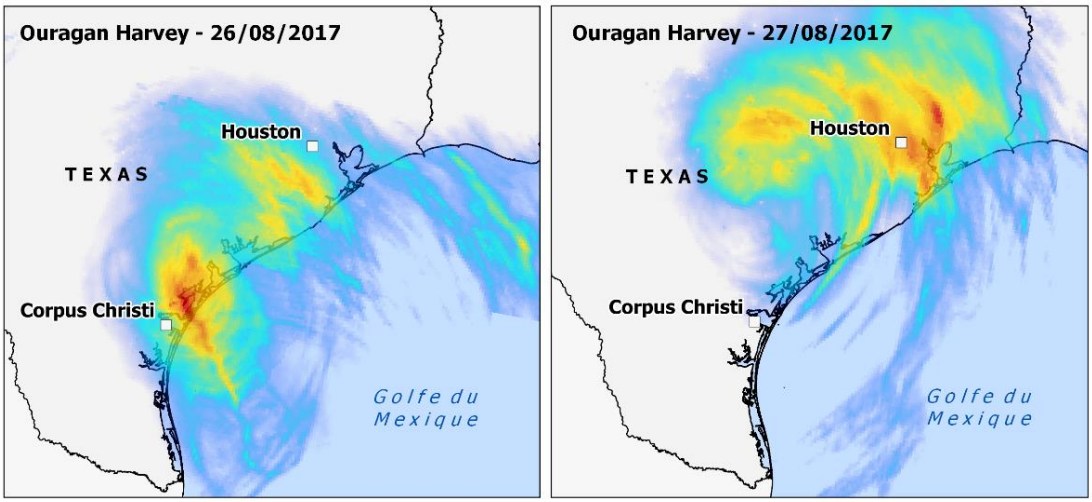


Figure 1.5 : Trajectoire d'un objet mobile dans le temps : l'ouragan Harvey au Texas (C.Cavalière)

- le *changement* qui caractérise une action directe du temps sur l'objet du territoire et inclut : l'apparition ou la disparition d'objets sur le territoire, le changement d'état d'un objet, ou encore le changement d'identité ou de forme d'un objet.

Risques naturels et usage des outils du numérique s'inscrivent ainsi dans les dimensions spatiale, thématique et temporelle de l'information géographique. Le tableau 1.1 suivant affiche différents thèmes capturés par l'information géographique dans ces deux champs et les classifie selon un type identifié :

Tableau 1.1 : Objets géographiques et attributs avec types d'informations associées

Champ	Objet géographique / attribut	Type d'information
Risques naturels	Perturbation atmosphérique / Crue / Inondation	Spatio-temporel et thématique : cartographie de la trajectoire d'un phénomène mobile et de la variabilité de son intensité
	Bâtiment détruit	Spatio-temporel : disparition d'un objet sur le territoire après la catastrophe
	Zone sinistrée déclarée non habitable	Spatio-temporel : changement d'état d'un objet après la catastrophe
Numérique	Le taux de foyers, par unité administrative, reliés au réseau très haut débit	Spatial et thématique : cartographie de la variabilité spatiale d'un phénomène technologique
	Le taux d'évolution, par unité administrative, du nombre d'individus équipés d'un smartphone, entre 2010 et 2018	Spatial et thématique incluant une dimension temporelle : cartographie de la variabilité spatiale de l'évolution d'un phénomène technologique (changement d'état)
	Individu équipé d'un GPS prenant une mesure de localisation toutes les trente minutes	Spatio-temporel : cartographie des trajectoires de l'individu dans l'espace

1.1.1.4. L'environnement de l'information géographique traditionnelle

Les jeux de données géographiques traditionnels, utilisés par les professionnels ou les étudiants en apprentissage des SIG, sont structurés et formalisés selon des normes précises, gages de leur qualité finale (Goodchild & Glennon, 2010). La production de la donnée est alors supervisée par un protocole rigoureux, les méthodes d'acquisition ainsi que la qualité des données sont documentées dans les métadonnées fournies simultanément : l'utilisateur peut alors évaluer la pertinence des données dont il dispose par rapport aux exigences de son projet (paragraphe 6 de la directive 2007/2/CE INSPIRE[11]). Ce matériau traditionnel et basique du géographe, qualifié de *scarce data* ou données éparses (Miller & Goodchild, 2015), expression caractérisant des données peu volumineuses[12] présente néanmoins trois inconvénients principaux :

- les processus d'acquisition et de formalisation sont codifiés, longs et coûteux ;
- la disponibilité instantanée de ces jeux de données est quasi inexistante ;
- leur mise à jour est irrégulière, dépendant grandement des budgets et du personnel disponible.

[11] Source : http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32007L0002&from=FR (Consulté pour la dernière fois le 14/03/2018)

[12] Par opposition aux Big Data, qui seront introduites dans la prochaine partie de ce chapitre et dont les volumes de données nécessitent des téraoctets d'espace de stockage, quand bien même certains jeux traditionnels, notamment des séries de données socio-démographiques ou des images satellites peuvent représenter quelques gigaoctets de données.

Ces jeux de données sont destinés à alimenter les bases de données géographiques des organismes institutionnels (Pornon, 2015) ; ceux-ci centralisent la collecte des données jugées utiles à la connaissance des territoires, quelle que soit l'échelle géographique des acteurs impliqués[13], par leurs propres agents ou par des organismes partenaires (Pornon, 2015) ainsi que leur diffusion *unilatérale*, via les géoportails de consultation ou les catalogues de données téléchargeables.

Tout jeu de données issu d'un organisme institutionnel s'accompagne d'une série de métadonnées qui correspondent, selon la formule consacrée, aux données décrivant les données[14]. L'organisation des métadonnées est généralement formalisée dans un fichier au format de page web, structurant les informations en différentes rubriques (Plumejeaud *et al.*, 2013). Les informations obligatoires comprennent la description du contenu de la qualité du jeu de données, le système de référence des données et leur résolution spatiale, la date de mise à jour des données ainsi que l'organisme responsable de leur production et de leur diffusion. Par exemple, les métadonnées fournies à l'utilisateur qui télécharge un jeu de données géographiques sur le portail d'un organisme officiel, par exemple la BD ADMIN EXPRESS® de l'IGN, contiennent en général un certain nombre de dossiers et de fichiers enregistrant les éléments suivants :

- un descriptif complet du jeu de données et de ses caractéristiques : méthode d'acquisition et de structuration des données, emprise spatiale, description des attributs, description des différents fichiers et de leur poids, date de livraison du produit, éléments sommaires d'évaluation de la qualité des données (actualisation des données, méthode de numérisation des données et sources de l'information sémantique) ;
- un fichier HTML[15] ou XML contenant les coordonnées de l'organisme producteur, le format ainsi que le système de référence spatiale des données, leur étendue géographique, leur résolution spatiale, l'encodage des caractères, la date de publication et la fréquence de mise à jour, les conditions et restrictions d'utilisation, etc.

[13] Source : https://www.ecologique-solidaire.gouv.fr/linformation-geographique (Consulté pour la dernière fois le 23/03/2018)
[14] Source : http://www.emse.fr/tice/uved/SIG/Glossaire/co/Metadonnees.html (Consulté pour la dernière fois le 14/03/2018)
[15] Le fichier HTML récapitulant l'ensemble des métadonnées relatives à la BD ADMIN EXPRESS de l'IGN est accessible via ce lien : http://professionnels.ign.fr/doc/m%C3%A9tadonn%C3%A9es%20de%20produit%20Unit%C3%A9s%20Administratives%20express.html (Consulté pour la dernière fois le 21/02/2019)

1.1.1.5. Des ambigüités persistantes à lever

Comme évoqué précédemment, les géographes ont coutume d'appréhender l'information géographique comme toute information localisée et décrite par ses attributs, et qui constitue la *matière première"*de la représentation cartographique (Lambert & Zanin, 2016). Cette définition aux limites non circonscrites soulève alors un problème crucial.

Si l'information géographique constitue la carte, qu'est-ce qui constitue l'information géographique ? En effet, comme le laissaient sous-entendre les paragraphes précédents, le concept d'information géographique tend à se confondre avec la définition de la donnée géographique (Béguin & Pumain, 2017). Pornon (2015) propose de définir la donnée géographique comme la représentation d'un objet du territoire localisé de manière explicite par une paire de coordonnées X et Y. La donnée géographique constitue ainsi un objet géographique géolocalisé, disposant d'une propriété géométrique (sa forme : ponctuelle, linéaire ou zonale) et d'informations descriptives (ses attributs) : en cela, les couches de données que l'on importe dans un logiciel SIG pour constituer un fond de carte et les stations de jaugeage géolocalisées évoquées précédemment disposent des mêmes composantes spatiales et descriptives, bien qu'elles soient radicalement différentes en termes de forme et de contenu. Pour lever ce flou entre la donnée et l'information, nous portons notre attention sur deux types de cartes du Web, consultables quotidiennement :

- *Les cartes de la pollution de l'air en Auvergne-Rhône-Alpes* : l'observatoire de la surveillance de la qualité de l'air de la région Auvergne-Rhône-Alpes publie une carte quotidienne de la prévision de l'indice de la qualité de l'air. Issu de calculs prenant en compte plusieurs polluants, cet indice correspond à une estimation du niveau global de pollution en tout point du territoire régional[16]. La figure 1.6 présente une carte de prévision des valeurs de cet indice de qualité de l'air pour la journée du 06/02/2019. Superposée au fond de carte, l'internaute visualise une matrice de pixels dont la valeur, identifiable par sa couleur, correspond à l'estimation de l'indice de pollution : lorsqu'elle est faible, l'indice tend vers des tonalités vertes ; lorsqu'elle est forte, il adopte des tons rouges.

[16] Source : https://www.atmo-auvergnerhonealpes.fr/article/indices-de-qualite-de-lair (Consulté pour la dernière fois le 06/02/2019)

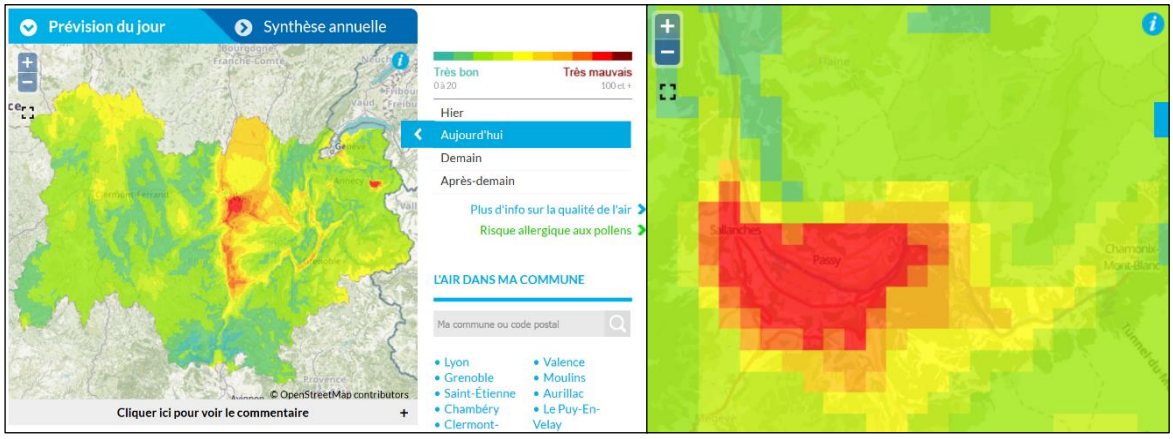


Figure 1.6 : Carte prévisionnelle de l'indice de pollution sur la région Auvergne-Rhône Alpes et la vallée de l'Arve, pour la journée du 06/02/2019 (ATMO Auvergne-Rhône Alpes)

Cette première carte restitue des données, sous forme de pixels, qui ont été *transformées* : l'internaute ne voit pas la donnée géographique *brute* qui est à la base de ce résultat. L'observatoire dispose d'un réseau de stations réparties sur le territoire régional : chaque station est alimentée d'un certain nombre de capteurs et d'analyseurs destinés à estimer le niveau de pollution d'un composé précis ; elle réalise donc une mesure *ponctuelle* de chaque polluant. La matrice de pixels diffusée au quotidien est donc le résultat d'une modélisation qui s'appuie sur les mesures des stations et les éléments de l'espace physique qui influencent l'émission et la dispersion des polluants (localisation des sources de polluants, topographie, conditions météorologiques) afin de proposer une couverture globale du territoire régional.

- *Les cartes de vigilance de Météo France* : Météo France émet et actualise chaque jour une carte de vigilance qui présente, pour neuf types de phénomènes d'origine météorologique (pluie, inondations, orages, vents violents, neige, verglas, avalanches, grand froid, canicule) un niveau de danger lié aux conditions météorologiques prévues, pour une durée de 24h[17]. Cette carte restitue une information traitée par un code de couleurs (soit quatre couleurs pour quatre niveaux de vigilance) et spatialisée à une échelle administrative générale, le département. Soulignons que le processus de construction de cette carte repose sur une transmission hiérarchique des informations à différents échelons territoriaux (d'un centre de prévisions départementales vers un centre de prévision interrégional avant validation par le centre national de prévisions de Météo France[18]) alors que le phénomène météorologique survient de manière plus ou moins intense sur un territoire dont les limites ne dépendent pas du maillage territorial administratif (figure 1.7).

[17] Source : http://vigilance.meteofrance.com/guide/vigilance.html (Consulté pour la dernière fois le 08/03/2018)
[18] Source : http://www.irma-grenoble.com/05documentation/04dossiers_articles.php?id_DTart=98&id_DT=11 (Consulté pour la dernière fois le 08/03/2018)

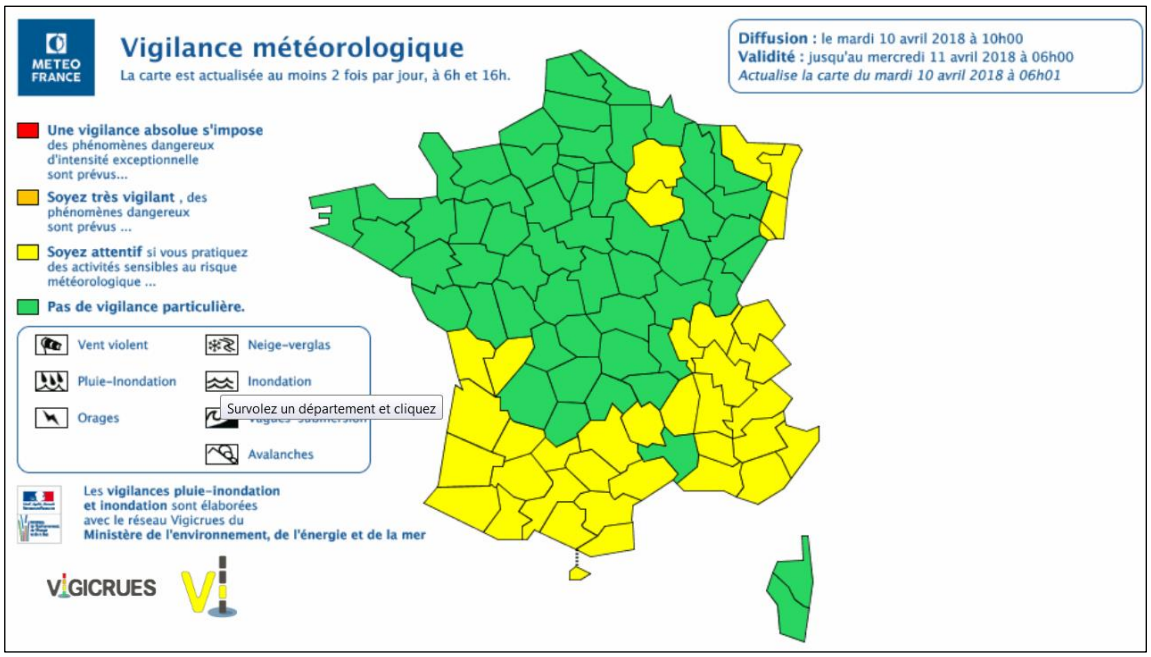


Figure 1.7 : Carte de vigilance météorologique pour la journée du 10/04/2018 (Vigilance Météo France)

La carte de vigilance transforme ainsi une information quantitative brute issue d'un réseau de capteurs de mesures relatives aux conditions environnementales et météorologiques particulières en une information qualitative ordonnée, les niveaux de vigilance, spatialisée à l'échelle des départements.

Comme pour l'exemple précédent, l'internaute qui consulte cette carte ne visualise pas la donnée brute ; de la même manière, le résultat ne lui donne pas accès aux données transformées des différents réseaux de capteurs, issues des modèles permettant aux météorologues d'estimer l'intensité et/ou la dangerosité d'un phénomène. En conséquence, cette carte ne restitue pas des données, mais une interprétation construite sur l'analyse de données (et il en était de même pour la carte du zonage des niveaux d'aléas, présentée par la figure 1.3) : cette interprétation est manifeste par la légende des cartes de vigilance : pas de chiffres ni de localisation précise mais des conseils sur les comportements à adopter.

Pour conclure sur ce premier point, nous envisageons la donnée géographique comme une donnée disposant d'une composante spatiale double (géolocalisation et forme) et d'attributs descriptifs et nous adopterons la terminologie suivante : la donnée géographique brute fait référence à la *première* donnée collectée et stockée, non modifiée par l'intervention du professionnel[19] ; la donnée géographique transformée provient d'une donnée brute ayant subi au moins une transformation (calcul, modélisation matricielle, agrégation, découpage, etc.). L'information géographique est localisée et décrite mais contrairement à la donnée, elle

[19] Bien qu'on puisse évidemment débattre sur ce point puisque les données de recensement par département correspondent à des données agrégées. Cela étant, nous conservons le qualificatif de brut car l'utilisateur des SIG, qui arrive "en fin de chaîne" n'a pas encore modifié les données de son propre chef.

ne peut être stockée dans une base de données sous forme de table (Lambert & Zanin, 2016). En outre, elle résulte d'un processus d'interprétation du constructeur de la carte, qui apporte du sens aux données et à la compréhension du phénomène spatial (Reix *et al*., 2011 ; Béguin & Pumain, 2017). L'information géographique fait donc référence, en premier lieu, à la capacité du constructeur de la carte à dégager des tendances spatiales perceptibles à partir de la distribution des données localisées (par exemple, Météo France décide de placer tel département à tel niveau de vigilance après une analyse combinée de diverses sources de données) ; dans un second temps, à l'interrogation des facteurs susceptibles d'être l'origine des structures perçues.

1.1.2. Acteurs et enjeux traditionnels des données et de l'information géographique

1.1.2.1. L'appropriation des territoires par des corps scientifiques et professionnels

La carte, définie comme une "*représentation géométrique conventionnelle, généralement plane, en position relative, de phénomènes concrets ou abstraits, localisables dans l'espace […]*" (Béguin & Pumain, 2017) dans les manuels académiques, reste pendant plus de vingt siècles de son histoire l'outil d'expression des données géographiques par lequel savants, topographes, géographes et ingénieurs communiquent leurs travaux. Production et représentation des données géographiques requièrent, quelles que soient les époques considérées, des connaissances théoriques auxquelles s'ajoute la maîtrise indispensable d'outils et de compétences techniques voire artistiques (Hofmann *et al*., 2012).

Bien que les premières traces de productions cartographiques connues remonteraient à la civilisation mésopotamienne et à la constitution de ce qu'on pourrait qualifier d'ancêtre du cadastre (Grataloup, 2011 ; Lambert & Zanin, 2016), le développement des sciences géographiques connaît son premier essor pendant l'antiquité grecque, ce qui correspond aux périodes de colonisation et de conquêtes des marges de l'*oecoumène*[20]. Chez les savants grecs, la cartographie s'impose d'ores-et-déjà comme une science fondée sur des méthodes rigoureuses et rationalisées, témoignant d'une préoccupation précoce en ce qui concerne les mesures, les distances et la représentation de l'espace terrestre connu (Grataloup, 2011). De la même manière, la cartographie de la Renaissance, qui amorce l'élargissement de la cartographie du monde connu, reste le fruit principal d'érudits parfois autodidactes, comme Mercator, et s'inscrit toujours dans une lignée scientifique de par le perfectionnement des

[20]Source : http://expositions.bnf.fr/globes/bornes/borne2.htm Exposition virtuelle de la BnF "Représenter la Terre" (Consulté pour la dernière fois le 11/02/2019)

instruments de mesure et le renforcement de la maîtrise technique et mathématique (Hofmann *et al*., 2012).

A partir du XVII^{ème} siècle, la cartographie s'institutionnalise : en France, sous l'Ancien Régime, l'intendance royale reconnaît la cartographie comme fonction d'Etat à part entière. On peut fréquemment rencontrer, dans les cartouches ornant les cartes des XVII^{ème} et XVIII^{ème} siècles, la mention *Ingenieur Geographe du Roy* qui distingue le corps professionnel créé par Colbert. A la fin du XVIII^{ème} siècle, la fabrique d'une carte est une opération qui exige le maniement d'instruments de précision, des connaissances mathématiques et géométriques ainsi que la collecte d'information sur le terrain. Arpenteurs et ingénieurs géographes, dont la formation professionnelle se normalise dès la fin du XVIII^{ème} siècle par la création d'écoles dédiées, participent à la collecte d'information et à la représentation cartographique des territoires (Hofmann *et al.*, 2012). Le positionnement de la cartographie comme discipline scientifique et champ de recherche institutionnalisé s'affirme au XX^{ème} siècle : les progrès de l'aviation consécutifs à la Première Guerre mondiale contribuent à généraliser l'utilisation de la photographie aérienne, permettant l'acquisition de données *à distance* (Hofmann *et al*., 2012). Dans la seconde moitié du XX^{ème} siècle, la constitution de réseaux de satellites artificiels mis en orbite autour du globe engage une révolution majeure dans la mesure de la Terre et la capture de l'information géographique (Béguin & Pumain, 2017). De par l'imagerie satellite, la mesure et la représentation de la Terre n'impliquent plus une présence humaine et physique sur l'espace étudié mais demeurent une tâche appelant des compétences professionnelles solides, ainsi qu'un certain coût financier (Hofmann *et al*., 2012).

En outre, la connaissance des outils de mesure du topographe et de l'arpenteur glisse rapidement vers la maîtrise de nouveaux outils informatiques professionnels, en l'occurrence les SIG et outils de télédétection, qui se substituent à la pratique du terrain et au papier. L'informatisation de la géographie, dès les années 1970, est un moteur de transformations des pratiques cartographiques (Guermond, 1991) : les logiciels de SIG permettent d'associer l'affichage et le traitement de données consignées dans divers formats : images *raster*, couches vectorielles, tables de base de données, etc. (Béguin & Pumain, 2017) ; en termes de traitement, ils intègrent des outils d'analyse statistique, facilitant ainsi les possibilités d'établissement de corrélations entre phénomènes et modèles de prévision (Lambert & Zanin, 2016) d'une part, et d'analyse et représentation des effets de distance et de discontinuités d'autre part (Béguin & Pumain, 2017).

La seconde partie du XX^{ème} siècle entame donc une rupture totale des pratiques cartographiques avec les siècles précédents : l'affichage automatique des données succède au dessin manuel, l'impression numérique ou l'écran à la gravure et à la presse à imprimer, la mesure effectuée à l'aide d'outils numériques ou à distance, aux instruments et calculs mathématiques et à la présence primordiale du géographe sur le terrain[21].

[21] Cette recherche doctorale a ainsi été menée en déconnexion complète du terrain réel.

1.1.2.2. Des méthodes graphiques rationnelles pour véhiculer un discours mis en valeur par la carte

Représenter les objets du territoire pour décider.

La production d'un savoir cartographique académique ou professionnel nécessite ainsi l'apprentissage de compétences techniques qui ont évolué au cours des siècles, de la triangulation à la base de données géoréférencées par GPS et aux SIG de nos jours. Pour autant, l'enjeu fondamental de la carte, au-delà de sa vocation à localiser une donnée sur un espace, qui perdure dans les représentations cartographiques professionnelles, c'est de communiquer un message (Béguin & Pumain, 2017). Cet exercice de communication entremêle, quelles que soient les époques considérées, besoins pragmatiques de connaissances des territoires et esthétique.

L'exposition virtuelle intitulée *Les enjeux de la cartographie* de la Bibliothèque nationale de France[22] (BnF) classe les enjeux de la cartographie en cinq thèmes : connaître, représenter, contrôler, agir et imaginer. En effet, depuis l'Antiquité, la cartographie s'ancre dans deux usages principaux rationnels, qui se sont perpétués dans nos sociétés : délimiter un territoire (d'abord en termes de propriétés de terres agricoles) et se déplacer. L'usage cadastral de la carte a subsisté puisqu'il constitue toujours un document de référence en termes de fiscalité foncière. La représentation des voies de communication, qu'elles soient terrestres ou maritimes, apparaît également comme un legs de l'Antiquité, qui connaît une première apogée avec les atlas portulans au XIV^{ème} siècle, et qui reste fondamentale dans l'ensemble des cartes distribuées auprès du grand public de nos jours, qu'elles soient routières ou topographiques.

Au XVII^{ème} siècle, la production et la cartographie de l'information géographique évoluent : la carte affirme son rôle comme outil de contrôle et d'administration du territoire. En conséquence, l'échelle des cartes change : la maîtrise du territoire par l'administration requiert un degré de connaissance fin (Hofmann *et al.*, 2012). Ainsi, du planisphère d'échelle globale, on progresse vers une cartographie du territoire national et multiscalaire. Elle introduit notamment la représentation de la topographie dans les cartes de Cassini dans la seconde moitié du XVIII^{ème} siècle. La figure 1.8 illustre les différents éléments du territoire que le cartographe Jean de Beins rend perceptible dans sa représentation cartographique du Dauphiné, à savoir les montagnes, le réseau hydrographique, les villes, villages et places fortifiées. Le graphisme est utilitaire : si la carte indique la localisation de différents éléments, elle permet également de les hiérarchiser sur le territoire. Les montagnes apparaissent d'autant plus grandes que leur altitude est élevée, une proportion entre le trait des cours d'eau et leur importance dans le réseau hydrographique est visible et enfin, dimensions et

[22] Source : http://expositions.bnf.fr/globes/bornes/borne4.htm Exposition virtuelle de la BnF "Les enjeux de la cartographie" (Consulté pour la dernière fois le 11/02/2019)

quantité d'édifices représentés différencient pôles urbains régionaux fortifiés et bourgs ruraux (Gal & Lazier, 2018).

Figure 1.8 : "Carte et description generale du Dauphiné avec les confins des Païs et provinces voisines le tout racourcy et réduicte par Jean de Beins ingénieur et géographe du Roy avec privilege de sa majesté", Gravure, ca. 1630 (Gal & Lazier, 2018)

Cette approche novatrice confère à la carte deux nouveaux usages qui persistent dans l'usage contemporain des cartes :

- les données géographiques inventoriées et localisées sur la carte deviennent un instrument de connaissance de l'espace, pour des administrateurs et décideurs qui ne sont pas nécessairement sur place (Gal & Lazier, 2018) ;
- la visualisation de l'espace construit et aménagé, dans le présent, constitue la base des connaissances indispensables à la projection de ce même espace dans l'avenir, soit à la prévision d'aménagements futurs qui répondent aux besoins des populations et aux contraintes exercées sur l'espace ; cette prévision peut également être exécutée à distance.

Une performance graphique destinée à communiquer.

Malgré sa dimension rationnelle et pragmatique, la construction de la carte mobilise un ensemble de méthodes graphiques (dessin, couleurs, formes, etc.) qui communiquent une certaine représentation, parfois orientée et mise en scène, de l'espace. Dès l'Antiquité, même si l'acquisition des savoirs géographiques se fonde sur des méthodes rationnelles, la

cartographie du monde demeure guidée par ses origines mythologiques (Grataloup, 2011). Ainsi, la représentation de l'*oecoumène* est toujours accompagnée de la représentation du monde céleste, usage qui persiste aux Temps Modernes.

Au cours du XIX^{ème} siècle, une transformation majeure des pratiques cartographiques coïncide avec l'apparition de la carte thématique consignant des données statistiques (Lambert & Zanin, 2016). Dès la première moitié du XIX^{ème} siècle s'engage une réflexion sur la forme de la carte : son esthétique graphique ne réside plus dans l'appel aux héritages antiques ou religieux mais dans les techniques visuelles mises en œuvre pour communiquer un message à des destinataires. Nous avons précédemment présenté la *Carte figurative de l'instruction populaire* publiée par Charles Dupin en 1826 (cf. figure 1.2), généralement appréhendée comme la première carte thématique de l'histoire de la cartographie (Lambert & Zanin, 2016). Son contemporain Charles-Joseph Minard a également amplement contribué au développement de la cartographie thématique, notamment par ses cartes de flux et cartes à diagrammes, toujours accompagnées d'une notice explicitant la lecture des informations. D'une part, cet ingénieur est parvenu à imposer définitivement la carte comme outil de connaissance du territoire et de décision[23] ; d'autre part, les géographes considèrent qu'il reste le pionnier des travaux de visualisation et de sémiologie graphique, dont Jacques Bertin formalisera les règles de communication un siècle plus tard, règles qui restent approuvées aujourd'hui.

La figure 1.9 présente ainsi la *Carte figurative et approximative représentant pour l'année 1858 les émigrants du globe* établie en 1862 par Minard, Régnier et Dourdet : le pays d'origine des émigrants est représenté par la couleur de la bande ; la largeur de cette bande varie en fonction du nombre d'émigrants recensés et ces mêmes valeurs sont inscrites au cœur des bandes colorées. Cette carte serait aujourd'hui qualifiée de carte de flux. La figure 1.10 soumet une représentation cartographique analogue mais contemporaine : les bandes de la figure 1.9 deviennent des flèches indiquant avec précision les pays d'accueil (alors que l'exactitude du fond cartographique apparaît secondaire sur la figure 1.9). La variation de la largeur de la flèche, proportionnelle à la valeur de la variable cartographiée, est une technique qui a persisté dans le temps, définie par Bertin comme une variable visuelle de taille.

[23] Source : https://visionscarto.net/charles-joseph-minard-cinquante-cartes (Consulté pour la dernière fois le 12/02/2019)

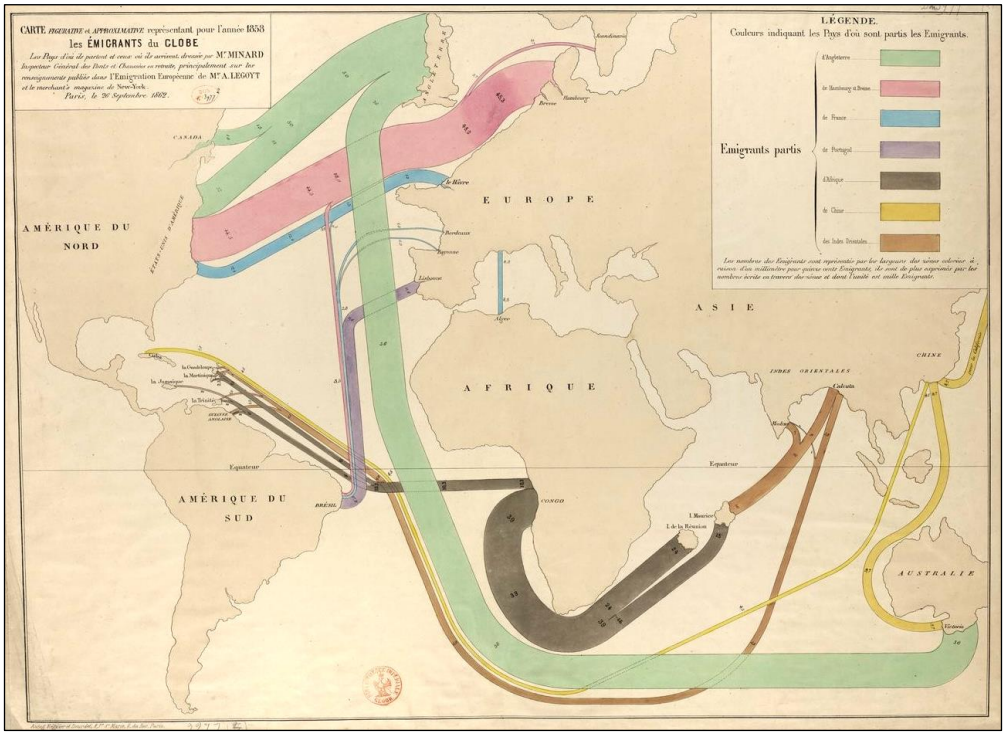


Figure 1.9 : Carte figurative et approximative représentant pour l'année 1858 les émigrants du globe. (Gallica)

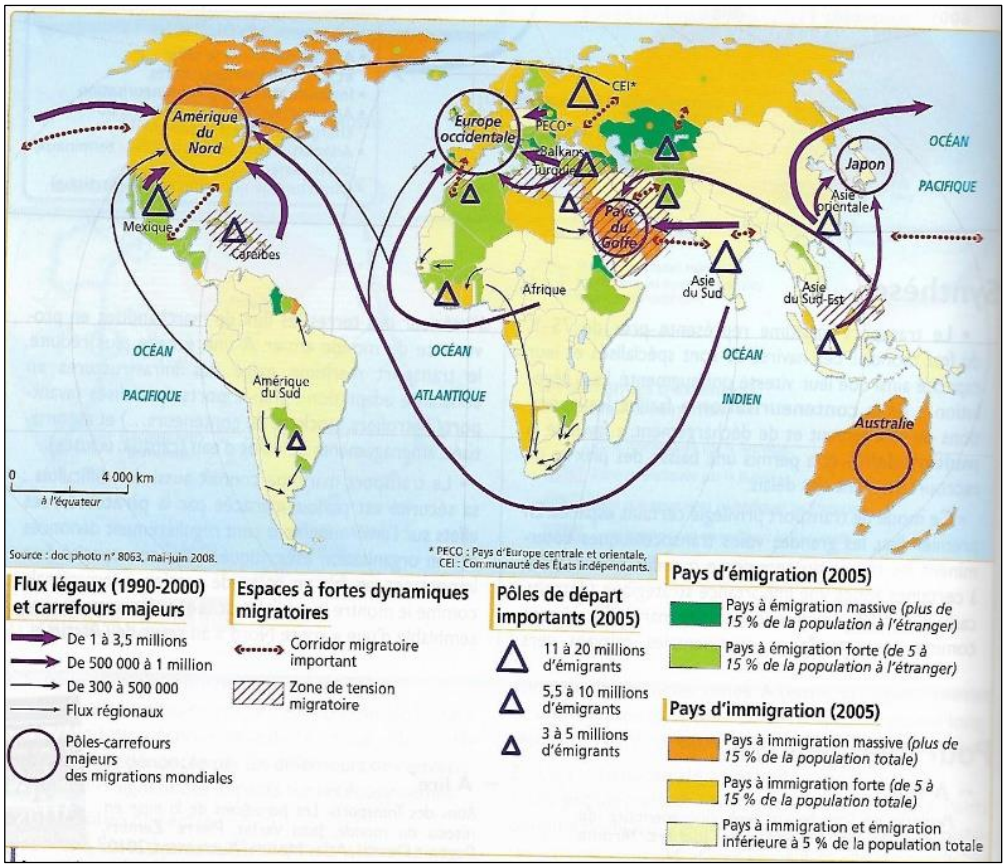


Figure 1.10 : Carte actuelle illustrant les dynamiques des populations à l'échelle mondiale, issue d'un manuel scolaire (Claude et al., 2013)

1.1.2.3. Les enjeux de l'introduction des outils numériques dans la cartographie thématique professionnelle

Aujourd'hui, la cartographie thématique reste le domaine de production majeur de la discipline : ses méthodes statistiques et visuelles demeurent l'un des piliers de l'enseignement (Béguin & Pumain, 2017). En outre, elle bénéficie d'un fort ancrage dans la société puisqu'elle est quasiment omniprésente dans les divers médias, qu'ils soient traditionnels (journaux, télévision) ou numériques (Lambert & Zanin, 2016). Au-delà de ces pratiques héritées du XIX^{ème} siècle, les enjeux de la cartographie professionnelle actuelle soulèvent un débat de fond et de forme :

- à l'heure où la construction d'une carte thématique peut être accomplie en quelques clics et que le numérique entraîne un foisonnement de ces cartes diffusées notamment sur le Web, il s'avère indispensable de discuter de la significativité d'une carte (Lambert & Zanin, 2016): une seule représentation cartographique n'est pas le reflet d'une vérité sur le terrain, d'autant plus lorsqu'intervient, dans le processus de construction de la carte, une étape de discrétisation qui a fréquemment tendance à être négligée alors qu'elle peut mettre en évidence des tendances spatiales radicalement différentes (Monmonier, 1996). La carte de la radicalisation islamique (figure 1.11) qui avait été publiée en 2016 au Journal du Dimanche (JDD) avait ainsi été l'objet de nombreuses critiques[24] : en premier lieu, elle ne respecte pas la règle de sémiologie logique de représentation de l'intensité des valeurs par un dégradé de couleurs. De plus, elle comporte une erreur statistique dans la mesure où elle affiche des valeurs absolues et non un taux d'individus radicalisés rapportés à une population.

[24] Source : https://www.lemonde.fr/les-decodeurs/article/2016/10/11/quelques-conseils-pour-reussir-vos-cartes-geographiques_5011945_4355770.html (Consulté pour la dernière fois le 12/02/2019)

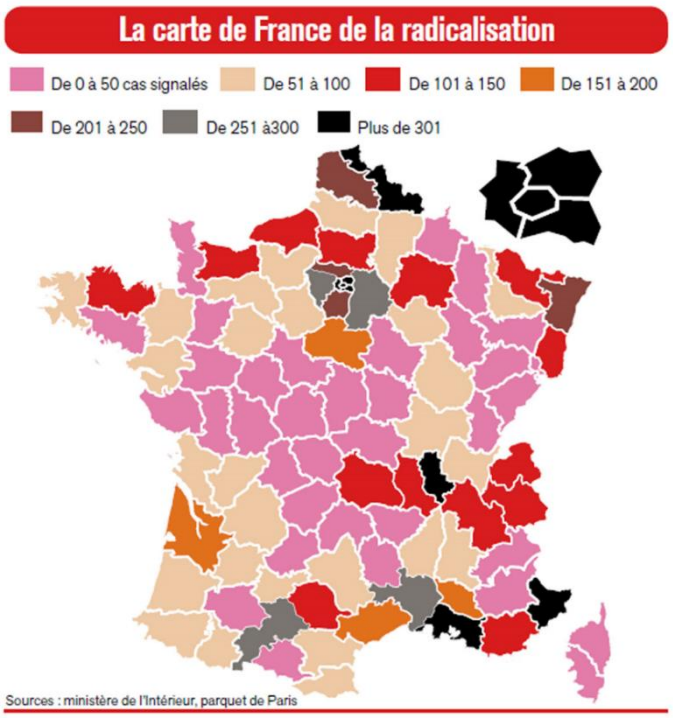


Figure 1.11 : Carte de la radicalisation islamique en 2016 (JDD)

- produire une carte implique désormais respect de règles statistiques et de sémiologie graphique d'une part, et d'autre part, la mise en scène, c'est-à-dire l'esthétique de la carte destinée à la rendre attrayante pour son destinataire[25] (Lambert & Zanin, 2016). La généralisation de l'utilisation des outils informatiques et du passage à l'écran dans la production des cartes décuple les possibilités d'amélioration de graphisme et de lecture des cartes : animation, interactivité, construction de cartes par anamorphose, recours à des styles graphiques variés, etc.

[25] Les questions conciliant mise en scène et sémiologie graphique ne sont pas pour autant l'apanage de l'intégration des outils numériques dans la cartographie : les cartes européennes de propagande ou d'instrumentalisation du sentiment national publiées entre 1870 et la Seconde Guerre mondiale emploient des graphismes particuliers afin d'imposer des représentations mentales, par toujours rationnelles des territoires auprès du grand public.

1.2. Les transformations des pratiques et acteurs des données géographiques à l'heure du numérique

Au XXème siècle, la carte s'enracine progressivement dans la société civile, se démocratise et devient un objet de consommation courante (Béguin & Pumain, 2017) : cartes routières, touristiques, plans de lignes ferroviaires, cartes postales, etc., constituent autant de représentations cartographiques destinées au grand public. Au début des années 2000, l'ancrage progressif de l'informatique et le récent accès au GPS dans la société civile donnent un nouveau souffle à la production de données géographiques et de cartes : de nouvelles formes de données et de nouveaux acteurs de la cartographie apparaissent.

1.2.1. Les mutations induites par les nouveaux outils numériques

1.2.1.1. Le Web social comme prélude à l'émergence de nouvelles pratiques

La démocratisation des ordinateurs et de l'accessibilité à une connexion Internet dans les foyers au début des années 2000 ne constituent pas des facteurs anodins en ce qui concerne le comportement des nouveaux utilisateurs, issus du grand public, vis-à-vis des différents types de contenu (dont les représentations cartographiques ou globes) disponibles sur le Web. En parallèle, dans cette même période, les technologies 2.0 et l'émergence du *Web social* font progressivement évoluer la posture de l'internaute. Il ne s'agit plus de "glaner" de l'information de page en page en cliquant sur des liens ; l'enjeu consiste alors à impliquer l'internaute dans la production d'un savoir partagé et ouvert à toute la communauté (Goodchild, 2007). Le *Web social* engage ainsi l'internaute dans une posture active : il consulte et enrichit le contenu existant. Par conséquent, les plateformes fondées sur les technologies du *Web 2.0* annoncent l'amorce d'une dynamique de participation active à la création du contenu Web, dans la période même où le numérique s'enracine dans le quotidien du grand public.

Le *Web social* (désormais remplacé par l'expression *réseaux* ou *médias sociaux*) s'appuie sur l'architecture alors nouvelle du *Web 2.0*, soit un ensemble d'outils, d'interfaces et de protocoles destinés d'une part, à faciliter les interactions entre les internautes et d'autre part, à rendre tout internaute, quel que soit son profil, acteur de la création du contenu mis à la disposition de toute la communauté d'internautes[26]. (Kaplan et Haenlein, 2010) proposent de définir les médias sociaux comme *"un ensemble d'outils en ligne construits sur les bases idéologiques et technologiques du Web 2.0, qui permet la création et l'échange d'un contenu généré par un utilisateur"*. Ces outils se diffusent sous plusieurs formes : l'exemple type du principe de fonctionnement collaboratif du *Web social*, souvent cité dans la littérature reste

[26] Source : https://en.wikipedia.org/wiki/Social_web (Consulté pour la dernière fois le 15/02/2019)

l'encyclopédie *Wikipédia*, lancée en 2001 (Goodchild, 2007). *Wikipédia* est un *wiki*, c'est-à-dire une interface conçue pour favoriser la collaboration et les interactions entre un groupe d'internautes acteurs, autorisés à accéder à la plateforme afin d'enrichir son contenu (créer, illustrer ou modifier une page hébergée par un site web). Un tel site constitue donc à la fois un espace de *co-construction* de connaissances et un espace *ressource* pour tout internaute qui recherche de l'information. La participation au *Web social* prend également d'autres formes : le blog, et avec l'adoption des smartphones, le partage massif et quasi continu de divers supports multimédias (Pirolli et Crétin-Pirolli, 2011).

L'approche de l'information par le *Web social* se distingue des médias et sources d'informations traditionnelles : la logique 2.0 est avant tout inscrite dans le partage et la collaboration entre des internautes acteurs du contenu (Pirolli et Crétin-Pirolli, 2011), et affranchie de toute approbation d'une institution. Les technologies et usages du *Web social* bouleversent l'ancienne logique *top-down* à sens unique (Goodchild, 2007 ; McDougall, 2012) : les masses d'individus connectés supplantent les acteurs traditionnels de la création de données ; ils sont devenus le moteur de la création et de la diffusion d'informations sur le Web. En outre, ils s'inscrivent dans un système de dialogue complexe (qui peut être à la fois horizontal et vertical) et dont les émetteurs et destinataires sont souvent multiples et difficilement identifiables (Pirolli et Crétin-Pirolli, 2011).

1.2.1.2. Le Web social devient géographique : néogéographie et Géoweb

La proximité entre carte et grand public n'est pas un phénomène récent, conséquent au numérique : dès la seconde moitié du XX^{ème} siècle, la carte s'affirme comme instrument d'usage quotidien ou de loisirs, notamment au travers des plans de transports en commun, des cartes météorologiques, des cartes de randonnée ou routières, et plus récemment, même si elle représente un territoire fictif, par sa fréquente présence dans les jeux vidéos.

Par le numérique, le grand public ne découvre pas la carte mais s'adapte et adopte les nouveaux supports qui transforment et facilitent l'accès à la carte et aux données géographiques. Au début des années 2000, conséquence de la nouvelle architecture du *Web 2.0*, apparaît une nouvelle pratique qualifiée de *néogéographie* : elle désigne l'ensemble des techniques et outils du *Web 2.0* qui permettent aux internautes de créer leurs propres cartes avec leur propres données localisées[27]. Mais l'acte considéré comme fondateur du Géoweb correspond à la naissance de *Google Maps* en 2005, soit d'une cartographie d'échelle mondiale englobant carte de type routier et imagerie satellite proposant des fonctions de localisation, de recherche d'itinéraire et un ensemble de services Web[28]. Le Géoweb se définit alors comme une tendance à l'indexation géographique systématique du contenu mis en ligne

[27] Source : https://www.slideshare.net/renalid/la-cartographie-sur-internet-de-la-nogographie-au-goweb (Consulté pour la dernière fois le 15/02/2019)
[28] Source : https://mondegeonumerique.wordpress.com/2010/06/24/le-geoweb-pour-les-nuls/ (Consulté pour la dernière fois le 15/02/2019)

sur les plateformes du *Web 2.0* par géoréférencement des données sur la surface terrestre (Joliveau, 2011).

On peut subdiviser le recoupement entre Géoweb et pratiques néogéographiques selon deux logiques[29] :

- une logique traditionnelle verticale : le Géoweb de Google ou des différents portails qui proposent d'afficher et de naviguer sur des couches d'information géographique relatives à des thèmes diversifiés, comme le Géoportail de l'IGN, s'inscrivent dans la logique héritée *top-down*. Ils peuvent aussi bien être associés à une utilisation privée (un internaute consulte le cadastre numérisé sur un portail gouvernemental) que professionnelle dans la mesure où ils fournissent des services (Joliveau, 2011) : un professionnel peut ainsi intégrer une application cartographique basée sur Google Maps au site web de son entreprise.
- une logique horizontale et contributive, inscrite dans la lignée du *Web social* et dont l'exemple le plus fréquemment cité est la cartographie collaborative *OpenStreetMap* (Goodchild, 2007 ; Joliveau, 2011), qui a pour objectif de reconstituer, par les contributions de bénévoles, une carte du monde libre de droits. Dans ce cas, la pratique du Géoweb se veut également sociale dans la mesure où elle assigne une posture active à l'internaute : celui-ci n'est pas qu'un simple spectateur du contenu web qui navigue sur la carte et soumet des requêtes au serveur, il est acteur de la création et du géoréférencement d'un contenu géolocalisé (Joliveau *et al.*, 2013).

Dans tous les cas, le Géoweb social se démarque de l'usage professionnel des SIG : les cartes du Géoweb assurent à tout individu disposant d'un ordinateur (ou smartphone) et d'une connection Internet, un accès à des fonctionnalités jusqu'alors monopole de corps professionnels, centrées sur la possibilité d'ajouter des données personnelles et de réaliser des cartes en fonction de ses centres d'intérêts (Joliveau *et al.*, 2013). En conséquence, l'avènement du Géoweb amorce l'ouverture et la diversification des acteurs et du rôle des données géographiques et de la cartographie[30].

1.2.1.3. L'adoption des dispositifs mobiles de géolocalisation : un bond dans la production des données géographiques

Si tout un chacun aura certainement au moins une fois, pour ses besoins particuliers ou professionnels, recherché un itinéraire sur une plateforme cartographique ou une adresse particulière sur une carte collaborative, seuls certains individus consacreront de leur temps libre à devenir des contributeurs *OpenStreetMap* réguliers (ou à partager tout autre type d'information sur un blog thématique). L'outil qui bouleverse les pratiques de production de données géographiques et qui constitue l'origine la plus vraisemblable de la situation présente

[29] ibid.
[30] Objet du paragraphe 1.2.3 de ce chapitre.

d'explosion du volume des bases de données spatialisées produites par des applications (Miller, 2007 ; Joliveau, 2011), est le *dispositif mobile* qui *suit* l'utilisateur dans son quotidien et *capture* des données spatialisées. Ces dispositifs, qualifiés de *location-aware technologies* (Miller, 2007), abrégées par l'acronyme de *LATs*, désignent tout appareil capable de se géolocaliser, c'est-à-dire de détecter automatiquement et en temps réel la position géographique d'un individu ou d'un objet (Joliveau, 2011). Les LATs ont recours à trois principes de géolocalisation principaux qui offrent la possibilité d'une émission régulière et donc d'un suivi précis de l'appareil : le signal GPS, la radiolocation par bornes WiFi et les systèmes de radiofréquence RFID (Joliveau, 2011).

Rapidement, ces technologies numériques sont intégrées dans les smartphones, qui embarquent un capteur GPS, et sont adoptés massivement par la société ; ils deviennent ainsi de véritables dispositifs mobiles de suivi de leurs utilisateurs. En parallèle, les plateformes du Géoweb, initialement conçues pour des ordinateurs, développent leurs propres applications mobiles dont certaines fonctionnalités s'exécutent par la géolocalisation (Joliveau, 2011) : le controversé *Facebook Nearby Friends*[31], l'ajout d'une localisation aux tweets mais aussi tout GPS de navigation utilisé sur la route. Ainsi, l'adoption générale des LATs et des applications mobiles associées par la société est à l'origine :

- de bases de données spatialisées qui représentent des téraoctets de volumes (Joliveau, 2011), dont une partie est mise à disposition d'un public averti, mais qui restent aux mains des géants du Web (Joliveau, 2011 ; Quesnot, 2016) ;
- du développement des *location-based services*, soit des services qui fournissent à l'utilisateur des informations ciblées proposées en fonction de la géolocalisation de l'appareil (Miller, 2007).

1.2.2. Typologie des pratiques de création de données géographiques autour du Géoweb

1.2.2.1. L'Open Data

La première philosophie ancrée dans l'esprit du Géoweb de logique horizontale correspond à l'émergence d'une posture générale prônant l'ouverture des données, en réponse à un besoin de transparence émanant des sociétés. L'*Open Data* désigne ainsi, de la part des institutions productrices de données, une politique de partage de données fondée sur deux axes[32] :

[31] Facebook Nearby Friends est une application basée sur la géolocalisation : elle avertit un utilisateur connecté lorsqu'un de ses "amis" se trouve à proximité et affiche sa position exacte sur une carte. Source : https://www.facebook.com/help/android-app/291236034364603?helpref=uf_permalink

[32] Source : https://www.gouvernement.fr/action/l-ouverture-des-donnees-publiques (Consulté pour la dernière fois le 18/02/2019)

- une mise à disposition gratuite, pour l'ensemble des acteurs, qu'ils soient professionnels ou issus de la société civile, des jeux de données que les instituts produisent dans le cadre de leur mission de service public ;

- cette mise à disposition doit s'effectuer dans des formats ouverts afin de faciliter la réutilisation des données.

Sur le Géoweb, l'*Open Data* se traduit par la diversité de portails cataloguant des données statistiques spatialisées ou des couches d'information géographique directement exploitables : aux Etats-Unis, comme en France, chaque échelon administratif dispose de son propre portail de données, du pays aux différentes villes ; il en est de même pour les instituts environnementaux et d'étude statistique de la population. Le tableau 1.2 ci-dessous propose quelques exemples de portails de données américains et des types de données téléchargeables par tout internaute :

Tableau 1.2 : Exemple de portails américains cataloguant des données ouvertes

Institut / Echelon Géographique	Lien	Formats de données
Data.gov (portail du gouvernement fédéral américain)	https://www.data.gov/	Divisé en thèmes comme son équivalent français. Fichiers aux formats CSV, XLS, SHP, XML ou simples rapports d'étude en format PDF
City of Houston	http://data.houstontx.gov/	Egalement divisé en thèmes (services, limites administratives, santé, environnement) Fichiers aux formats CSV, XLS, SHP, XML ou simples rapports d'étude en format PDF
Census Bureau	https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml	Jeux de données CSV relatifs à la population et aux conditions de vie (structure, habitat, revenus, santé, usage des TIC, moyens de transports, etc.). Inclut un outil de cartographie en ligne
NOAA	https://data.noaa.gov/	Jeux de données météorologiques, hydrologiques ou océanographiques dans des formats XML, TIFF, XYZ

1.2.2.2. La Volunteered Geographic Information (VGI) et le Crowdsourcing

La Volunteered Geographic Information.

En 2007, Goodchild introduit le concept de *Volunteered Geographic Information* (abrégée par l'acronyme VGI) pour désigner une pratique émergente initiée par la rencontre

entre les technologies du Web 2.0 et le grand public, pratique par laquelle certains internautes se positionnent comme nouvelle "source" de production de données géographiques. La VGI est ainsi définie comme *l'implication d'un grand nombre d'individus, qui, dans la sphère de leurs activités privées, créent de l'information géographique, qu'ils aient peu ou pas de qualification dans ce champ disciplinaire* (Goodchild, 2007) ; elle s'inscrit donc dans la lignée des pratiques initiées par le Géoweb 2.0, dans la mesure où elle transfère des compétences et prérogatives traditionnellement restreintes à des cercles professionnels, auprès d'un public amateur.

Les chercheurs appréhendent la VGI comme une donnée à composantes multiples, illustrées dans le tableau 1.3 ci-dessous (Capineri, 2016). Contrairement à une donnée géographique formelle issue d'un producteur officiel, la VGI a pour caractéristique principale d'être une donnée protéiforme plus ou moins précise, aussi bien dans son contenu que dans son référencement géographique :

Tableau 1.3 : Les composantes de la VGI

Composante	Rôle	Formes
Spatiale	Localiser la donnée	Coordonnées GPS, Géotag (marqueur textuel à caractère géographique), nom de lieu
Contenu	Support de la donnée, par lequel on peut construire de l'information	Le contenu est protéiforme : texte, image, vidéo, check-in (pointage qui signifie la présence de l'individu en un lieu particulier)
Attributs	Ils correspondent en fait à des métadonnées	Horodatage de la création de la donnée, numéro d'identifiant de l'utilisateur, langue identifiée, etc.

Par ailleurs, nous considérons ici que la VGI se distingue d'autres types de contenus spatialisés générés par les utilisateurs des plateformes du *Web social* - et principalement des applications comme Twitter ou BlaBlacar - de par son caractère *délibéré* qui, dans les pratiques actuelles de production de données géographiques 2.0, ne s'avère pas toujours identifiable (Capineri, 2016) : en effet, la création d'une donnée qualifiée de VGI est un acte contributif[33], c'est-à-dire que tout internaute qui crée une donnée sur une plateforme est *conscient* que sa contribution est visible par l'ensemble des utilisateurs du Web et peut être modifiée par les

[33] Le qualificatif de participation demeurant sujet à débat dans la communauté, nous employons le terme de contribution. Le terme de participation a en effet été introduit, dans un contexte officiel, pour désigner l'intégration des connaissances et expériences du public aux processus de décision dans l'aménagement du territoire. Or, les pratiques géographiques 2.0 s'affranchissent de tout cadre institutionnel ; c'est pourquoi le terme de contribution paraît plus approprié pour qualifier l'implication d'individus à la création de données libres (Joliveau, 2013).

autres contributeurs de la plateforme. Parmi les motivations qui engendrent la création de la VGI, on peut donc discerner d'une part, une *intention* d'informer un public d'internautes et, d'autre part, une *conscience* de la réutilisation de la donnée créée par un tiers. La figure 1.12 affiche un exemple de saisie de données VGI de la plateforme *Harveyneeds*, développée par la communauté ouverte Sketch City à Houston[34], et destinée à collecter et diffuser les données relatives à l'aide d'urgence pendant l'ouragan Harvey en 2017[35]. Sur cette plateforme, le contributeur peut notamment géoréférencer les refuges : il dispose de consignes précises afin d'orienter la saisie des nouvelles données, affichées sur une cartographie ouverte et publique, intitulée *Houston Shelter Map* :

[34] Source : http://sketchcity.org/ (Consulté pour la dernière fois le 18/02/2019)
[35] Source : http://harveyneeds.org/#maps (Consulté pour la dernière fois le 18/02/2019) Attention, l'accès à la cartographie globale et à la base de données n'est à ce jour plus possible.

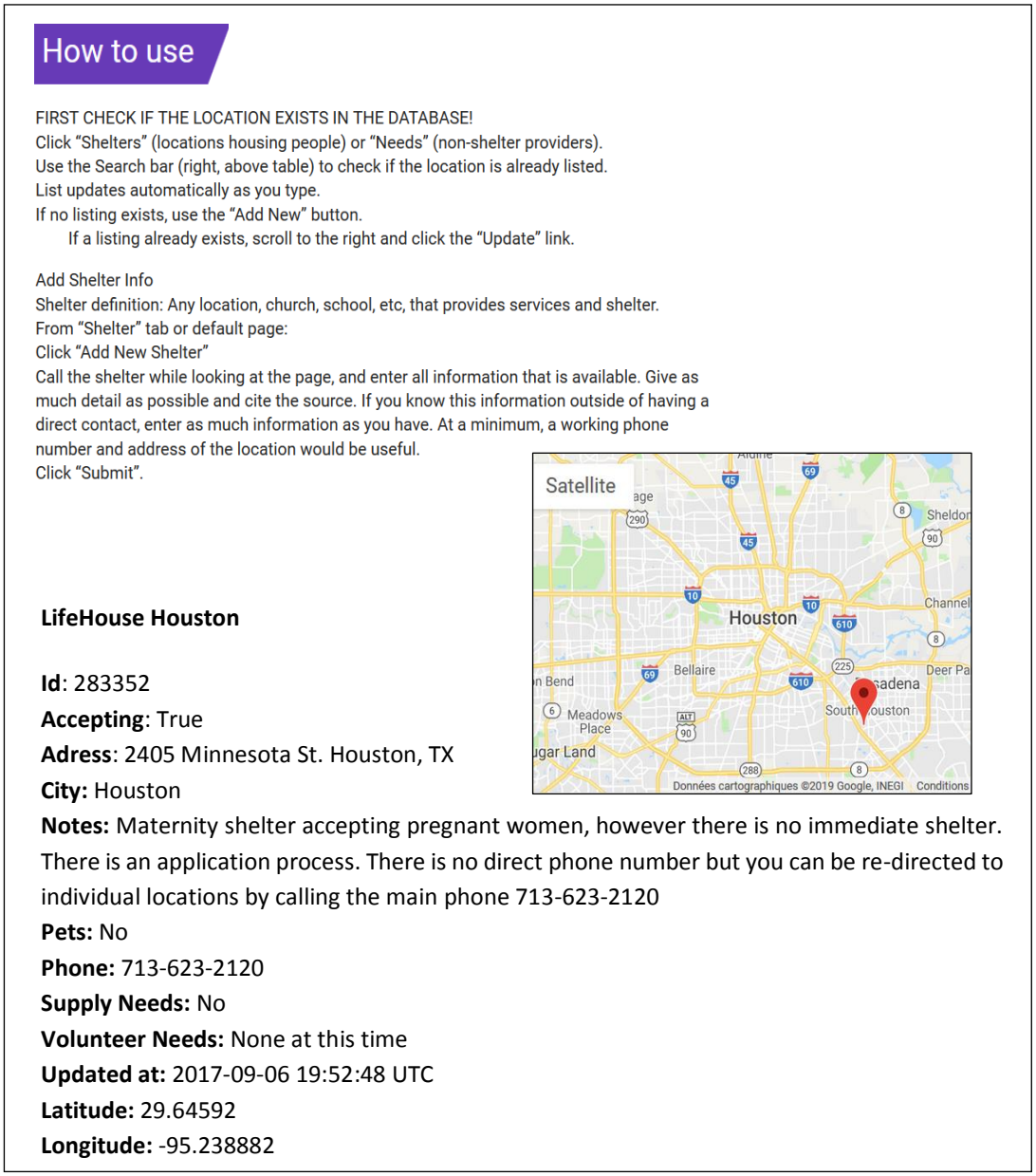


Figure 1.12 : Saisie de données contributives - géolocalisation et attributs - relatives à la localisation de refuges en période de crise (Sketch city , Houston recovers)

Le crowdsourcing.

Merriam Webster définit en 2005 le *crowdsourcing* comme le processus d'acquisition et de traitement d'informations de manière collective. Il fait généralement appel à la contribution pour la réalisation de tâches en sous-traitance, via la sollicitation d'un groupe important de contributeurs, en général des communautés d'utilisateurs des outils du *Web social*. Le *crowdsourcing* est ainsi souvent utilisé pour subdiviser des tâches fastidieuses entre plusieurs types de participants : volontaire ou employé, chaque contributeur traite, sur son initiative, une petite partie de l'information qui s'ajoute à toutes les informations traitées par

d'autres contributeurs[36]. Au premier abord, le *crowdsourcing* ne semble pas se démarquer de la VGI ; dans les pratiques, on peut distinguer deux différences fondamentales : un ensemble de contributeurs de *crowdsourcing* accomplit des *microtâches*, définies comme des tâches simples et faciles, de différents types, qui peuvent être exécutées en un laps de temps très court[37] (alors que la création d'une donnée relative à un refuge cité plus haut va certainement requérir un peu plus de temps : il faudra sans doute vérifier les coordonnées et contacter le personnel du refuge pour compléter les attributs descriptifs). Le contributeur de la microtâche n'est pas un producteur de données mais un individu qui apporte une primo-analyse de données collectées. En outre, il n'a pas besoin de se déplacer sur le terrain pour vérifier l'exactitude d'une donnée ou accomplir une microtâche. La plateforme académique *Crowd4U* (figure 1.13) se présente comme une interface permettant à tout internaute d'accomplir autant de *microtâches* qu'il le souhaite : dans cet exemple, une microtâche consiste à observer l'image aérienne et indiquer l'état du bâtiment marqué d'une croix rouge, les réponses potentielles étant déjà fournies par des boutons dédiés.

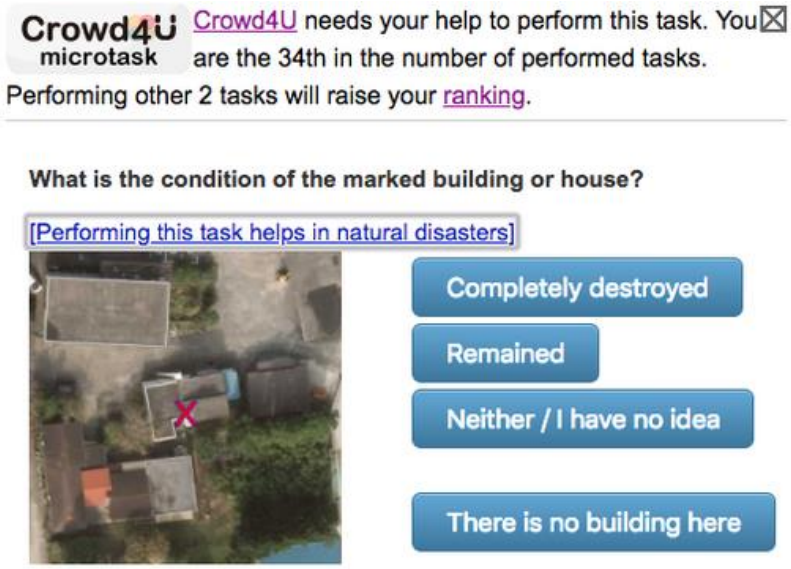


Figure 1.13 : Exemple de microtâche exécutée par un contributeur (Crowd4U)

VGI et *crowdsourcing* reposent sur l'hypothèse selon laquelle un groupe de personnes peut s'avérer plus efficace qu'un expert pour résoudre un problème (Goodchild & Glennon, 2010), dans la mesure où la démocratisation des outils numériques tend à réduire le fossé entre experts initialement seuls à maîtriser les TIC et le grand public, désormais tout autant consommateur de ces TIC[38]. Ces pratiques impliquent alors deux postulats : d'une part,

[36] Source : https://en.wikipedia.org/wiki/Crowdsourcing#In_geography (Consulté pour la dernière fois le 18/02/2019)
[37] Source : https://crowd4u.org/en/ (Consulté pour la dernière fois le 18/02/2019)
[38] Théorie de Jeff Howe, énoncée dans l'article "The rise of crowdsourcing", publié sur le site du magasine *Wired* en 2006.

l'information créée par un groupe composé de multiples observateurs reflète davantage la réalité qu'une information créée par une seule personne (Goodchild & Glennon, 2010). D'autre part, l'information collectée et diffusée par des contributeurs qui coopèrent et ont un intérêt commun dans le domaine en question est plus précise et plus fiable qu'une information provenant d'une minorité extérieure (Dashti *et al.*, 2014).

1.2.2.3. Le problème des traces numériques

Le Centre National de Ressources Textuelles et Lexicales définit la *trace* comme la *suite d'empreintes, de marques laissées par le passage d'un individu, d'un animal, d'un véhicule […]* qui entre autres, permettent de le *suivre en se guidant sur ce qu'il a laissé derrière lui*[39]. Dans le monde numérique, le recours à ce substantif conserve sa signification originelle : en effet, bien que la définition des traces numériques ne fasse pas encore l'objet d'un consensus dans la communauté scientifique (Galinon-Ménélec & Zlitni, 2013), nous pourrons accepter la définition proposée par Laflaquière en 2009 : *"une trace numérique peut être vue comme un ensemble d'enregistrements de données dont l'existence est provoquée par des interactions utilisateur dans le cadre de la réalisation de son activité instrumentée".* Dans leurs conditions actuelles d'utilisation, les *cookies* sont devenus des traceurs très répandus sur le web : ils collectent des données sur un internaute pendant sa navigation, lesquelles sont ensuite traitées par des algorithmes afin de faire un ciblage, c'est-à-dire de proposer un contenu susceptible d'intéresser l'internaute en fonction des centres d'intérêts identifiés à partir de ses traces (ce qui correspond à la rubrique "*recommandé pour vous*" des sites de vente en ligne ou des réseaux sociaux). Conséquence de la diversité et de la généralisation des dispositifs capables de détecter la position d'un utilisateur, les traces numériques produites à l'occasion de l'utilisation d'outils et d'applications diverses ont désormais tendance à être systématiquement géolocalisées (Mericskay *et al.*, 2018).

A l'instar des données produites ou validées par VGI ou *crowdsourcing*, la trace numérique se révèle hétérogène, dans son contenu comme dans la nature de sa géolocalisation. Le tableau 1.4 propose quelques exemples de traces numériques à dimension spatiale laissées plus ou moins consciemment sur divers types de réseaux :

[39] Source : http://www.cnrtl.fr/definition/trace (Consulté pour la dernière fois le 18/02/2019)

Tableau 1.4 : Quelques exemples de traces numériques hétérogènes impliquant une dimension spatiale

Exemple de trace numérique	Type de géolocalisation, contenu et enjeux
Carte Bancaire (achat en présence physique)	Localisation textuelle : nom du magasin et nom de la ville Attributs descriptifs : montant de la transaction, date Acceptation tacite de la trace par facilité de paiment
Tweet géolocalisé	Localisation précise : coordonnées GPS Attributs descriptifs : contenu (texte, image, lien), horodatage, numéro d'identifiant Acte volontaire
Blablacar	Localisation textuelle : adresse du lieu de départ / adresse du lieu de destination Attributs descriptifs : horodatage, type de véhicule, nombre de passagers, villes desservies, etc. Trace involontaire ou acceptation tacite par besoin
StreetPass des consoles Nintendo 3DS	Flou : l'entreprise présente la fonction comme un moyen d'échange automatique de données d'une console à une autre, dans un rayon de 100 mètres. Aucune information précise ne filtre quant à l'enregistrement éventuel de ces traces au-delà des consoles des utilisateurs. Les seules données auxquelles le propriétaire de la console a accès concernent les temps de jeu et le nombre de pas effectués pendant que la console est en veille au quotidien.
Trace GPX de randonnée postée sur Camptocamp	Localisation précise : *waypoint* acquis par GPS Mise en ligne volontaire et dans un esprit collaboratif : le contributeur à l'origine de la trace a conscience de sa visibilité et de sa réutilisation par les autres internautes.

Comme le soulignait le tableau 1.4, le positionnement de la trace numérique s'avère problématique : si la trace GPX porte ce nom de *trace* dans la mesure où elle restitue l'itinéraire du randonneur, elle reste néanmoins le seul exemple cité dans le tableau dont la création est délibérée et assumée par l'individu qui en est l'origine (et en cela, le comportement collaboratif du randonneur est assimilable à de la VGI). Les autres exemples mentionnés se démarquent de cette logique de par le contexte même de leur acquisition et de leur diffusion : le paiement par carte est bien commode, de même que le service offert par BlaBlacar et quel joueur se soucie du contenu des données échangées par sa console ? Ainsi, dans la plupart des cas, la capture d'une trace ne traduit pas nécessairement une intention contributive car elle est fréquemment acquise aux dépens de l'utilisateur (Pélissier, 2015 ; Quesnot, 2016). Ce contexte de capture de la donnée s'avère en fait complexe à saisir : si une partie de ces traces géolocalisées sont acquises indépendamment de la volonté de l'internaute (Mericskay *et al.*, 2018), comme lors de l'utilisation du GPS routier via *Google Maps*, d'autres sont la conséquence directe d'actions de leur part. Un individu peut accepter de participer à une expérience pendant laquelle il conserve une puce GPS qui enregistrera l'ensemble de ses déplacements dans une période donnée. Si un utilisateur de Twitter géolocalise ses tweets, c'est qu'il a lui-même activé la fonction de géolocalisation dans les paramètres de son compte.

A l'heure actuelle, il est difficile de savoir si ce comportement peut être associé à une attitude fataliste des internautes, résultant en une acceptation tacite vis-à-vis de la capture

des traces, ou s'il s'agit d'un manque d'information et donc une inconscience des problématiques éthiques associées à la création de ces traces.

On pourra également mettre en évidence une seconde différence fondamentale : un contributeur d'une plateforme 2.0 de VGI renseigne les objets du territoire : par exemple, dans *OpenStreetMap*, un contributeur peut numériser des données et éditer des attributs[40]. Il renseigne ainsi les objets du territoire (de la même manière qu'un contributeur ajoutant des données relatives à un refuge, évoqué dans la figure 1.13). Les traces numériques géolocalisées, dont le dispositif de capture est différent, ne renseignent pas tant sur les objets du territoire que sur les comportements des individus connectés sur le territoire : autrement dit, elles renseignent les pratiques spatiales des individus. En conséquence, nous considérons ici la trace numérique géolocalisée comme un témoin (ou marqueur) qui atteste de la présence physique d'un individu connecté en un point précis du territoire, à un moment connu. Elle témoigne ainsi d'une interaction entre le lieu réel et l'individu, dont nous ne connaissons pas le contexte précis, qui est sauvegardée au travers de ce témoin virtuel.

1.2.3. Le renouveau géographique du Géoweb

1.2.3.1. Les nouveaux usages consécutifs à l'ouverture des acteurs des données et de la carte

"La carte n'est plus du seul ressort des experts cartographes et la diversification des acteurs impliqués révèle la fin des grands récits cartographiques" (Noucher, 2017). Les outils du Géoweb déclenchent une rupture radicale avec les codes et pratiques séculaires d'une discipline historiquement monopole de savants et de corps professionnels. Dans l'ère de la transparence numérique, tout un chacun bénéficie d'un accès à la donnée géographique et peut ainsi devenir acteur de la production de la carte. Le Géoweb foisonne ainsi de deux types de produits. D'un côté, on trouve les nombreux portails d'organismes institutionnels qui offrent des interfaces cartographiques permettant à tout internaute, moyennant quelques clics, de visualiser des couches de données sélectionnées au choix, relatives à un thème particulier : sur la figure 1.14, l'utilisateur a choisi d'afficher les données relatives aux stations de jaugeage qui jalonnent les cours d'eau ainsi que les capteurs qui surveillent la hauteur d'eau des lacs au Texas.

[40] Source : https://wiki.openstreetmap.org/wiki/FR:Contribuer_aux_donn%C3%A9es_cartographiques (Consulté pour la dernière fois le 19/02/2019)

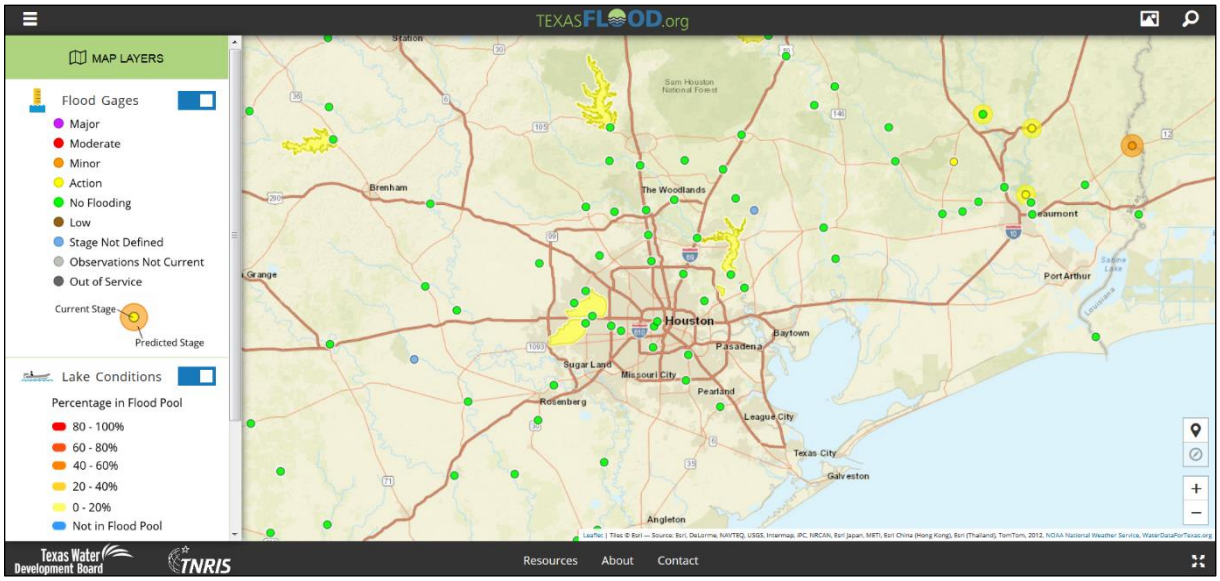


Figure 1.15 : Interface cartographique de visualisation des données relatives à la prévision et à la surveillance des inondations (Texas Water Development Board)

D'autres portails institutionnels intègrent des fonctionnalités permettant de créer très rapidement des cartes thématiques : sur la figure 1.15, l'internaute a utilisé les données de recensement pour créer une carte thématique du nombre de foyers ne disposant d'aucun abonnement à Internet et dont les revenus annuels sont inférieurs à 20 000$, par *census tract* sur le comté de Harris (métropole de Houston).

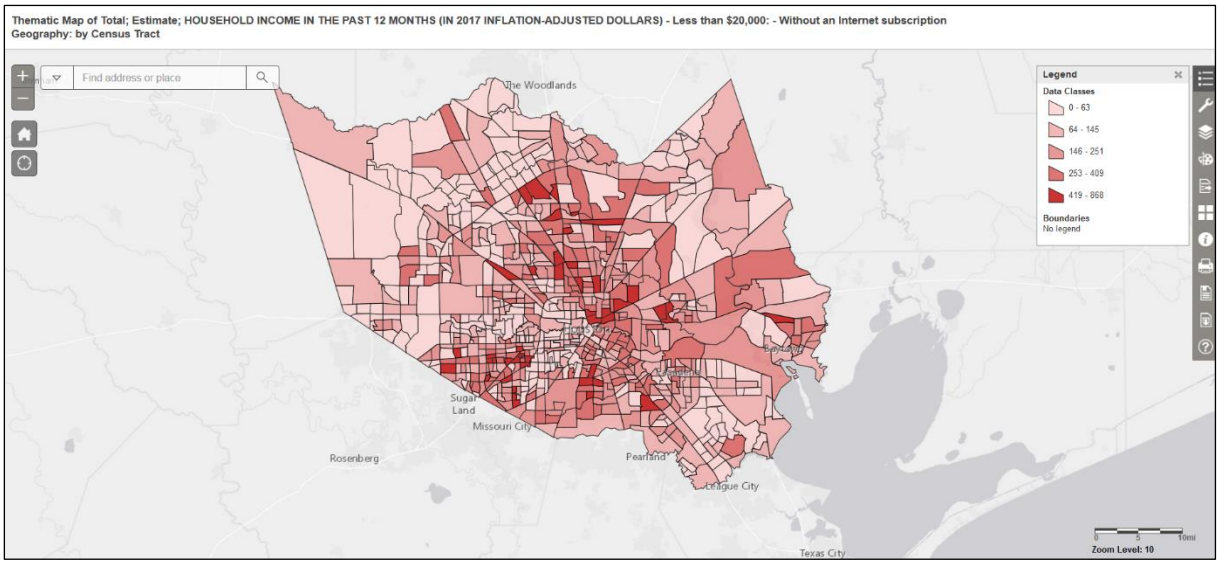


Figure 1.14 : Carte thématique représentant le nombre de foyers dont les revenus annuels sont inférieurs à 20 000 $ et ne disposant d'aucun abonnement à Internet (American Fact Finder)

Cohabitant aux côtés de ces portails diffusant des données géographiques traditionnelles, fleurissent les cartes du Géoweb social qui sont l'expression personnalisée de l'usager et de ses pratiques (Noucher, 2017). La carte n'est plus un produit qui donne une vision du monde unicentrée, celle de l'expert cartographe, mais un objet multisource qui donne une vision égocentrée des pratiques territoriales (Noucher, 2017). De par son foisonnement sur le Web, elle crée un nouvel espace virtuel de cartes, dans lequel chaque objet carte est porteur des représentations et usages de son contributeur (Joliveau *et al*., 2013 ; Noucher, 2017). La figure 1.16 soumet l'exemple d'une trace numérique représentant l'itinéraire d'une randonnée pédestre. La trace, qui dans ce cas précis est une production volontaire, est dessinée à la main par les outils mis à disposition des internautes sur le Géoportail de l'IGN ; elle se superpose ainsi au fond topographique traditionnel de l'IGN. Par les fonctionnalités de ce portail, on peut certes exporter une carte dans un format d'image. Mais surtout, la trace créée peut devenir donnée géographique et rapidement être diffusée et partagée sur le Web : ainsi, du simple dessin, elle devient un fichier KML (format de données de Google Earth), lequel fichier peut alors être importé dans l'interface cartographique d'un site collaboratif dédié aux échanges d'itinéraires de randonnées, puis téléchargé et enregistré dans le GPS de tout autre internaute. En conséquence, quel que soit le type de trace considéré, sa représentation cartographique informe les internautes de l'existence d'une pratique spatiale.

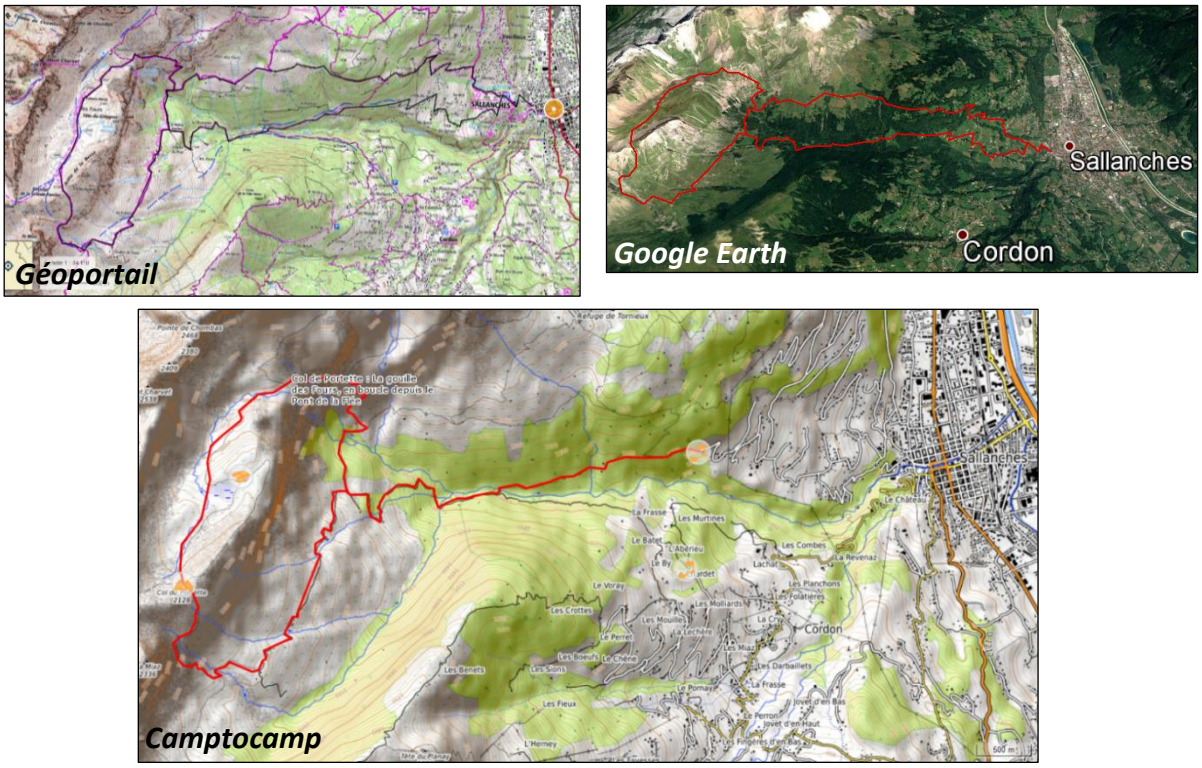


Figure 1.16 : Exemples de traces de randonnée importées dans différentes interfaces cartographiques du Géoweb et sites collaboratifs : (Géoportail, Google Earth, Camp to Camp)

L'internaute n'a donc plus besoin de compétences particulières pour produire des données ou une carte, pas plus que pour utiliser des fonctionnalités de recherche de données ou de navigation : la cartographie a ainsi glissé d'une discipline rigoureuse et productrice de contenus construits vers un objet ouvert et spontané, de multiples producteurs vers de multiples consommateurs (Joliveau *et al.*, 2013 ; Mericskay, 2016). L'ouverture des acteurs et des types des données (Mericskay, 2016) signifie-t-elle alors la dégradation du sens de l'objet carte (Joliveau *et al.*, 2013) ? Dans les faits, une majorité des cartographies du Géoweb social se réduisent à de simples fonds de cartes sur lequels se superposent les points d'intérêt (donc les données, comme les traces présentées dans la figure 1.16) du territoire renseignés par les internautes contributeurs (Joliveau *et al.*, 2013 ; Mericskay, 2016). Ces cartes sont avant tout néogéographiques dans la mesure où elles ne sont pas destinées à produire de l'information géographique, au sens où nous l'avons définie dans le premier axe de ce chapitre. Au contraire, elles s'accommodent de l'illustration la plus sommaire des rapports du contributeur au territoire (dans la figure précédente, pourquoi choisir tel itinéraire plutôt qu'un autre ?) par la production d'une donnée brute. En effet, l'analyse de l'information géographique implique la recherche de corrélations entre les manifestations spatiales des phénomènes étudiés. Le Géoweb social étant ancré dans une logique prédominante de géolocalisation systématique de tout contenu numérique créé par un internaute, ces cartes s'inscrivent davantage dans une posture de représentation spatiale visuelle du contenu web géolocalisé (Joliveau *et al.*, 2013 ; Noucher, 2017), comme l'indique la figure 1.17 :

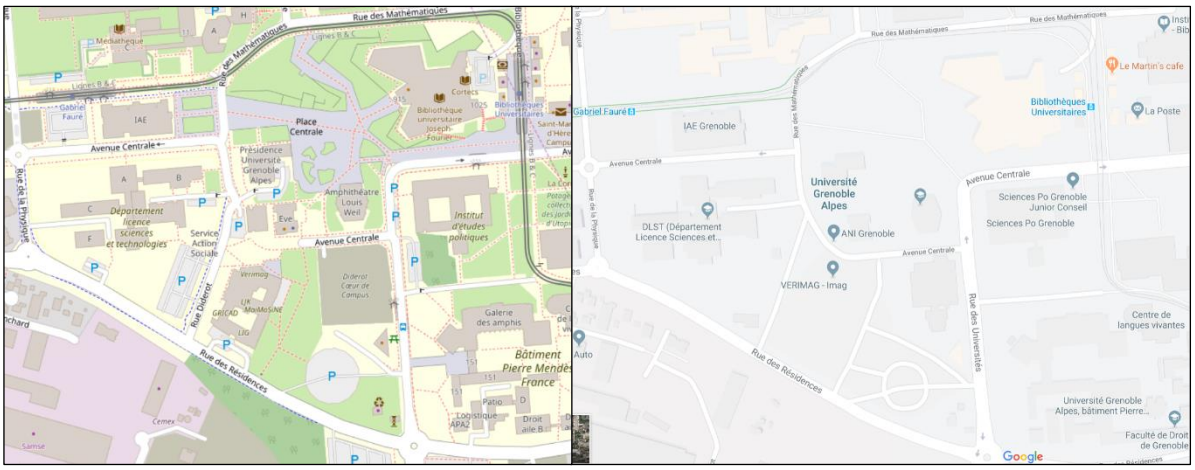


Figure 1.17 : Deux cartes du Géoweb pour lesquelles les contributeurs renseignent les objets "habillant" le territoire représenté (OpenStreetMap à gauche et Google Maps à droite)

Par ailleurs, si la cartographie traditionnelle s'accompagne historiquement d'un discours ou sert d'appui à la prise de décision, les nouveaux rapports du grand public à la donnée géographique et à la carte semblent s'insérer dans le temps présent, voire même dans l'instantanéité. Un automobiliste peut consulter des données relatives à l'état du trafic routier en temps réel et ce, avant même son départ et pendant sa navigation. De la même manière,

le *géocaching*[41] ou plus récemment, les jeux en réalité augmentée évoqués dans l'introduction, constituent des exemples par lesquels les usagers entrevoient et parcourent le monde physique par l'utilisation, en temps réel, de la carte et des données du Géoweb. Données et cartes du Géoweb social correspondent ainsi, en premier lieu, au partage d'expériences individuelles du territoire.

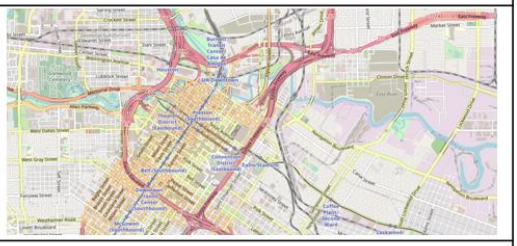
1.2.3.2. Les formes des cartes du Géoweb

Bien qu'elles aient été construites sur des interfaces web institutionnelles ou libres, les cartes présentées dans les figures précédentes ont deux points communs évidents : elles témoignent du passage de la société du papier vers la société de l'écran (Mericskay, 2016) ; inscrites dans la continuité de la posture de l'internaute acteur/contributeur, elles reposent sur le principe d'une interactivité se déclinant à plusieurs degrés. En effet, celle-ci peut apparaître sous sa forme la plus sommaire (cf. figure 1.14 : l'internaute peut déplacer la carte, zoomer/dézoomer, sélectionner les couches qu'il souhaite visualiser), ou intégrer des outils de création de données (cf. figures 1.16 et 1.17), ou des assistants de construction de cartes thématiques, à la manière des SIG (cf. figure 1.15). D'une cartographie expertisée à vocation planificatrice et diffusée auprès d'autres professionnels ou élus, on passe à une cartographie de navigation et de visualisation d'informations centrée sur les attentes de l'internaute. Plus précisément, la cartographie du Géoweb a rapidement développé ses propres normes et son esthétique graphique.

La carte du Géoweb est en premier lieu constituée d'un fond cartographique caractéristique : du fond classique de *Google Maps* ou d'*OpenStreetMap* pullule une pléthore de référentiels libres ou produits par des sociétés, faisant désormais partie intégrante du *design* numérique de la carte et ayant même été assimilés par les acteurs des cartes officielles (Mericskay, 2016) : Géoportail propose ainsi aussi bien ses propres fonds de carte traditionnels de l'IGN que des fonds développés par l'entreprise ESRI et également un fond *OpenStreetMap*. Le tableau 1.5 soumet quelques exemples, non exhaustifs mais fréquemment rencontrés, de la diversité des fonds de carte du Géoweb :

[41] Le géocaching est une forme de chasse au trésor pratiquée à l'aide du GPS, à l'échelle du globe, qui consiste à rechercher des caches (boîtes contenant un registre des visites et des objets divers) géolocalisées. Source : https://fr.wikipedia.org/wiki/G%C3%A9ocaching

Tableau 1.5 : Fonds cartographiques du Géoweb

Type de référentiel cartographique	Exemple de fond de carte	Illustration
Fond « classique »	OpenStreet Map	
Fond topographique	Stamen Terrain	
Fond design Visualisation scientifique	ESRI Dark Gray Canvas	

Dans un second temps, étant donné que la géolocalisation et l'annotation des objets du territoire, généralement d'implantation ponctuelle, constituent le fondement de la cartographie du Géoweb, son *design* cartographique s'illustre par la prépondérance des pointeurs et punaises (Mericskay, 2016), quelle que soit leur forme, permettant d'identifier les lieux et objets du territoire renseigné. Tout objet de la carte est cliquable : un clic affiche généralement un *mash-up*, c'est-à-dire une application qui développe un contenu provenant de l'agrégation de données issues d'applications tierces du Web (Noucher, 2017). Pour autant, en réponse vraisemblable au flux croissant de données, et principalement des traces géolocalisées produites en permanence, les plateformes cartographiques du Géoweb intègrent des méthodes de représentation graphique ainsi que des codes sémiologiques destinés à alléger le graphisme de la carte (Mericskay, 2016). En outre, ces modes de représentation des données s'inspirent des outils et des normes qui pré-existaient dans les SIG :

- les données d'implantation ponctuelle peuvent être représentées sous forme de carte de chaleur (soit une fonctionnalité traditionnelle du mode raster des SIG) ou encore sous forme agrégée par des *clusters* : la figure 1.18 ci-dessous affiche une carte des tweets relatifs à l'ouragan Harvey, dans le centre de Houston. Les cercles représentent chacun un *cluster,* soit un ensemble de tweets géolocalisés agrégés en une seule entité en fonction d'un critère

spatial de distance. Si un point est trop éloigné d'un nuage de tweets, il apparaît sous la forme d'un marqueur bleu[42].

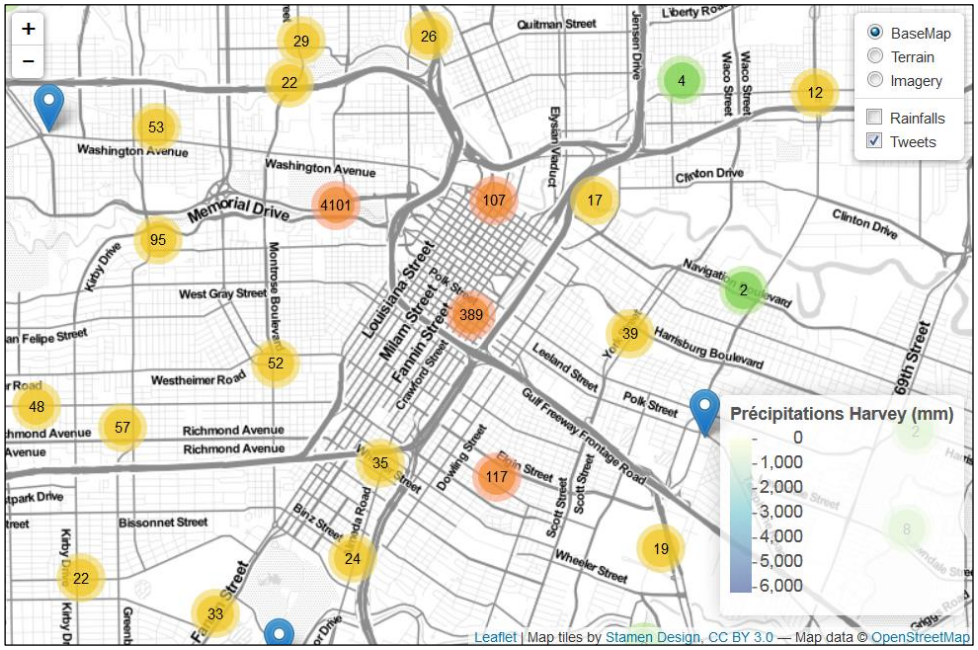


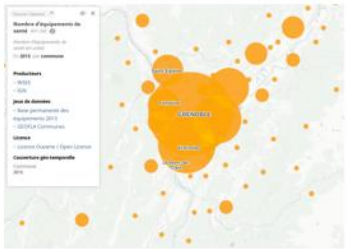
Figure 1.18 : Tweets géolocalisés agrégés sous forme de clusters (Carte réalisée avec le package leaflet de R)

La carte est navigable et zoomable : si l'internaute zoome, l'agrégation des données en clusters est automatiquement recalculée. Pour autant, cet exemple montre typiquement le paradoxe actuel de la carte : comme le veut la tradition, la carte témoigne de la localisation d'objets sur un territoire mais, comme la confection des méthodes de construction échappent au savoir du géographe, elle s'affranchit du questionnement et des contraintes de la discipline : ainsi, on ne connaît pas le critère de distance retenu pour constituer les clusters et la sémiologie graphique de ces agrégats ne correspond pas aux normes de la discipline.

- dans certains cas au contraire, la représentation des données suit en partie les normes énoncées par Bertin : le tableau 1.6 affiche un panel de cartes du Géoweb : parmi les variables visuelles utilisées, on retrouve la forme, adaptée à l'identification d'un objet, les valeurs de teintes pour témoigner de l'intensité d'un phénomène, ou encore la taille qui indique également une gradation représentant des valeurs absolues.

[42] Une représentation cartographique en clusters, comme proposée sur la figure 1.18, n'est donc pas sémiologiquement identique à une carte en cercles proportionnels : une carte de tweets géolocalisés en cercles proportionnels indiquerait le nombre de points recensés sur une entité géographique fixe alors que la clustérisation agrège les points pour en faciliter la visualisation et est recalculée automatiquement en fonction du niveau de zoom de l'utilisateur.

Tableau 1.6 : Exemples de cartes du Géoweb adoptant une sémiologie graphique professionnelle

Carte du Géoweb	Exemple	Sémiologie
Savoie Routes – Etat du trafic routier http://www.savoie-route.fr/		Valeurs de couleurs : plus la couleur est rouge foncée, plus la circulation est ralentie. Forme : panneaux identifiant les sections en travaux, les incidents, les cols fermés etc.
Pokémon Go – Quelle créature/arène/centre Pokémon trouver dans quel lieu ? https://pokewebgo.com/fr/#		Forme : pointeurs identifiant les créatures et lieux d'intérêt du jeu
DataFrance – Nombre d'équipements de santé par commune en 2013 http://map.datafrance.info/		Taille : cercle dont la taille varie en fonction du nombre d'équipements par commune

La cartographie du Géoweb peut néanmoins se soustraire aux règles sémiologiques, statistiques et graphiques, qui sont pourtant indispensables à la valorisation des données en information géographique et à l'intelligibilité du message cartographique (Mericskay, 2016). Les cartes en question peuvent aussi bien être le fruit de professionnels que d'internautes dans le cadre d'un usage récréatif ou informatif : si la carte collaborative représentant la localisation des Pokémon a tendance à paraître surchargée, la carte des équipements de santé issue du portail de visualisation de données DataFrance, construite à partir de jeux de données diffusés par l'IGN et l'INSEE témoigne également d'un défaut de lisibilité. Les cartes du Géoweb arborent ainsi deux types d'aberrations fréquentes :

- l'erreur statistique : la carte présentée dans la figure 1.15 (*Carte thématique représentant le nombre de foyers dont les revenus annuels sont inférieurs à 20 000 $ et ne disposant d'aucun abonnement à Internet*) est fausse : l'internaute a sélectionné la variable à cartographier exprimée en valeurs absolues, qui se trouvent représentées par la variable de Bertin définie comme *valeurs de teintes* ; pour assurer la justesse de la représentation, il aurait dû sélectionner les valeurs exprimées en pourcentage. De plus, on ne sait pas s'il est familier

des méthodes de discrétisation proposées par l'outil ou s'il s'est contenté de valider les paramètres proposés par défaut.

- l'erreur sémiologique : la carte de la figure 1.19[43] provient d'un site officiel mais peut être retravaillée : elle représente l'Indice de Vulnérabilité Sociale[44] par comté des Etats-Unis. Plus cet indice est faible, moins la population de comté est vulnérable face aux risques. Plus le score est élevé, plus la population du comté est vulnérable. Le géographe aurait utilisé, de préférence, la variation d'une même teinte pour faire transparaître cette gradation, soit la variable visuelle valeurs de teintes définie par Bertin.

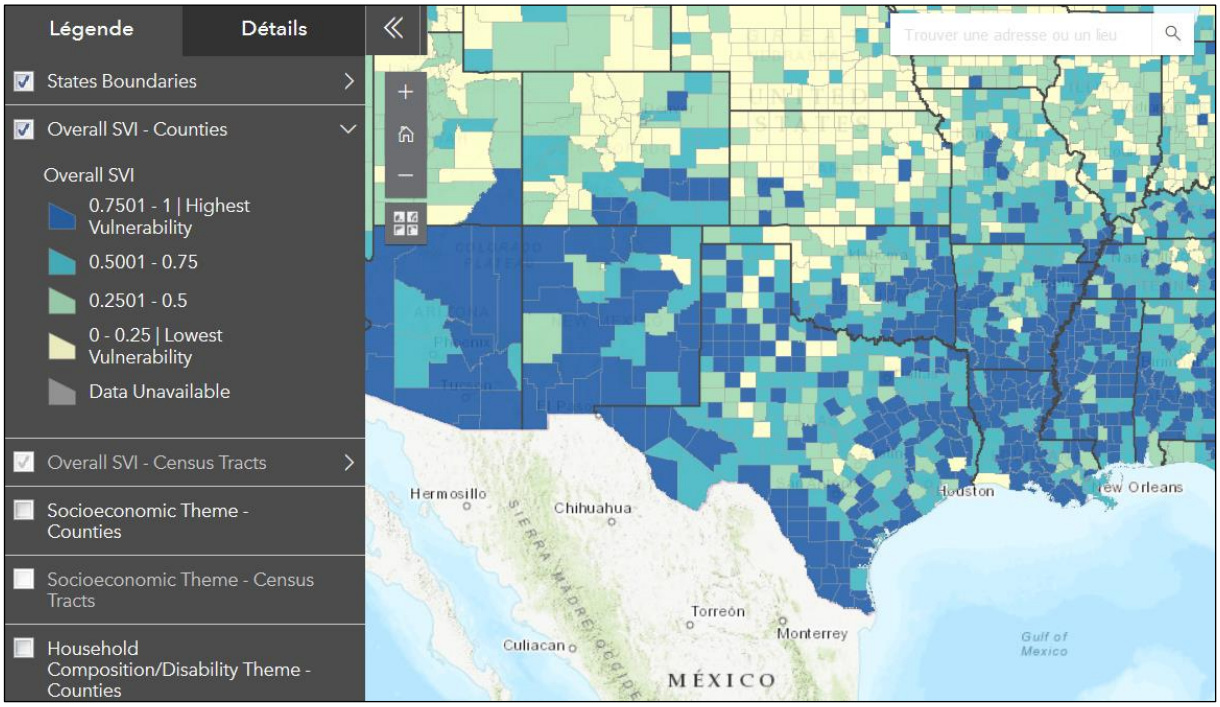


Figure 1.19 : Carte de l'Indice de Vulnérabilité Sociale des populations par comté (Centers for Disease Control and Prevention)

Les représentations cartographiques foisonnantes du Géoweb arborent ainsi de multiples facettes : elles peuvent aussi bien être créées par des professionnels que par des internautes amateurs, de la même manière qu'elles peuvent constituer le support de visualisation de données expertisées comme de données collaboratives. Si les enjeux des données expertisées ont été décrits précédemment, comment se positionnent les données collaboratives dans l'information géographique ?

[43] Source : https://svi.cdc.gov/map.html (Consulté pour la dernière fois le 25/02/2019)
[44] Il sera plus longuement question de cet indice dans le chapitre 4 qui présente les données officielles mobilisées dans le cadre de la recherche.

1.2.4. Enjeux et positionnement des nouvelles données du Géoweb

1.2.4.1. Un exemple d'école pour poser les enjeux

Au cours de ses sorties estivales, un randonneur a l'habitude de photographier les espèces herbacées qu'il considère comme *remarquables*. Il ne se contente pas d'une simple photographie : comme il est équipé d'un smartphone, il prend également le soin d'enregistrer les coordonnées GPS de la photo, auxquelles sont ajoutées d'autres données : la date et l'heure de la photographie, le lieu-dit et un commentaire. Ainsi, à la fin de la saison, la randonneur a photographié, c'est-à-dire sauvegardé la présence d'un certain nombre de plantes, qu'il a jugées dignes d'une attention particulière. Le tableau 1.7 affiche quelques exemples des données collectées. Le randonneur, qui n'est ni botaniste, ni professionnel de la production de données, assure-t-il une collecte consciencieuse et fiable de données de terrain ?

Tableau 1.7 : Un type potentiel de données botaniques géolocalisées par un randonneur

Photographie	Genre/Espèce	Date	Latitude	Longitude	Commentaires
	Orchidée	2018/06/02 11:03:50	45.944818	6.581072	Alpage à vaches, juste après la traversée du torrent qui descend du Pas de Monthieu
	Dactylorhize de mai	2018/07/18 13:12:28	45.912882	6.543110	Marécage en contrebas du col de Niard / une quinzaine de tiges + prêle et ciboulette
	Digitale jaune	2018/07/14 11:03:49			Pierre de l'alpage des Tarines, sous le chemin qui monte aux Allènes
	Orchis globuleux	2018/08/21 10:26:39	45.946567	6.569836	Alpage à moutons du Lancheron, une seule tige
	Lis Martagon	2018/07/18 14:45:37	45.937065	6.560995	3 tiges après la traversée de la passerelle enfoncée (direction Pierre Fendue) ; au milieu des éboulements des Fours; **Vu ici pour la première fois**

Les données consignées dans le tableau 1.7 illustrent les difficultés à qualifier, d'un point de vue professionnel, la qualité et la fiabilité des données créées sur le Géoweb social par les internautes contributeurs :

- en premier lieu, le randonneur choisit de sauvegarder l'existence de plantes particulières : il en *sélectionne* certaines et en exclut d'autres, sans qu'il indique les raisons qui le poussent à choisir une plante précise (à l'exception probable du lis martagon, rencontré

pour la première fois dans l'espace indiqué). Le jeu de données produit est donc *non exhaustif* et *subjectif* ;

- des données *manquent* : un professionnel apprécierait de voir figurer les caractéristiques de la station (type de sol, type de milieu, ensoleillement, etc.). De plus, les coordonnées GPS de la digitale n'étant pas renseignées, la donnée est perdue si l'on l'affiche sur une carte (pour assurer la sauvegarde de la plante, il serait alors pertinent de la géolocaliser manuellement) ;

- des données restent *imprécises* car le randonneur n'a pas reconnu l'espèce de l'orchis à odeur de sureau (orchidée jaune) ;

- une donnée est *fausse* : la digitale, bien que de fleurs jaunes, est une digitale à grandes fleurs ;

- la rubrique des commentaires manque de *cohérence* : le randonneur évoque brièvement les conditions environnementales du terrain ou la présence d'un alpage mais ces informations ne sont pas fournies ensemble de manière systématique.

Le tableau de données affiche ainsi un certain nombre d'imperfections, qu'on peut lier à trois facteurs principaux : les difficultés à identifier une plante, le non-recours systématique à l'usage du GPS et le fait que le randonneur ne pratique pas de relevé standardisé par une grille de critères précis, à la manière d'un professionnel. Néanmoins, si le comportement de ce randonneur n'est pas une pratique isolée et si un certain nombre d'individus l'adoptent et la perpétuent dans le temps, alors les données acquises offrent le potentiel de constituer une base de connaissances utile à l'étude de l'évolution des espèces et des milieux[45].

1.2.4.2. Des concepts-clés : le citizen-as-sensor et la familiarité au territoire

On peut effectivement mettre en exergue la pertinence des données générées par les contributeurs du Géoweb 2.0 par deux concepts : le *citizen-as-sensor* (qui sera ici traduit par l'expression individu-capteur) et la familiarité à l'espace.

Introduit en 2007 par Goodchild, le concept de l'individu-capteur entremêle deux idées : il désigne en premier lieu tout individu localisable, soit parce qu'il détient un appareil équipé d'un GPS embarqué, soit parce qu'il est connecté à un réseau qui va estimer la localisation du périphérique via l'adresse IP : tout usager d'un GPS routier, de même que tout usager du Web laisse ainsi des traces de sa navigation. Dans un second temps, il qualifie tout individu fournissant des indications sur des objets localisés ou des lieux particuliers : c'est le cas d'un contributeur *OpenStreetMap* ou du randonneur présenté dans le paragraphe 1.2.4.1. L'usage généralisé des outils du numérique et de l'Internet peut ainsi attribuer un nouveau rôle central à l'individu, dans la mesure où il suffit de posséder l'appareil adéquat et une connexion à un

[45] C'est notamment le fondement des *Citizen Sciences* (ou sciences participatives) qui correspondent à la production d'un savoir scientifique à laquelle des non-scientifiques participent volontairement, notamment dans les phases d'inventaire et de collecte des données. Source : https://fr.wikipedia.org/wiki/Sciences_participatives#Principes (Consulté pour la dernière fois le 01/03/2019)

réseau pour être un contributeur potentiel. Autre conséquence de la généralisation de l'accès à ces outils : des milliers de contributeurs potentiels sont d'ores-et-déjà déployés et parcourent les territoires (Schade *et al*., 2013).

Les données produites par l'individu-capteur ont-elles alors du sens ? Goodchild envisage l'individu comme l'élément clé de son environnement. En effet, tout individu qui entretient une expérience et des pratiques quotidiennes avec un espace particulier développe un certain degré de *familiarité* (Goodchild, 2009) avec cet espace, ce qui le rend capable de détecter et de rendre compte précisément de tout changement environnemental. Cela signifie que les individus utilisent leurs facultés d'observation pour acquérir de la connaissance relative à leur espace quotidien et sont donc sensibles aux changements subits. Plus la fréquentation d'un espace est longue, plus le degré d'expertise apporté par l'individu-capteur peut être précis. Par conséquent, tout individu impliqué et connecté peut produire une information de terrain, pertinente et valorisable, de la même manière qu'un groupe d'individus qui assistent à un événement naturel est capable de produire un jeu de données plus précis qu'un expert non familié du terrain (Goodchild et Glennon, 2010), celui-ci étant généralement dépêché après la fin de l'événement (Dashti *et al.*, 2014).

1.2.4.3. Les promesses des données du Géoweb

L'émergence de la disponibilité des nouvelles formes de données numériques produites par les individus-capteurs a suscité un enthousiasme tel qu'il inspire le sentiment, à la lecture des publications réalisées au début des années 2010, que ces données allaient bouleverser le monde de la recherche et marquer l'avènement d'un nouveau paradigme de construction de la connaissance (Anderson, 2008 ; Kitchin, 2013 ; Miller & Goodchild, 2015) dans une nouvelle posture *bottom-up* (la donnée remonte du producteur vers le chercheur) : ce n'est donc pas le chercheur qui capture des données en fonction de ses besoins mais c'est le contenu du réseau qui dicterait les pistes sur lesquelles le chercheur oriente sa recherche (Miller & Goodchild, 2015). Dans la littérature, on constate ainsi une assertion redondante : le renouvellement de nos sources d'information traditionnelles par les nouvelles formes de données numériques 2.0 offre des perspectives inédites pour refonder notre compréhension des phénomènes socio-spatiaux et spatio-temporels (Andrienko *et al*., 2013 ; Steiger *et al*., 2015 ; Lucchini *et al*., 2016 ; Cebeillac *et al*., 2017). Pourquoi une telle affirmation ?

En premier lieu, alors que le processus de création de données géographiques traditionnelles est resté le domaine d'activité d'experts peu nombreux, isolés et éloignés de la réalité de terrain, les individus-capteurs ont la faculté de produire des jeux de données sur des thématiques qui restaient marginalisées et ignorées par les producteurs de données géographiques traditionnelles (Elwood *et al*., 2011 ; McDougall, 2012), notamment toute donnée pouvant concerner directement les habitudes de vie, les opinions et les pratiques spatiales des individus. Dans un second temps, ces jeux de données s'avéreraient complets et exhaustifs à la différence des enquêtes qui s'appuient sur un échantillon issu d'une population

mère (Goodchild, 2013). Ce second point peut en revanche être discuté : si l'on reprend l'exemple botanique du paragraphe 1.2.4.1, le randonneur sélectionnait les plantes à capturer ; mais s'il n'est pas seul à adopter cette démarche, un groupe de contributeurs pourrait parvenir à inventorier l'ensemble des végétaux herbacés. Par ailleurs, si l'on peut constituer un sondage représentatif d'une population mère et généraliser des résultats avec une certaine marge d'erreur, nous verrons que la caractérisation précise des profils de populations régulièrement productrices de données et traces numériques géolocalisées, et notamment des tweets géolocalisés, reste une question épineuse.

En ce qui concerne les questions géographiques, la discipline place les rapports entre l'individu et le territoire comme pilier : comprendre un territoire ne se réduit pas à la simple description quantitative ou qualitative de ses objets. En cela, les données de l'individu-capteur, et plus particulièrement les traces numériques, introduisent une dimension personnelle dans les données (Miller, 2007) qui offre l'opportunité de saisir les représentations et pratiques de l'espace par ceux qui le vivent et le parcourent en temps réel (Andrienko *et al*., 2013). Pour (Elwood *et al*., 2011), ces données représentent alors la chance d'intégrer une *approche anthropocentrée* à l'analyse spatiale traditionnelle en recentrant l'objet d'étude sur l'individu et ses activités dans le temps et dans l'espace.

1.2.4.4. Quelles caractéristiques pour les données du Géoweb ?

L'avènement des données issues des individus-capteurs marque une rupture entre deux époques (Elwood *et al*., 2011) : un temps où la production de données géographiques était un processus long, codifié, centralisé et dispendieux et un temps récent où la géolocalisation généralisée des contenus web transpose tout type de donnée en donnée géographique potentielle (Mericskay, 2016).

Quelle place accorder alors aux données géolocalisées produites par les technologies 2.0 parmi les données géographiques ?

Au premier abord, les apparences des données géolocalisées 2.0 inciteraient à les juxtaposer au niveau des données géographiques traditionnelles des géographes : en effet, lorsqu'on regarde les différents exemples de données contributives ou de traces numériques géolocalisées énumérées dans les paragraphes précédents, on constate qu'elles se conforment aux différentes caractéristiques qui confèrent leurs spécificités aux données géographiques : elles disposent d'une composante spatiale puisqu'elles sont géolocalisées par leurs coordonnées GPS, mais également d'attributs qui les décrivent[46]. En outre, elles peuvent constituer des données de référence (c'est le cas des données *OpenStreetMap*), des données

[46] Notons néanmoins que les données que nous appelons ici "attributs" correspondent à ce que les développeurs considèrent comme métadonnées. Nous retenons la terminologie des SIG pour qualifier ces données car elles décrivent les propriétés de l'objet géolocalisé. En fait, il n'y a pas de métadonnées au sens professionnel du terme.

thématiques (carte de localisation des refuges en période de crise, carte des traces de randonnée, carte collaborative de localisation des points d'intérêt du jeu en réalité augmentée), voire des données temporelles (à travers l'exemple des données botaniques si la collecte est régulière).

Malgré cela, la majorité des données géolocalisées du Web, en dehors de la VGI, correspondent à une production spontanée, consciente ou inconsciente des individus capteurs : en conséquence, les données géolocalisées 2.0 sont fréquemment produites sans intention particulière de diffuser une information pertinente (ce qui est le cas des traces numériques géolocalisées). C'est pourquoi (Quesnot, 2016) critiquait le recours systématique aux données géolocalisées et les espérances de renouveau des théories qu'elles suscitaient dans la communauté académique, citées plus haut :

- comme indiqué dans l'introduction de ce manuscrit, le propre des données numériques géolocalisées du Web 2.0 est leur acontextualité : en effet, le chercheur qui est totalement absent du processus d'acquisition de la donnée (Audard *et al.*, 2014), ne connaît pas les motivations qui incitent un internaute à créer, à tel endroit et à tel moment, une donnée géolocalisée. En conséquence, la présence de données en un lieu précis ne permet pas d'extrapoler précisément le sens du lieu pour l'ensemble des individus qui le parcourent.
- le deuxième argument découle directement des effets de cette acontextualité : les données géolocalisées 2.0 ne sont pas des données géographiques au sens de la discipline : Henri Lefebvre envisageait l'espace géographique comme une entité à trois dimensions : l'espace physique, l'espace mental et l'espace social (Joliveau, 2010). Les espérances formulées vis-à-vis des données géolocalisées 2.0 reposent ainsi sur le fait qu'elles sont, d'une part, localisées dans l'espace physique, et d'autre part, qu'elle constituent la porte d'entrée, de par leur contenu, sur les représentations et pratiques de l'espace vécu par les individus. Or, le fait que la donnée géolocalisée soit coupée de son contexte de capture coupe court à la possibilité de restituer le sens de la pratique associée à toute donnée[47]. La dimension géographique de la donnée géolocalisée 2.0 se réduit ainsi au simple ajout des coordonnées XY.
- enfin, ces données sont avant tout des produits commerciaux capitalisés par les sociétés surnommées les *Géants du Web* dans un but purement commercial. Le chercheur est absent des processus de régulation et de diffusion des données ; en fait, il ne collecte gratuitement qu'une partie des masses de données qui restent entre les mains de ces entreprises.

Ainsi, M.Goodchild, après avoir été lui aussi plutôt enthousiaste face à l'avènement de l'ère de la donnée géolocalisée 2.0, adressait une mise en garde dès 2013 : il indiquait alors que le flot de recherches s'appuyant sur ce nouveau matériau, et en particulier les tweets

[47] Il s'agit du problème que nous avons d'ores-et-déjà identifié précédemment : la trace de randonnée n'explique pas les choix d'itinéraire du randonneur ; de la même manière, on ne sait pas ce qui motivait la sauvegarde de la présence d'une espèce herbacée particulière.

géolocalisés, reposent sur la *croyance* (terme repris par Quesnot en 2016) qu'ils constituent une nouvelle source d'informations sur la dimension sociale de l'espace.

Des Big GeoData aux propriétés floues.

En fait, les données géolocalisées numériques, qui sont aujourd'hui le cœur du problème, se démarquent des données géographiques traditionnelles de par leurs propriétés bien distinctes : elles peuvent être qualifiées de *Big Geodata* (Goodchild, 2013). Les contenus spatialisés générés sur le Web par les internautes sont en effet à l'origine de nouvelles bases de données spatialisées particulièrement volumineuses (Joliveau, 2011 ; Pornon, 2015). Ces données, qui ont la caractéristique d'être collectées en continu, ont rapidement été intégrées au phénomène des *Big Data*, ce terme que les informaticiens et *data analysts* emploient pour désigner les masses de traces laissées plus ou moins consciemment par les internautes au cours de leur navigation ou capturées par les dispositifs intelligents qui accompagnent les individus au quotidien, et dont les propriétés dépassent les capacités techniques de nos moyens traditionnels de stockage et d'analyse (Quesnot, 2016). Bien que la définition et la portée des *Big Data* demeurent toujours floues (Goodchild, 2013 ; Kitchin, 2013), les auteurs s'accordent à décliner leurs caractéristiques principales selon la formule des 3V :

- le volume (*Volume*) : ces données représentent des téra voire pétaoctets de stockage et génèrent une situation de surabondance de données : en l'espace de quelques jours, on peut produire une masse de données bien plus conséquente que l'ensemble de données existant avant les années 2000 (Kitchin, 2013) ;
- la rapidité (*Velocity*) : les données sont acquises en temps réel et collectées de manière quasi continue ;
- la variété (*Variety*) : les données se présentent sous des formes diversifiées.

Leur intégration à des problématiques de recherche en sciences humaines, et particulièrement en géographie n'est pas sans conséquence méthodologique. En premier lieu, elles concourent à la transition d'un environnement de données éparses et coûteuses vers un environnement où la donnée est ubiquiste mais non régulée, non structurée, non standardisée (Kitchin, 2013), ce qui soulève un certain nombre de questions épistémologiques[48]. Ensuite, si les informaticiens appréhendent les *Big Data* comme un phénomène exhaustif, il n'en est pas le cas pour les géographes ; comme indiqué précédemment, le nombre, qui est mis en avant comme élément significatif par Chris Anderson, n'est pas un gage de sens en géographie. En outre, les *Big Data* ne permettent plus de travailler à partir de données issues d'un échantillon représentatif. Les traces et données numériques géolocalisées sont produites par les usagers des TIC qui utilisent des fonctionnalités et plateformes particulières : on peut ainsi tenter d'établir des profils de populations productrices et consommatrices de ces outils (Li *et al*., 2013) mais cet exercice ne

[48] Ces questions font l'objet d'une présentation et d'une discussion plus approfondies dans les chapitres 3 et 4.

se révèle par forcément concluant. Au final, l'enjeu consiste à savoir si la connaissance qu'on peut acquérir à partir des traces et données géolocalisées produites par une population particulière, c'est-à-dire les producteurs, est le reflet des pratiques de l'ensemble des individus qui vivent un territoire (Goodchild, 2013).

Quid de leur qualité ?

Si les données traditionnelles, qu'on qualifie également de *hard data* en raison des contraintes et normes qui régissent leur production, sont renseignées sur divers paramètres de qualité par leur série de métadonnées associées, qu'en est-il des données et traces géolocalisées du Géoweb qui peuvent être produites par tout internaute ? Des études ont d'ores-et-déjà indiqué que des jeux de données issus d'*OpenStreetMap* étaient spatialement précis, bien que leurs attributs s'avèrent parfois incomplets (Haklay, 2010) ; de la même manière, les données collectées par des programmes de sciences participatives intégrant donc des individus non experts de la société civile à des programmes scientifiques ont également prouvé leur valeur, dans le domaine de l'écologie mais également des risques naturels, domaine dans lequel cette participation peut engager la création d'une culture du risque améliorant la résilience des populations (Paul *et al*., 2018). Mais qu'en est-il des traces numériques géolocalisées produites inconsciemment ou sans intention particulière d'informer un tiers ?

Si l'on cherche à caractériser les biais éventuels portés les traces numériques, on peut mettre en évidence plusieurs facteurs d'incertitude :

- L'incertitude spatiale concernant la géolocalisation de la trace par le GPS (Elwood *et al*., 2011) ;
- l'incertitude sémantique : lorsque l'individu-capteur décrit un lieu, un événement, une activité ou un objet du territoire, un processus d'interprétation entre en jeu[49] : l'individu décrit avec son propre vocabulaire, en fonction de son expérience personnelle et de sa familiarité au lieu ou à l'objet. Pour (Arnaud, 2009), cette incertitude sémantique est directement imputée à la subjectivité de l'observateur qui interprète l'objet ou le lieu en fonction de son degré de compréhension et de connaissance.

Pour pallier ces difficultés, certains chercheurs ont alors mis en évidence deux pistes envisageables :

- la véracité de l'information (la trace indique-t-elle une réalité objective ou constitue-t-elle une fausse donnée, qu'on appelle du *bruit*) peut être vérifiée par la première loi de

[49] Si l'on reprend l'exemple du randonneur et de ses données botaniques, il y a un processus d'interprétation lors de l'identification de l'espèce et de la description des lieux.

Tobler[50] : une trace dont la véracité semble douteuse peut être comparée aux traces localisées dans ses environs, voire même à des données officielles (Goodchild & Glennon, 2010) ;

- le nombre de traces peut également être un indice : dans la même logique, plus on trouve de données géolocalisées rassemblées en un lieu précis du territoire, mieux on pourra vérifier que les traces se recoupent et apportent une description précise de l'objet, événement ou lieu en question (Elwood *et al*., 2011).

De notre point de vue, la question qui demeure centrale et qui devra faire l'objet d'une discussion plus approfondie en fin de manuscrit sur ce volet de la qualité des traces reste la suivante : faut-il considérer et juger de la qualité et de l'incertitude des traces numériques géolocalisées avec des critères identiques aux données géographiques traditionnelles ?

Pour conclure, peuvent-elles servir à créer de l'information géographique, ancrée dans les attentes et usages professionnels du géographe ?

Un producteur de données géolocalisées 2.0, qu'il s'agisse d'un contributeur VGI ou de traces plus ou moins volontaires et conscientes, n'est pas, au premier abord, un créateur d'information géographique : un contributeur d'*OpenStreetMap* ne fait pas d'analyse spatiale à la manière d'un professionnel des SIG (Joliveau, 2010) : son champ d'action se restreint à la description partagée des objets de la surface terrestre ; en conséquence, la représentation de ces objets est focalisée sur leur localisation et non dans une perspective d'analyse d'un phénomène spatial (Joliveau *et al*., 2013). Ainsi, tout internaute qui signale un moustique tigre[51], une nuisance olfactive liée à la pollution atmosphérique[52], photographie une plante ou tout objet du territoire à un instant donné, crée une donnée numérique géolocalisée et son action s'arrête, en apparence, à cette tâche. Pour autant, son rôle peut s'amplifier : grâce aux outils des plateformes du Géoweb, l'internaute peut, toujours sans compétence particulière, valoriser ses traces ou données : on peut ainsi calculer le nombre de kilomètres parcourus, établir un profil altimétrique, cartographier la localisation des plantes photographiées et revenir, une année plus tard, pour constater la présence/absence/migration de l'espèce inventoriée l'année précédente et l'ajouter dans les données.

Et paradoxalement à toutes les limites mises en évidence précédemment, en particulier en ce qui concerne la dimension géographique des traces, on dispose déjà d'une certain nombre de résultats témoignant des possibilités de leur valorisation cartographique[53]. Le problème qui mérite alors toute l'attention des chercheurs en sciences humaines reste le suivant : quelle est la représentativité de ces résultats ? Indiquent-ils des tendances qui

[50] « *Everything is related to everything else, but near things are more related than distant things.* »
[51] Source : http://www.signalement-moustique.fr/ (Consulté pour la dernière fois le 04/03/2019)
[52] Source : https://www.atmo-odo.fr/odoaura (Consulté pour la dernière fois le 04/03/2019)
[53] Cf. Chapitre 3

peuvent être généralisées à d'autres lieux et à d'autres populations et qui persistent à travers le temps ?

Conclusion du chapitre 1

Si la production de données géographiques et de cartes est traditionnellement restée l'expression d'experts compétents, l'adoption du numérique et de la géolocalisation par la société civile a brisé les barrières qui restreignaient ces tâches à un cercle d'initiés et, en conséquence, démultiplié les acteurs du domaine ainsi que la nature des données. En fait, deux mondes de données cartographiables cohabitent désormais : les données expertisées construites par des instituts reconnus à destination de professionnels, qui alimentent la cartographie officielle des rapports et portails de visualisation de données, côtoient le flot de données géolocalisées produites sur le Géoweb, par tous, sans que leur création ne soit nécessairement associée à l'intention d'informer des destinataires spécifiques, ni de produire un contenu valorisable en information par la carte. Face à la pluralité de l'origine de ces nouvelles données et de leurs formes, il nous semblait alors essentiel de nous positionner sur le flou ambiant qui persiste dans la qualification des données géographiques. Nous considérons ainsi que :

- toute pratique de VGI ou de *crowdsourcing* aboutit à la création d'une donnée numérique contributive géolocalisée : elle constitue un apport volontaire, non contraint par l'utilisation d'un outil particulier, et témoigne d'une prise de conscience qu'il existe une visibilité par tous ainsi qu'une possibilité de réutilisation par un tiers. Cette pratique représente un comportement assumé : l'utilisateur choisit de localiser et de renseigner un objet du territoire.

- En ce qui concerne le cas enchevêtré et confus des traces numériques géolocalisées : nous avons défini la trace comme l'élément attestant de la présence d'un individu en un lieu précis à un moment donné et permettant un suivi régulier de son parcours spatial. Nous n'emploierons pas le terme de "données", ni le terme de "contributif", qui restera propre à la VGI, car la trace présente l'inconvénient d'être plus ou moins volontairement et consciemment produite par tout utilisateur des TIC et du Web. La trace peut être assumée ou subie : l'utilisateur d'un smartphone peut être rapidement localisé car son appareil va borner, indépendamment de sa volonté. S'il a conscience de ce fait, il assume la production inéluctable de traces ; dans le cas contraire, il la subit.

Dans tous les cas, nous restons prudents quant à l'application du qualificatif de "géographique" face à la nature des données ou traces numériques géolocalisées : c'est là une première limite du concept de l'individu-capteur. A la différence des données produites dans un programme de sciences participatives, les traces numériques géolocalisées sont capturées sans mention de contexte environnemental précis du moment, et sans prise de conscience de la nécessité de fournir un contenu précis et valorisable par un tiers. L'enjeu primordial reste alors le suivant : *quid* de la possibilité de construire une information géographique à partir de traces numériques géolocalisées mais non scientifiques ?

2. Twitter et les tweets géolocalisés : une plateforme de production de données géographiques ?

Twitter constitue l'une des plateformes du Web social qui brise les contraintes traditionnelles des communications, et par laquelle tout internaute peut endosser le rôle de créateur d'informations, dans de multiples logiques de diffusion puisque divers types d'acteurs sont inscrits sur le réseau : un utilisateur représentant un service ou une institution peut ainsi diffuser de l'information dans une logique *top-down*, tout comme un internaute de la société civile peut informer les acteurs du territoire dans une logique *bottom-up*[1]. Cette communication fonctionne également selon un sens horizontal : tout acteur du réseau, qui participe à la création de contenu à titre privé, peut communiquer vers un destinataire précis ou s'adresser à l'ensemble des membres du réseau. Twitter n'est certes pas la plateforme du Web social la plus prolixe et populaire parmi les internautes[2] mais elle offre l'avantage considérable d'être fondée sur un principe de communication rapide et spontanée, et d'intégrer des fonctionnalités de géolocalisation directe.

L'accès au contenu créé sur Twitter n'est pas réservé au seul groupe constitué des quelques 330 millions d'utilisateurs actifs mensuels du service puisqu'il existe diverses applications voire visualisations en accès libre qui permettent de filtrer et de sélectionner des tweets particuliers, ou encore d'observer des phénomènes de diffusion spatiale ou de variabilité temporelle des rythmes d'émission des tweets. Enfin, moyennant des compétences fondamentales en informatique et quelques lignes de code, tout internaute peut extraire et stocker sur sa machine, un jeu de tweets répondant à des critères qu'il aura sélectionnés, via des interfaces de programmation applicative.

Ce deuxième chapitre présente le type de contenu web que constitue le matériau brut de cette recherche. Il vise à circonscrire le tweet géolocalisé dans le monde des nouvelles données numériques géolocalisées, en rapport avec les réflexions engagées dans le chapitre précédent. Par ailleurs, il nous paraît essentiel de positionner les attentes particulières des géographes vis-à-vis de ces données, qui se démarquent de l'approche des *data analysts*. Ce chapitre se structure alors de la manière suivante : la première partie présente les principes de fonctionnement de Twitter ainsi que les différentes formes des tweets ; la deuxième partie présente les outils qui gravitent autour de la plateforme et la troisième partie positionne le tweet géolocalisé par rapport aux données traditionnelles ainsi que dans le Web social.

[1] Cette logique verticale de communication à double sens est d'ailleurs très fréquente dans le cas de la gestion des catastrophes naturelles.

[2] En 2018, Twitter est la 5ème plateforme de réseaux sociaux la plus populaire parmi les internautes, avec 330 millions d'utilisateurs mensuels actifs enregistrés dans le monde ; la plateforme reste loin derrière les milliards d'utilisateurs de Facebook et Youtube. Source : https://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2018/ (Consulté pour la dernière fois le 05/03/2019).

2.1. Une plateforme du Web social dédiée à la création et au partage d'informations

La plateforme Twitter étant ancrée dans la logique du Web social, tout individu peut rapidement devenir membre de la communauté et créer de l'information, toujours sans besoin de compétences professionnelles spécifiques : il suffit de disposer d'un outil TIC (smartphone, ordinateur ou tablette) et d'une connexion à Internet. Cette première partie présente les grands principes du fonctionnement technique de la plateforme, qu'il s'agisse de l'internaute utilisateur du service ou du développeur. Elle expose ainsi les modes de communication entre utilisateurs et les différents types d'informations véhiculées par les tweets d'une part, et d'autre part, elle décrit les outils de recherche et d'extraction d'informations mis à disposition du public. Pour autant, il faut noter que Twitter constitue avant tout une plateforme du Web social et non du Géoweb social : l'objectif originel d'un tweet ne consiste pas à représenter un objet du territoire, ni à figurer sur une carte[3].

2.1.1. La communication comme premier principe de fonctionnement

2.1.1.1. Qu'est-ce qu'un tweet ?

Lancé en 2006, Twitter est une plateforme de *microblogage*, via laquelle les membres de la communauté rédigent et diffusent de courts messages d'information[4], limités à 140 caractères jusqu'en novembre 2017, et désormais à 280 caractères[5]. L'activité tweeting est ainsi conçue comme un système de *microblogage,* c'est-à-dire un ensemble de messages brefs et concis renseignant les activités du moment, l'opinion de l'utilisateur ou diffusant un message d'information auprès du public[6]. Le tweet constitue alors un matériau multidimensionnel incluant fréquemment (figure 2.1) :

- un contenu sémantique de forme hétérogène : il peut mêler texte, image PNG ou JPEG, image animée GIF, émoji, vidéo, lien URL redirigeant vers une page web tierce, etc. ;
- le nom associé au compte de l'utilisateur à l'origine du tweet : ce nom est identifié par le signe @ qui le précède ;

[3] Certains tweets peuvent néanmoins décrire un objet du territoire.
[4] A la différence des blogs traditionnels, qui consistent à publier des articles plus conséquents mais moins fréquemment mis à jour.
[5] Source : https://www.lemonde.fr/entreprises/article/2017/11/07/twitter-generalise-les-messages-en-280-caracteres_5211616_1656994.html (Consulté pour la dernière fois le 05/03/2019)
[6] Source : https://fr.wikipedia.org/wiki/Twitter (Consulté pour la dernière fois le 05/03/2019)

- un *timestamp* (ou horodatage) : chaque tweet est renseigné par sa date et son horaire d'émission ; par défaut, l'horodatage affiché avec le tweet correspond au fuseau horaire UTC-8 (fuseau Pacifique des Etats-Unis).

Figure 2.1 : Exemple d'un tweet au contenu hétérogène, relatif à la tempête Freya, publié le 05/03/2019

- le tweet peut éventuellement inclure une mention d'utilisateur (figure 2.2) : lorsqu'un utilisateur souhaite attirer l'attention d'un destinataire particulier, il lui suffit d'inscrire le nom d'utilisateur associé au compte destinataire dans son tweet avant d'ajouter le reste de son message ; celui-ci sera alors notifié et libre de répondre au compte émetteur ;

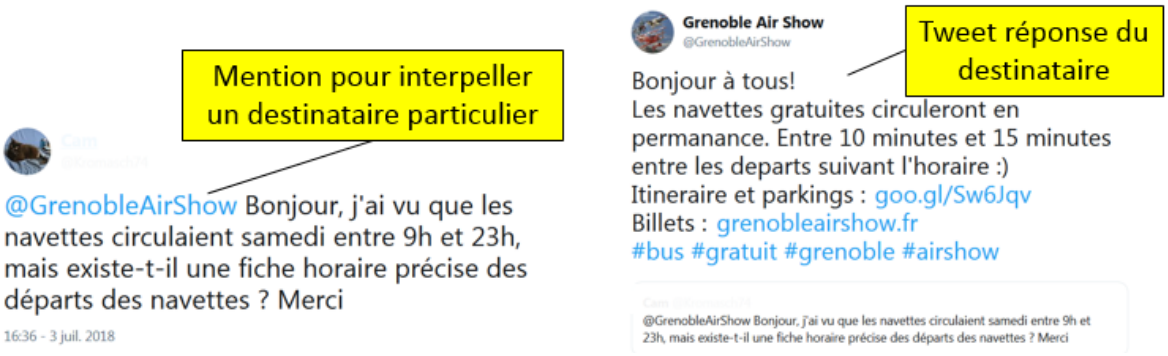


Figure 2.2 : Exemple d'un tweet adressant une demande de renseignements auprès d'un destinataire particulier et de sa réponse

2.1.1.2. Principes de communication et de diffusion des contenus

Un utilisateur enregistré sur la plateforme Twitter peut s'occuper à deux types d'actions : soit consulter de l'information en fonction de thématiques ou d'utilisateurs qui le concernent, soit créer de l'information.

Créer de l'information.

L'utilisateur qui souhaite créer un nouveau tweet se rend directement sur la page d'accueil de son profil. Pour publier, il lui suffira de cliquer sur le bouton *Tweeter*, de taper son texte, d'ajouter éventuellement une photographie, un GIF animé ou, fonctionnalité mise en service en 2015, de créer un sondage (une seule question pour deux réponses possibles mais fermées) en vote anonyme pour une durée maximale de sept jours.

Consulter de l'information.

Avant de tweeter, l'utilisateur peut également prendre connaissance des sujets signalés comme tendances au moment de sa connexion, c'est-à-dire les sujets qui drainent à ce moment-là des flux de tweets conséquents. Ces sujets s'identifient généralement par les *hashtags :* un hashtag est un mot-clé (ou une série de mots-clés accolés) permettant à un utilisateur de recontextualiser un tweet par rapport à une thématique précise, un sujet d'actualité ou encore à un événement particulier (Lin *et al.*, 2013). Le hashtag est identifié par le signe # qui le précède. Tout compte, qu'il soit associé à un simple individu, à une institution ou encore à un média, peut créer des hashtags au quotidien ; tout hashtag créé peut être réutilisé et diffusé par n'importe quel membre de la communauté, qu'il soit simple utilisateur dans sa sphère privée, acteur d'un service public ou qu'il représente une association. Sur la figure 2.3, l'utilisateur prend connaissance des tendances virtuelles observées dans les tweets émis en France pour la journée du 5 mars 2019 (partie gauche de la figure).

Figure 2.3 : Exemple des tendances et de tweets contenant le hashtag #MardiGras émis depuis les comptes de divers acteurs de Twitter

En cliquant sur le hashtag *#MardiGras*, il applique un filtre de recherche lexicale fondée sur la présence de ce hashtag dans les tweets, qui sont alors retournés et visibles sur la plateforme (partie droite de la figure 2.3). Parmi ces tweets, on distingue des messages émis par des individus inscrits sur Twitter à titre privé (et dont les propriétés du profil ont été volontairement masquées), mais également un acteur d'un service public (Ina) et un acteur représentant le monde associatif (L214). De nouveau sur la partie gauche de la figure 2.3, un clic sur le lien "*Modifier*" permet d'actualiser les tendances en fonction de diverses échelles géographiques sélectionnables par l'utilisateur – monde entier, pays, métropole (on pourra indiquer par exemple, les tendances enregistrées autour de Genève ou de Lyon, mais Grenoble ne fait pas partie des villes proposées). Si l'utilisateur souhaite publier un tweet contenant un hashtag ou un sujet tendance, il lui suffira de cliquer sur le mot en question.

De plus, si un utilisateur est abonné à d'autres comptes du réseau, il peut prendre connaissance des tweets publiés par ces comptes et les *retweeter*, c'est-à-dire rediffuser depuis son compte un message déjà émis par un compte précédent.

Une communication en réseau.

La diffusion des messages fonctionne en réseau : tout membre de la communauté disposant d'un compte est libre de s'*abonner* aux comptes d'autres membres, auquel cas l'utilisateur abonné est *notifié* sur les tweets émis par les comptes auxquels il s'est abonné. L'abonnement est généralement libre : si un utilisateur B souhaite s'abonner au compte d'un utilisateur A, il n'a pas besoin de son autorisation pour suivre l'activité de son compte. Ces usages inhérents au modèle du web social soulèvent ainsi la question des droits de propriété intellectuelle : en pratique, ils paraissent, sur Twitter, quasi inexistants : tout tweet créé par un utilisateur A peut être recopié, intégralement ou partiellement, par l'utilisateur B (qui est abonné au compte A et peut donc rapidement lire les messages associés à ce compte) ou par l'utilisateur C, qui lit le tweet de l'utilisateur A mais qui n'est pas abonné à son compte (ces deux utilisateurs, B et C risquent même de déformer, d'interpréter, voire de diffamer l'utilisateur A). En outre, ces mêmes utilisateurs B et C peuvent effectuer un *retweet* de l'utilisateur A, c'est-à-dire qu'ils utilisent une fonctionnalité automatique pour diffuser auprès de leur propre réseau d'abonnés le tweet de l'utilisateur A ; dans ce cas précis, l'auteur d'origine est mentionné. Choisir un compte privé constitue l'option permettant de pallier *a minima* les risques liés à la diffusion rapide et non régulée de l'information : tout utilisateur disposant d'un compte privé *autorise* l'abonnement d'autres membres à son compte et ses tweets seront uniquement visibles pour ses utilisateurs abonnés. Néanmoins, le maître mot du réseau restant l'ouverture, une majorité de tweets appartiennent ainsi au domaine public et sont, par conséquent, visibles par tout internaute quand bien même celui-ci ne dispose pas de compte sur la plateforme. D'ailleurs, il arrive que certains articles de quotidiens nationaux,

publiés en ligne, diffusent des revues de tweets sélectionnés pour afficher les opinions du public ou son ironie sur une question environnementale, sociale ou politique[7].

2.1.1.3. Proposition de classification des tweets

L'utilisation de la plateforme et le contenu des tweets sont d'autant diversifiés et hétérogènes qu'il existe différents profils de membres animés par des motivations plurielles. (Java *et al.*, 2007) ont établi la première typologie des pratiques de communication rencontrées sur le réseau en fonction du sens des tweets et de leurs usages : le *daily chatter* faisait ainsi référence à une information relative à une activité en cours ; l'*information sharing* concernait le partage d'informations par l'échange d'URL renvoyant à diverses sources de médias ; enfin, l'usage *conversation* qualifiait les messages portant des mentions (le @ précédant un nom d'utilisateur indiquant toujours l'existence d'un destinataire particulier) témoignant de discussions entre un groupe d'utilisateurs. Ces trois premières catégories pouvaient être qualifiées d'information *routinière* de fond qui circule quotidiennement sur le réseau (Goodchild & Glennon, 2010). Enfin, le *news reporting* caractérisait les utilisateurs rapportant ou commentant des événements en cours.

Les dernières années écoulées ayant été abondantes en événements naturels ou sociaux, on se propose ici de revisiter cette typologie[8] en fonction du profil des tweets publiés sur la plateforme d'une part, et d'autre part, des différentes logiques de communication qu'on peut mettre en exergue :

- la *diffusion d'un contenu personnel, centré sur l'individu* membre de la communauté. Le tweet est créé par un utilisateur dans le cadre de la sphère privée : il est centré sur ses intérêts propres. Il peut alors être simplement destiné à exprimer une action quelconque (*je suis en train de faire, voir, écouter*, etc.) ou à faire parler de soi sur le réseau. La communication sur Twitter pouvant également être utilisée dans un but de planification, à la manière des SMS, il peut également servir à interpeller un ou des destinataires particuliers afin de partager des informations précises.

- *la diffusion d'un contenu focalisé sur des événements exceptionnels* : le message peut toujours émaner d'un individu (ou du représentant d'un groupe) mais ne concerne pas ses activités, intérêts ou attentes propres. Ce message est centré sur un élément du monde réel qui perturbe la routine du quotidien et suscite une réaction de la part de l'utilisateur qui va alors le commenter et exprimer son opinion : ces tweets témoignent alors de l'engagement d'une dynamique collective sur le réseau. Ces dynamiques concernent généralement des événements ayant la capacité d'aiguiser la motivation à tweeter et contiennent un hashtag

[7] Le site de France 3 régions ou du Huffingston Post en proposent sur les sujets variées :
https://france3-regions.francetvinfo.fr/decouverte/internet/revue-de-tweets
https://www.huffingtonpost.fr/news/revue-de-tweets/

[8] Cette typologie ne prend pas en compte les spams et annonces publicitaires émis par des comptes automatiques.

permettant d'identifier l'événement en question : une rencontre sportive (*#IREFRA* pour le match de rugby entre l'Irlande et la France du tournoi des Six Nations de l'hiver 2019), un événement politique ou social qui soulève un débat dans l'opinion publique (*#ActeXVII* pour la 17ème journée de mobilisation du mouvement des gilets jaunes), les difficultés éprouvées en période de grève (*#GrèveSNCF* pendant la grève des cheminots au printemps 2018), des émissions télévisées qui invitent les spectateurs à twitter (*#Cdanslair* pour l'émission politique quotidienne de France 5), ou encore tout événement de l'actualité (*#Boeing737* lors du crash de l'avion de la compagnie Ethiopian Airlines en mars 2019).

- *la diffusion d'un contenu destiné à engager une dynamique dans le monde réel* : dans cette troisième catégorie, le tweet peut être rattaché à l'intérêt propre d'un utilisateur ou présenter un intérêt collectif. Il peut provenir d'un seul individu mais incite à une action dépassant le cadre de la conversation virtuelle : un utilisateur peut informer, interroger, voire lancer une alerte via le réseau. Ces tweets sont davantage destinés à inciter une dynamique dans le monde réel qu'à simplement commenter une situation ou un événement particulier.

Cette proposition doit ensuite être complétée en s'appuyant sur un autre facteur : le sens de la communication, qui se révèle plus complexe à saisir : par exemple, un tweet centré sur les activités d'un individu peut s'avérer insignifiant pour les autres membres de la communauté dans la mesure où le contenu tweeté n'est d'aucune utilité pour un tiers. Pour autant, ce même type de tweet peut parfois dépasser le seul intérêt de son auteur et concerner tout autant d'autres membres du réseau (c'était le cas du tweet présenté dans la figure 2.2 "*@GrenobleAirShow Bonjour, j'ai vu que les navettes circulaient samedi entre 9h et 23h, mais existe-t-il une fiche horaire précise des départs des navettes ? Merci*" ; la réponse fournie à l'utilisateur d'origine a été retweetée deux fois et a obtenu quatre mentions *J'aime*, indiquant ainsi que la question exprimée par un seul utilisateur dépassait son intérêt et concernait un ensemble d'individus ayant apprécié ces informations complémentaires). Nous avons alors défini quatre logiques de communication en fonction de l'origine et du contexte d'émission des tweets : la logique *horizontale*, les logiques *top-down* et *bottom-up,* et pour finir, la logique *multiple*. Ces logiques sont illustrées et détaillées en page suivante, dans le tableau 2.1 qui fournit des exemples de tweets[9] ancrés dans ces logiques et ces différents profils d'activité, ainsi que dans les explications qui accompagnent le tableau.

[9] Dans le tableau 2.1, le texte tapé correspond au texte original des tweets ; par conséquent, les éventuelles fautes de grammaire et d'orthographe sont conservées. Les numéros indiqués entre parenthèses avant les tweets permettent de les identifier dans les explications accompagnant le tableau.

Tableau 2.1 : Exemples de tweets diffusant différents types de contenus et ancrés dans des logiques de communication variées

Type d'information / Destinataire	Diffusion dans le cadre de la sphère et d'activités privées	Commenter un phénomène de société ou événement	Engager une dynamique dans le monde réel
Compte destinataire ciblé	*(1) "@... Bon anniversaire !!!!!!!!!!"*	*(4) "@Tag_Grenoble Ligne B"* *(5) "Ligne B manifestation régionale gilets jaunes Depuis 8h00 – durée indéterminée La ligne ne circule pas entre les stations Sainte-Claire et Gares."*	*(9) "Joyeux anniversaire @... on se retrouve ou tu sais pour une séance de shopping. Love"*
Destinataire non ciblé - Réseau	*(2) "Mon #MardiConseil "Au Revoir Là-Haut" de Pierre Lemaître. Un récit sombre des ravages psychologiques, humains et sociaux de la guerre."* *(3) "c'est qui le chanceux qui doit aller aux courses alors qui vient juste de rentrer de soirée ?"*	*(6) "Le fait d'en arriver à un Grand débat n'est ce pas la preuve flagrante de l'échec de nos politiques, parlementaires qui ne sont plus à l'écoute de leur peuple ?! #Cdanslair"* *(7) "Le stand de proximité du #GrandDébat est à la gare de #Grenoble pendant 2 jours le 6 et 7 mars, profitez-en pour partager vos propositions d'avenir pour la #France : #Etat, #TransitionEcologique, #Fiscalité, #Démocratie"* *(8) "Le #MardiGras du village : avant le défilé, le spectacle, le concours, et on fait brûler Monsieur Carnaval..."*	*(10) "Les gens d'abbeville si vous voyez un chat noir avec des tâches blanche vous pouvez me le dire svp, je suis rentree ce matin et mon chat était plus la, elle est pas habituer à etre dehors..."* *(11) "Le #Bas_Rhin est en #VigilanceOrange pour #VentViolent de 12h à 18h aujourd'hui. Evitez les déplacements, notamment en forêt, et soyez vigilants aux chutes d'objets. Plus d'informations sur le site de la préfecture du Bas-Rhin."*

Légende des couleurs : logique horizontale / logique top-down / logique bottom-up / logique multiple

Dans une logique *horizontale*, l'utilisateur est un individu qui tweete vers un destinataire particulier de la société civile ou vers l'ensemble du réseau, généralement de manière impersonnelle et à titre privé : les tweets (1) et (9) sont deux exemples de conversation entre deux utilisateurs ; le tweet (9) motive cependant une action directe amorcée sur le réseau et qui sera concrétisée dans le réel. Les tweets (2) et (3) n'ont aucun destinataire particulier : alors que le tweet (3) ne présente aucun intérêt pour un tiers (il témoigne de l'aspect égocentrique fréquemment critiqué des réseaux sociaux), le tweet (2), bien qu'il soit centré sur l'activité du moment de son auteur, peut intéresser un autre utilisateur. De la même manière, le tweet (10), bien qu'il soit centré sur une requête privée, interpelle directement un ensemble d'usagers. En revanche, le tweet (8) décrit un événement local sans mention particulière (et on ne peut prédire sa popularité parmi les utilisateurs). Dans une logique *bottom-up*, l'utilisateur est un individu faisant remonter une information vers des acteurs de la société : il peut ainsi interpeller un acteur public ou privé via son compte Twitter. Par le tweet (4), l'utilisateur souhaite savoir si des perturbations sont signalées sur la ligne de tramway qu'il doit emprunter, vraisemblablement afin d'organiser son trajet. La logique *top-down* implique que le tweet soit créé par un acteur public ou privé de la société qui fait descendre de l'information auprès des utilisateurs de la société civile : le tweet (5) constitue la réponse à la question formulée du tweet (4) ; les tweets (7) et (11) proviennent de comptes

associés à des préfectures départementales et diffusent des informations d'intérêt collectif auprès des citoyens membres de la plateforme. Certaines dynamiques peuvent néanmoins croiser ces différents sens de communication et construire une logique de communication *multiple* : le tweet (6) est créé par un téléspectateur qui réagit à une émission télévisée invitant explicitement son public à discuter de sujets politiques et sociaux via l'utilisation d'un hashtag particulier : la dynamique engagée fait alors remonter des opinions répondant à une proposition d'acteurs officiels par des messages qui peuvent être rediffusés et commentés entre utilisateurs. De la même manière, les tweets marqués en vert (5), (7) et (11) peuvent, en fonction de leur popularité sur le réseau, s'inscrire dans cette logique multiple : un utilisateur abonné aux comptes émetteurs qui considère que l'information donnée est importante et doit être partagée peut retweeter l'un de ces messages officiels, ajouter son propre commentaire, ou répondre directement au compte émetteur.

Malgré tout, certains types de tweets restent difficiles à classer. Les tweets publiés en réaction à des événements violents et/ou traumatisants (attaques terroristes, incendie de Notre-Dame de Paris) peuvent être inclus dans l'ensemble des catégories : la dimension émotionnelle s'inscrit dans une logique horizontale et collective ; d'un autre côté, une pléthore de tweets sont publiés par divers acteurs dans des logiques de débat (débat sur la cohésion nationale, la gestion de la crise après les attentats de Paris en novembre 2015 ou sur le financement de la restauration des monuments historiques) ou d'engagement de dynamiques dans le monde réel (appel aux marches et rassemblements citoyens après les attentats de janvier 2015, appel aux dons). De la même manière, les tweets relatifs aux rumeurs peuvent s'inscrire dans l'ensemble des catégories et logiques de communication décrites : en février 2018, des utilisateurs de Twitter aux Etats-Unis reçoivent des messages d'alerte au tsunami émis depuis différents comptes Twitter associés à la NOAA. Vraisemblablement, les messages ont été trop rapidement lus : il s'agissait alors d'un test mais les usagers ont massivement réagi sur le réseau pour demander la confirmation de ces alertes. La figure 2.4 montre un ensemble de tweets qui témoignent des acteurs et des sens de communication impliqués : on retrouve ainsi de simples individus pouvant s'adresser au réseau dans une logique horizontale ou aux acteurs institutionnels dans une logique ascendante et ces mêmes acteurs répondant à l'ensemble du réseau dans une logique descendante.

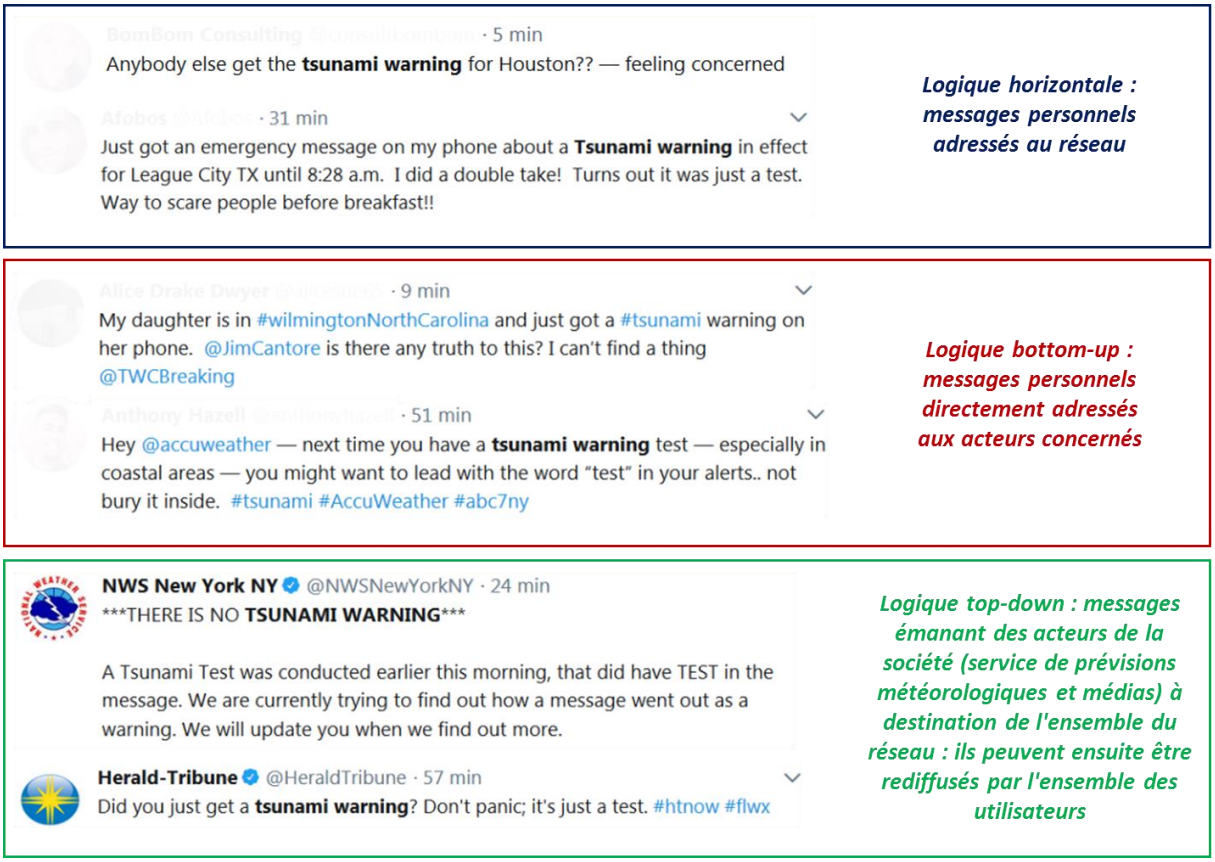


Figure 2.4 : Exemples de tweets relatifs à la diffusion d'une alerte test au tsunami sur les villes côtières des Etats-Unis en février 2018.

2.1.2. La dimension spatiale des tweets

2.1.2.1. Déclinaison des types de géoréférencement des tweets

Le nombre de tweets émis au quotidien est estimé à cinq cents millions, soit un total de trois cents milliards de tweets émis dans le monde entier (à l'exception des pays bloquant l'accès à la plateforme depuis un certain nombre d'années : Chine, Corée du Nord, Vietnam) entre 2006 et 2018[10]. Une partie de ce flux comporte, en plus de son contenu sémantique et de son horodatage, une information permettant de connaître la localisation du lieu d'émission du tweet. Tout comme le contenu sémantique du message, cette composante spatiale peut se décliner sous diverses formes et divers degrés de résolution spatiale[11] :

- Une entité spatiale peut apparaître sous la forme d'un toponyme ajouté à la composante sémantique du tweet, en fonction de divers degrés de précision et sous

[10] Source : https://www.blogdumoderateur.com/chiffres-twitter/ (Consulté pour la dernière fois le 05/03/2019)
[11] La résolution spatiale est ici considérée selon la granularité de l'information, c'est-à-dire son niveau de détail en termes de précision spatiale pour localiser l'information : pays, ville, quartier, coordonnées GPS.

différentes formes : nom de ville, nom d'un quartier ou d'une rue ou d'un objet précis du territoire (monument, musée, stade, salle de spectacle, etc.) ; nom qui peut être tapé sous la forme d'un hashtag ou apparaissant comme un simple mot du tweet. Il s'agit d'une pratique qualifiée de *Géotagging*. Dans le tweet de la figure 2.1, cette localisation sémantique était précise : *"#Tempête #Freya – Océan démonté hier à la Pointe de la #Torche à Plomeur (#Finistère, #Bretagne) – Rémi Lemenicier*[12] *:* [URL de la photographie]" ; ce tweet décline les informations sémantiques de localisation selon plusieurs niveaux de granularité : hashtag de la région (*#Bretagne*), hashtag du département (*#Finistère*), nom de la commune (*Plomeur*), nom et hashtag du lieu depuis lequel la photographie a été capturée (*Pointe de la #Torche*). En outre, grâce au point de vue offert par la photographie, l'internaute qui dispose d'une familiarité suffisante avec ce territoire pourra identifier le lieu précis depuis lequel la photographie a été prise. *A contrario*, les tweets de la figure 2.2 ne contiennent pas d'information de localisation, ou trop peu précise et biaisée : *"@GrenobleAirShow Bonjour, j'ai vu que les navettes circulaient samedi entre 9h et 23h mais existe-t-il une fiche horaire précise des départs des navettes ? Merci*", "*Bonjour à tous ! Les navettes gratuites circuleront en permanance. Entre 10 minutes et 15 minutes entre les departs suivant l'horaire ☺Itineraire et parkings :* [URL vers la carte] *Billets : grenobleairshow.fr #bus #gratuit #grenoble #airshow*". L'événement en question, le meeting aérien *Grenoble Air Show*, fait directement référence à la ville de Grenoble, y compris dans le hashtag *#grenoble* mentionné par le compte officiel de l'événement. En réalité, l'événement a lieu à l'aérodrome du Versoud, soit à une douzaine de kilomètres en amont de Grenoble. Seule la consultation du lien fourni dans le tweet indiquera l'adresse précise du lieu en question.

- si l'utilisateur active, dans les paramètres de son compte, l'option intitulée *Tweeter avec une localisation*, il peut fournir, lorsqu'il tape son tweet, une donnée relative à sa localisation plus précise, qui peut prendre deux formes en fonction du support numérique d'émission :

  - si l'utilisateur crée un tweet depuis un ordinateur connecté à Internet : la localisation est estimée via le point d'accès Wi-Fi et apparaît sous la forme de l'ajout d'une *place* : Twitter soumet le nom de plusieurs échelons géographiques (de la commune au pays) à l'internaute qui peut indiquer celui de son choix. L'information spatiale indiquée apparaît sous forme d'une *bounding box*, soit un rectangle qui délimite grossièrement l'emprise spatiale de l'échelon géographique ajouté (figure 2.5).

[12] Le photographe en question étant professionnel et son nom étant directement inscrit dans le tweet et par conséquent visible par tous, nous ne l'avons pas masqué (contrairement aux informations identifiant les profils des utilisateurs actifs dans leur sphère privée).

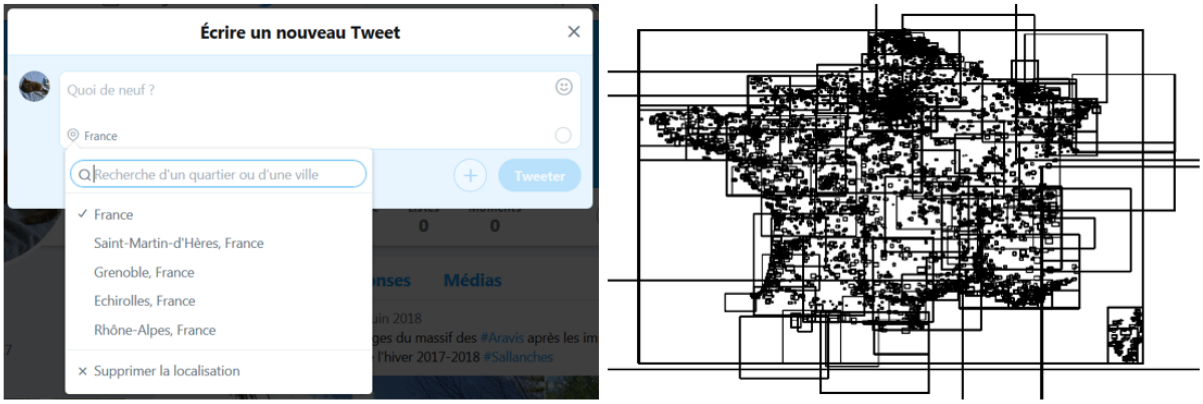


Figure 2.5 : Ajout d'une localisation de type place à un tweet créé par ordinateur et exemple de carte de tweets géolocalisés par bounding box sur la France

- si l'utilisateur crée un tweet depuis un smartphone équipé d'un récepteur GPS, sur lequel est installé l'application Twitter : l'ajout d'une localisation au tweet active le GPS de l'appareil. L'utilisateur a alors deux choix : soit il sélectionne l'une des adresses listées proposées par l'application (qui lui en soumet un certain nombre en fonction des coordonnées GPS fournies par le smartphone), soit il indique directement les coordonnées exactes de l'appareil (figure 2.6, NB : les adresses sont volontairement masquées). De tels tweets apparaissent alors sous forme de points.

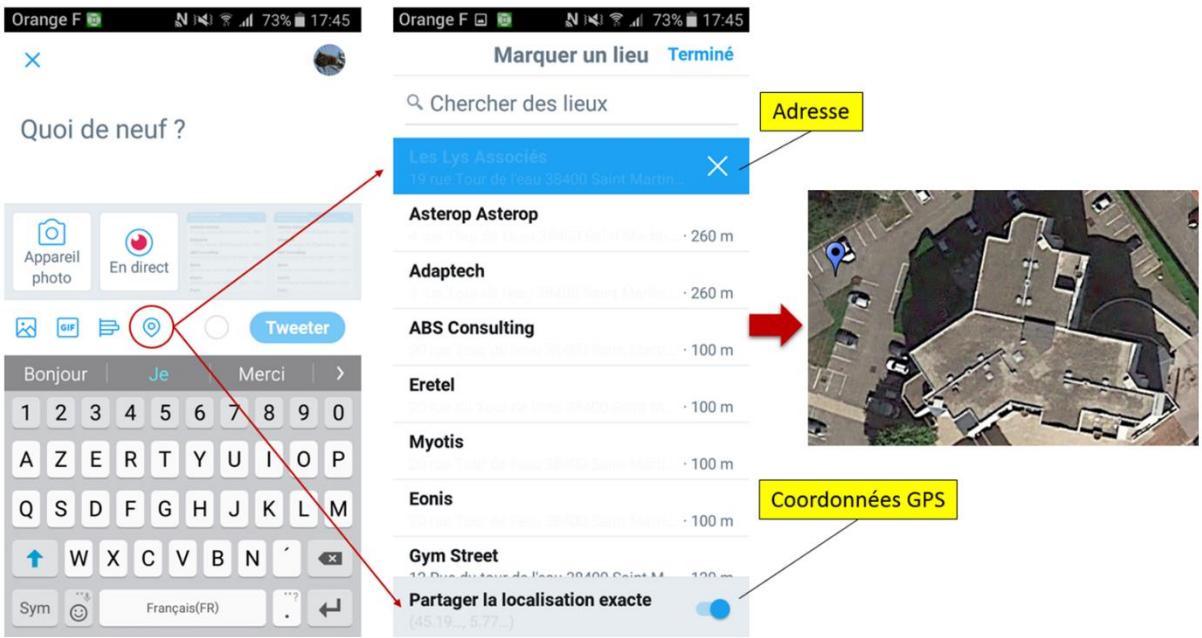


Figure 2.6 : Création d'un tweet géolocalisé par coordonnées GPS envoyé depuis un smartphone

L'usage de cette fonctionnalité présente néanmoins un inconvénient majeur : c'est l'utilisateur qui choisit les modalités de sa géolocalisation : en conséquence, tout tweet géolocalisé est soumis à une incertitude spatiale qui dépend de choix individuels dont le chercheur ne connaît pas les motivations. On peut d'ores-et-déjà envisager un certain nombre de facteurs définissant cette incertitude spatiale :

- dans la figure 2.5 (page précédente), on voit que l'utilisateur peut sélectionner diverses échelles géographiques de localisation. Ces échelles se remarquent nettement sur la carte juxtaposée au tweet : certains rectangles englobent le pays entier, d'autres une région, un département ou, au plus précis, une ville ou un quartier (à condition de zoomer).
- dans la figure 2.6, l'utilisateur peut sélectionner une adresse proposée dans la liste ou fournir ses coordonnées GPS exactes : si l'utilisateur ne souhaite pas donner son adresse (et en particulier s'il s'agit du domicile), il peut volontairement biaiser la localisation en choisissant un lieu voisin. Si toutefois il choisit d'indiquer ses coordonnées exactes, comme dans l'exemple de la figure, on doit encore prendre en compte l'incertitude liée à l'instrument GPS : ici, les coordonnées indiquées par l'appareil marquent l'emplacement du tweet à une vingtaine de mètres de la position réelle de l'utilisateur au moment de l'émission.
- l'exemple de la figure 2.7 ci-après soumet une autre manière d'influencer la géolocalisation des tweets : l'ordinateur à l'origine du tweet se trouve physiquement sur la commune de Saint-Martin-d'Hères en Isère. L'application identifie et liste une série de communes susceptibles de convenir à l'utilisateur. Or l'information donnée et les photographies acquises concernent un lieu différent de celui où se trouvent l'utilisateur et son ordinateur (l'action de tweeter est décalée spatialement et temporellement par rapport à l'action d'acquisition de l'information). Au-delà des lieux proposés par Twitter, l'utilisateur peut rechercher et sélectionner manuellement une commune. S'il choisit le lieu approprié, la localisation est juste mais ce n'est pas du temps réel ; de la même manière, si l'utilisateur choisit le lieu de sa présence physique au moment de l'émission du tweet, on sort également du temps réel et la localisation est fausse par rapport au contenu.

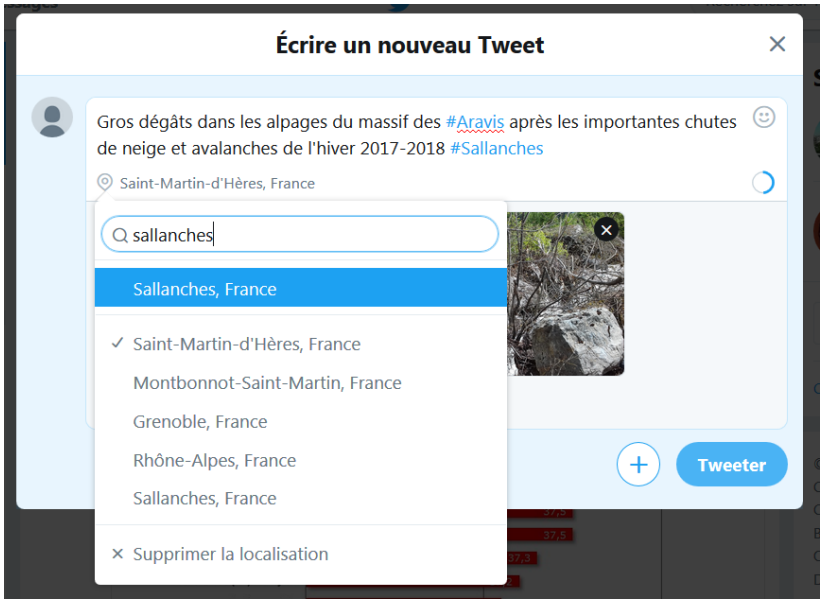


Figure 2.7 : Choix manuel de la localisation d'un tweet non émis en temps réel

Si 80% des utilisateurs de Twitter utilisent leur smartphone pour créer de l'information ou naviguer sur la plateforme[13] (et ils seraient quasiment 85% aux Etats-Unis[14]), le flux de tweets comportant une information de localisation ajoutée à partir de la fonctionnalité de géolocalisation par GPS ne représente qu'une proportion minime du flux global drainé quotidiennement sur le réseau : (Brovelli *et al.*, 2014) estiment que la proportion de tweets émis avec des coordonnées longitude/latitude varie entre 1% et 2% du flux total de tweets émis quotidiennement. Il s'avère en revanche difficile d'obtenir des chiffres plus précis ou plus récents : pour (Hawelka *et al.*, 2013), il s'agissait plutôt de 1% du flux total (le chiffre est de nouveau cité par [Mericskay *et al.*, 2018]). Chez (Sloan & Morgan, 2015), cette proportion diminuait à 0,85%. En conséquence, le chercheur ne dispose que d'une infime partie des données qui circulent sur la plateforme, ce qui soulève un certain nombre de biais en termes de représentativité ainsi qu'un risque de surinterprétation de résultats qui ne reflètent qu'une maigre part du terrain (Mericskay *et al.*, 2018).

2.1.2.2. Les enjeux de la géolocalisation des tweets

L'ajout d'une information de localisation par l'intermédiaire de la fonctionnalité appropriée n'est pas un fait accompli aux dépens de l'utilisateur. En effet, comme l'indiquaient les figures 2.5 et 2.6, si cette composante spatiale figure dans un tweet, sous forme de nom de ville ou de coordonnées GPS, c'est que l'utilisateur a validé les deux étapes indispensables :

(1) l'option *Tweeter avec une localisation* est activée ;

[13] Source : https://www.blogdumoderateur.com/chiffres-twitter/ (Consulté pour la dernière fois le 06/03/2019)
[14] Source : https://www.statista.com/statistics/293764/number-of-smartphone-twitter-users-in-the-united-states/ (Consulté pour la dernière fois le 06/03/2019)

(2) lorsque l'utilisateur est en train de tweeter, il a cliqué sur le bouton approprié pour autoriser la recherche de son emplacement.

Ce comportement reflète ainsi un acte volontaire et conscient à l'instant où le tweet est émis (mais qu'en est-il dans sa portée *a posteriori* et de l'intrusion dans la sphère privée ?) qui reste le fruit d'une minorité de membres de la communauté. Pourquoi le font-ils et pourquoi acceptent-ils d'ajouter un instrument de suivi supplémentaire à leur quotidien avec les risques qu'il comporte ? Mais surtout, la part des utilisateurs choisissant de se géolocaliser est-elle représentative du réseau ou négligeable ?

Qui utilise la géolocalisation des tweets et pourquoi ?

(Graham *et al*., 2013) soumettaient l'hypothèse qu'au regard de la très faible proportionnalité de tweets géolocalisés, il était très peu probable que les utilisateurs de cette fonctionnalité soient représentatifs de l'ensemble des usagers de la plateforme. (Sloan & Morgan, 2015) testaient cette hypothèse et publiaient un article focalisé sur cette question précise de l'utilisation du service de géolocalisation par GPS ou de *géotagging* en fonction des caractéristiques démographiques des utilisateurs de Twitter (les données collectées pour cette expérience sont néanmoins réduites au Royaume-Uni pour le mois d'avril 2015). Ces auteurs ont alors mis en évidence deux phénomènes :

- il existe des facteurs démographiques[15] (genre, âge, catégorie socio-professionnelle) qui introduisent une variabilité (existante mais ténue) des profils d'utilisateurs activant la fonctionnalité de géolocalisation : les femmes seraient plus enclines à activer la géolocalisation ; les utilisateurs les plus jeunes (13-20 ans) la délaissent ; les individus qui diffusent des tweets dans le cadre de leurs activités professionnelles (pratique généralement associée à l'appartenance à une catégorie professionnelle du haut de la hiérarchie) affichent une tendance moindre à la géolocalisation que les individus occupant des postes impliquant moins de responsabilités. Néanmoins, les tests de significativité pratiqués ne se révélant guère concluants, les résultats ne peuvent pas être généralisés à la population entière des utilisateurs de Twitter (et l'hypothèse énoncée par [Graham *et al*., 2013] se confirmerait).
- un seul facteur introduit des différences statistiquement significatives dans l'usage des fonctionnalités de géolocalisation : la langue du tweet. Parmi les tweets géolocalisés collectés pour l'expérience, les langues les plus représentées sont le portugais, l'indonésien, le thaï et les langues austronésiennes, soit des langues présentes dans les Etats et ensembles régionaux qui arborent les plus forts taux d'utilisation de Twitter à l'échelle du globe[16].

[15] L'âge, le genre et la catégorie socio-professionnelle d'un utilisateur sont déterminés à partir d'algorithmes d'apprentissage développés par ces mêmes auteurs (cf. bibliographie de l'article cité dans ce paragraphe).

[16] En 2015, le Brésil et les Etats de l'Asie du Sud (Indonésie, Philippines, Malaisie, Thaïlande) font partie des pays présentant les plus forts taux de pénétration du média social Twitter. Source : https://www.statista.com/statistics/279539/twitter-reach-in-selected-countries/ (Consulté pour la dernière fois le 12/03/2019)

Ce défaut (ou absence ?) de représentativité statistique du tweet géolocalisé, qui reste une question peu étudiée dans la recherche en sciences humaines (Mislove *et al.*, 2011 ; Li *et al.*, 2013) sera également en question en ce qui concerne notre terrain d'étude dans les paragraphes suivants (notre sujet étant localisé en un territoire différent, et dans des temporalités différentes). Le second enjeu majeur de toute fonctionnalité de géolocalisation reste, quelle que soit la plateforme considérée, le risque d'intrusion dans la vie privée de l'internaute.

Enjeux éthiques.

Les tweets géolocalisés peuvent s'apparenter à la face subreptice du Géoweb[17] qui qualifie une tendance systématique et automatique de géolocalisation de l'ensemble des contenus web créés ou des traces de navigation laissées par les internautes, acquises à leurs dépens[18]. En 2010, lorsque Twitter a intégré la géolocalisation à son interface, cette fonctionnalité était en effet automatiquement activée à l'ouverture d'un compte. Désormais, les directives de la CNIL contraignent à une mise en conformité, quelle que soit la plateforme : la géolocalisation d'un contenu ne peut pas être activée sans l'autorisation de l'utilisateur. Comme indiqué précédemment, il s'agit alors d'un acte volontaire : les utilisateurs des réseaux et médias sociaux sont-ils pour autant conscients des impacts de la géolocalisation de leurs contenus ? En 2015, les déplacements d'un djihadiste originaire de Nouvelle-Zélande ont été suivis pendant plusieurs mois par l'intermédiaire des tweets qu'il postait régulièrement sans avoir désactivé la fonction de géolocalisation[19]. Quelques mois après son arrivée à la Maison Blanche, Donald Trump continuait à tweeter depuis son smartphone personnel : un internaute a ainsi réussi à télécharger l'ensemble des métadonnées relatives au compte personnel du Président, notamment le modèle de smartphone utilisé et l'ensemble des coordonnées GPS des tweets émis[20].

Sur le web, on peut trouver de nombreuses applications de recherche et d'affichage de tweets géolocalisés : l'application *Geosocial Footprints* permet de lancer une collecte, d'afficher, et éventuellement de télécharger (sous forme de fichier CSV) les tweets géolocalisés de tout membre de Twitter : il suffit de renseigner un nom d'utilisateur. L'application détermine également un niveau de risque lié à la géolocalisation du contenu. La figure 2.8 propose deux recherches sur des utilisateurs aux profils différents. L'utilisateur de gauche n'a émis que cinq tweets : trois d'entre eux contiennent une donnée de localisation de type *place* et un seul tweet est géolocalisé par ses coordonnées GPS. L'application lui montre

[17] Expression de Thierry Joliveau. Source : https://mondegeonumerique.wordpress.com/2010/06/24/le-geoweb-pour-les-nuls/ (Consulté pour la dernière fois le 12/03/2019)

[18] Ou par un consentement tacite puisque la plupart des internautes vont *accepter les conditions d'utilisations* des plateformes sans même les avoir lues.

[19] Source : https://www.bfmtv.com/international/un-jihadiste-repere-sur-twitter-grace-a-la-geolocalisation-855388.html (Consulté pour la dernière fois le 12/03/2019)

[20] Source : https://www.frandroid.com/android/applications/securite-applications/410088_la-geolocalisation-de-twitter-permet-de-suivre-donald-trump-a-la-trace (Consulté pour la dernière fois le 12/03/2019)

l'emplacement de ce tweet sur la carte, estime un risque élevé d'intrusion dans sa vie privée et lui propose de désactiver la fonction de géolocalisation. En ce qui concerne l'utilisateur de droite, l'application a collecté un total de 200 tweets (son seuil maximal) ce qui suggère une activité plus intense ; en revanche, cet utilisateur n'a manifestement jamais eu recours à la géolocalisation, que ce soit sous forme d'ajout d'une *place* ou de coordonnées GPS : sa localisation n'est donc pas visible sur la carte et le risque d'intrusion estimé est nul.

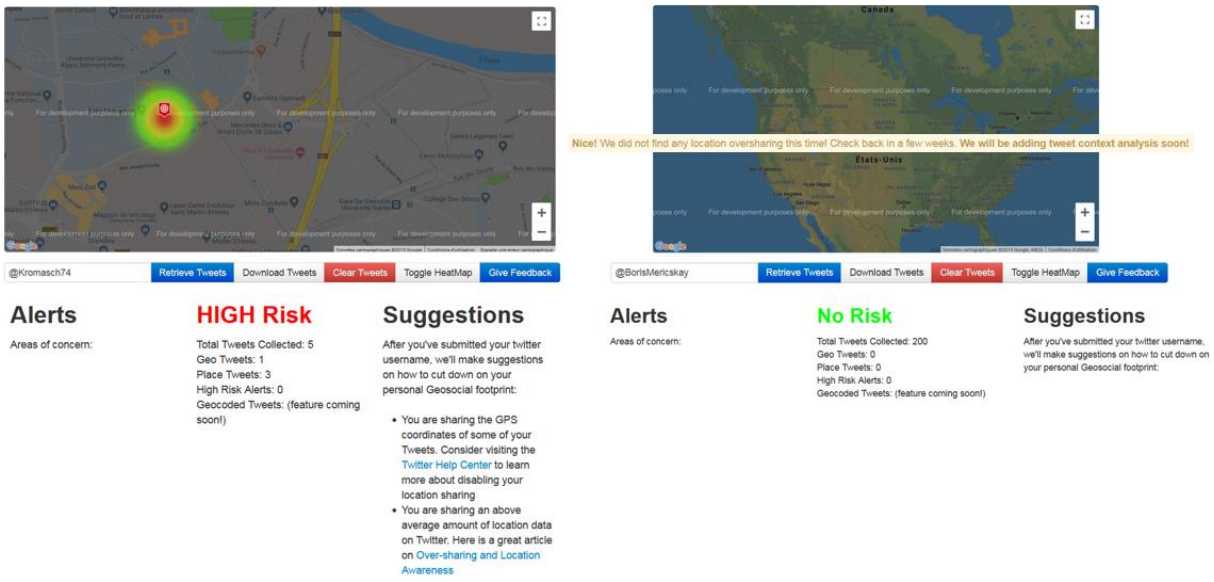


Figure 2.8 : Recherche de tweets par nom d'utilisateur sur l'application Geosocial Footprints

Bien que cette plateforme se présente comme didactique, elle reste intrusive dans son fonctionnement puisque *tout internaute* peut saisir le nom d'utilisateur de *tout membre* de la communauté Twitter et, si le membre en question utilise la géolocalisation par GPS, il peut même *télécharger son archive de tweets géolocalisés* par le bouton *Download tweets.*

Les enjeux éthiques liés à la géolocalisation concernent une multiplicité d'acteurs. En premier lieu, les propriétaires hébergeurs des plateformes, soit des entreprises privées dont la valeur repose sur ce capital de données mais que les institutions juridiques tentent de canaliser. Comme le rappelait (Quesnot, 2016), l'ensemble des contenus générés sur les plateformes de réseaux et médias sociaux sont avant tout des produits commerciaux appartenant aux *Géants du Web*, et vendus à d'autres sociétés privées partenaires qui les achètent et les analysent à des fins de publicité ciblée. L'ajout de fonctionnalités de géolocalisation révèle ainsi l'intention d'augmenter la valeur marchande de ces contenus. En France, bien que la loi contraigne les développeurs de plateformes et d'applications à désactiver par défaut la géolocalisation, la CNIL procède régulièrement à des mises en

demeure des sociétés développant des applications collectrices de données pour non-conformité à cette norme[21].

Pour le chercheur, les enjeux éthiques liés à la géolocalisation reposent sur deux points. La collecte de tweets géolocalisés constitue tout d'abord un risque d'intrusion dans la vie privée des utilisateurs : par la géolocalisation, il est possible d'estimer rapidement la localisation du domicile de tout utilisateur qui émet régulièrement des tweets (Kounadi *et al.*, 2015) ; dans un second temps, cette géolocalisation est complétée par une donnée rendant possible l'identification de chaque utilisateur[22]. En conséquence, tout utilisateur qui tweete abondamment sur le réseau prend non seulement le risque de divulguer l'emplacement de son domicile mais encore d'être suivi *à la trace* dans ses déplacements et activités. Le second enjeu qu'il convient d'exposer et qui sera discuté dans les chapitres de la seconde partie du manuscrit concerne une problématique d'ores-et-déjà exposée dans l'introduction : les usages sélectifs du numérique. L'ensemble des individus parcourant un territoire ne sont pas tous des usagers de Twitter ; ceux qui en font partie n'utilisent pas tous des fonctions de géolocalisation. Par ailleurs, nous avons vu qu'il était difficile de savoir quels profils d'individus se cachent derrière les utilisateurs qui se géolocalisent et surtout, que ces individus constituent une minorité (non représentative ?) de l'ensemble de la communauté. Construire une connaissance des lieux et de leurs usagers à partir de tels contenus géolocalisés, c'est alors prendre le risque de se focaliser sur des pratiques excluant l'ensemble des individus et des lieux non représentés sur le réseau.

Au-delà de l'usage expressif et parfois spatial de la plateforme, gravite un ensemble d'outils offrant à différents profils d'internautes diverses possibilités de recherche, d'extraction et de visualisation des tweets publiés sur le réseau.

2.2. Les outils du Web et du Géoweb associés à la plateforme Twitter

Ce deuxième axe inventorie les outils que l'internaute peut mobiliser sur le web afin de rechercher ou extraire des tweets répondant à des caractéristiques particulières. Dans un second temps, il présente une série d'outils apparentés au Géoweb, intégrant des fonctionnalités basiques de cartographie et d'analyse des tweets géolocalisés.

[21] Source : https://www.cnil.fr/fr/applications-mobiles-mise-en-demeure-absence-de-consentement-geolocalisation-ciblage-publicitaire-2 (Consulté pour la dernière fois le 13/03/2019)

[22] Tout utilisateur de Twitter dispose d'un numéro identifiant propre à son compte mais Twitter peut également livrer le nom d'utilisateur associé à tout compte. Si certains utilisateurs utilisent des pseudonymes, d'autres utilisent leur nom civil.

2.2.1. Outils à l'accès restreint aux utilisateurs de Twitter

2.2.1.1. Pour tout utilisateur : télécharger son archive de tweets

Tout utilisateur de la plateforme peut, à n'importe quel moment, adresser une demande de téléchargement de son archive personnelle de tweets (il suffit de passer dans les paramètres du compte) : cette archive contient l'ensemble des tweets qui ont été émis depuis le compte de l'utilisateur et propose un certain nombre de fichiers :

- des photographies incluant l'image utilisée par le profil de l'utilisateur ainsi que les images postées dans les tweets ;
- un ensemble de fichiers JS contenant des métadonnées décrivant : le compte de l'utilisateur, les appareils de connexion à Twitter, la liste des abonnés et des abonnements, les connexions éventuelles à des réseaux sociaux tiers, et un fichier contenant l'ensemble des tweets émis (intitulé *tweet.js*), décrits par leurs métadonnées complètes : numéro du tweet, présence de hashtags, de mentions d'utilisateurs ou encore d'URL, coordonnées géographiques si le GPS était activé, nombre de caractères, langue, horodatage, ainsi que des métadonnées indiquant la popularité du tweet sur le réseau : nombre de retweets, de mentions *J'aime*, de favoris ou de réponses. La figure 2.9 affiche un extrait du fichier *tweet.js* présent dans une archive d'utilisateur, ouvert sur le Notepad++ :

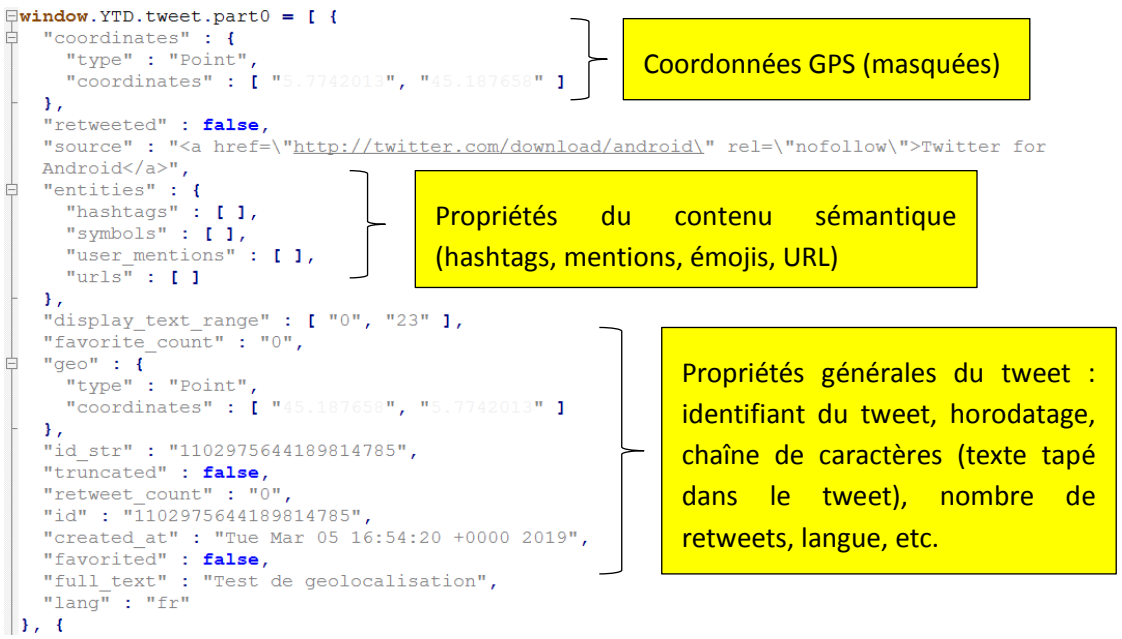


Figure 2.9 : Ensemble des métadonnées associées à un tweet géolocalisé, disponibles dans l'archive personnelle téléchargeable

2.2.1.2. Pour les développeurs : les API

Twitter dispose de deux API[23] principales accessibles aux développeurs et assurant la collecte de tweets : une *API Streaming* permet la collecte des tweets émis en temps réel sur le réseau et fonctionne selon deux flux :

- le flux *sample* fournit un échantillon de 1% du trafic mondial de tweets, sans application d'autres critères de sélection ;
- le flux *filter* permet de collecter des tweets en appliquant des critères sélectifs (géolocalisation, présence de certains hashtags).

L'*API Search* met à disposition du développeur les tweets qui ont été créés pendant les sept derniers jours précédents la connexion à l'API et l'exécution de la requête du développeur. Tout accès à une API requiert une seule condition préalable : le client doit disposer d'un compte Twitter et s'authentifier sur le site des développeurs de Twitter *developer.twitter.com* ; il peut alors demander la création d'une application en renseignant quelques champs : le nom de l'application, une brève description de sa thématique et de son objectif et si l'utilisation n'est pas personnelle, l'URL du site internet via lequel l'application sera disponible ainsi que le nom de l'organisme pour lequel il travaille. La validation de ces informations génère deux codes propres à l'application, une *consumer key* ainsi qu'un *consumer secret*, qui sont nécessaires pour la connexion aux API et le téléchargement de tweets.

Dans un second temps, toute connexion à l'API et soumission de requête client passe par des interfaces de programmation : par exemple, les langages python ou R disposent de librairies permettant au client de se connecter à une API et de lancer une requête destinée à télécharger des tweets filtrés en fonction de critères définis par le client. La figure 2.10 affiche un exemple de script rédigé avec les fonctions de la librairie *twitteR*[24] de R. Le client doit d'abord s'authentifier par les clés fournies lors de l'enregistrement de l'application sur le site des développeurs. La requête soumise ici demande un maximum de 50 000 tweets géolocalisés émis autour de la Nouvelle-Orléans, entre le 08/08/2017 et le 14/08/2017 et contenant au moins l'un des mots-clés ou hashtags suivants : *flood*, *floods*, *#flood* ou *#floods*. Les tweets collectés peuvent alors être formatés en un tableau puis injectés dans une base de données sur l'ordinateur du client.

[23] Une API est une interface de programmation applicative qui désigne une bibliothèque de fonctions offrant des services à d'autres applications ou logiciels. Source : https://fr.wikipedia.org/wiki/Interface_de_programmation (Consulté le 14/03/2019)

[24] Source : https://www.rdocumentation.org/packages/twitteR/versions/1.1.9 (Consulté pour la dernière fois le 14/03/2019)

```
# Chargement du package

library(twitteR)

# Authentification (nécessite d'avoir enregistré l'application sur twitter dev)

consumer_key <- [illegible]
consumer_secret <- [illegible]
access_token <- [illegible]
access_secret <- [illegible]

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

# Extraction des tweets sur la Nouvelle-Orléans

query <- searchTwitter(
                    "flood OR floods OR #flood OR #floods",
                    n=50000,
                    lang=NULL,
                    since='2017-08-08',
                    until='2017-08-14',
                    geocode='29.962526,-90.100890,32mi'
                                        )

# Convertir les résultats de la requête en dataframe

df.<- do.call("rbind", lapply(query, as.data.frame))
```

Clés de sécurité à saisir (fournies lors de l'enregistrement de l'application sur twitter.dev) masquées

Critères de la requête client : filtrage sémantique, effectif maximal de 50 000 tweets émis entre le 08/08/2017 et le 14/08/2017, toutes langues confondues et ajout d'une zone tampon d'extraction (32 miles autour du point aux coordonnées saisies)

Figure 2.10 : Script R de connexion et de requête à l'API Search de Twitter par le package twitteR

Comment les tweets sont-ils alors retournés au client ? La collecte gratuite est limitée par l'entreprise mais varie en fonction des API : le nombre de requêtes *API Search* par heure et par utilisateur est limité à 150[25]. En utilisant le flux *filter* de l'*API Streaming*, les tweets retournés au client peuvent être également soumis à un échantillonnage (si les critères de recherche indiqués par le client concernent un flux de tweets supérieur à la limite du 1%) ; néanmoins, si les critères de filtrage fournis par la requête du client sont suffisamment sélectifs (comme lorsqu'on demande uniquement des tweets géolocalisés), il échappe à cette contrainte d'échantillonnage[26].

Malgré tout, tout client connecté aux API, qu'il soit chercheur ou acteur privé, subit la dépendance aux politiques des Géants du Web en matière de redistribution des données (Quesnot, 2016) ; depuis 2018, Twitter a développé ses offres en matière d'API afin de promouvoir l'utilisation d'interface *premium* payantes acceptant davantage de requêtes et ouvrant l'accès à des flux plus conséquents[27].

[25] Source : https://www.blogdumoderateur.com/les-limites-de-twitter/ (Consulté pour la dernière fois le 14/03/2019)

[26] Si les tweets retournés au client sont échantillonnés, celui-ci sera averti par une notification.

[27] Source : https://www.zdnet.fr/actualites/twitter-presente-ses-api-premium-pour-les-developpeurs-39860034.htm (Consulté pour la dernière fois le 14/03/2019)

2.2.2. Outils de recherche et de suivi des tweets

2.2.2.1. L'interface Twitter de recherche avancée

L'entreprise Twitter dispose d'une interface de recherche avancée[28], accessible à tout internaute, qu'il soit membre de la communauté ou non, et qui permet de rechercher et d'afficher des tweets en fonction de critères de sélection. L'interface soumet un certain nombre de filtres qui trient les tweets en fonction des critères indiqués par l'internaute. Ce tri s'applique à tous les tweets créés et stockés par la plateforme, à l'exception des tweets émanant de comptes privés. L'utilisateur ou internaute peut ainsi rechercher et consulter des tweets en fonction (figure 2.11) :

- de la présence d'un certain nombre de mots-clés, d'un ou de plusieurs hashtags ;
- de la langue d'origine des messages ;
- de comptes d'utilisateurs émetteurs ;
- d'un intervalle temporel précis, à partir d'un calendrier soumettant des dates qui remontent jusqu'à l'envoi du premier tweet public ;
- de la proximité à un lieu : ville, région, pays (la requête retourne alors des tweets émis avec une composante spatiale).

Figure 2.11 : Tweets filtrés par l'interface de recherche avancée de Twitter (critère de sélection entrée par la requête client : #texasfloods

[28] Source : https://twitter.com/search-advanced?lang=fr

2.2.2.2. Les outils de suivi de la communication en réseau

Pour tout membre de la communauté, Twitter dispose d'une fonctionnalité de suivi du réseau personnel, intitulé *Statistiques,* accessible depuis le compte utilisateur : l'outil prend la forme d'un tableau de bord mensuel par lequel l'utilisateur peut suivre le nombre de vues associées à ses tweets publiés, le nombre de nouveaux abonnés, le nombre de tweets qu'il a publiés ou encore le nombre de visites de son profil par des internautes tiers. Pour l'utilisateur dans sa sphère privée, l'outil est bâti dans une logique de promotion de soi par la popularité sur le réseau ; pour les entreprises, il s'agit davantage d'identifier les attentes et intérêts d'un public potentiel.

En effet, aux côtés des simples fonctionnalités de recherche, des sociétés ont développé des applications permettant d'assurer le suivi des tendances et notamment des hashtags. Dans une logique de partenariats entre entreprises, Twitter peut sponsoriser des hashtags associés à des activités, événements ou produits développés par des entreprises privées ; et réciproquement, les responsables marketing peuvent *tracker* la popularité d'un hashtag afin d'orienter leurs activités. Si l'idée de la veille des messages émis sur la plateforme et d'une communication maîtrisée à travers la diffusion et l'adoption de hashtags précis pourrait s'avérer utile dans le domaine de la gestion des catastrophes naturelles, ces outils restent pour l'instant développés dans une logique commerciale et non de participation citoyenne volontaire.

2.2.3. Outils de visualisation des tweets accessibles à tout internaute

2.2.3.1. Les interfaces cartographiques de recherche et de visualisation des tweets géolocalisés

Le Géoweb héberge des applications cartographiques interactives permettant à tout internaute de visualiser, de rechercher et d'explorer la localisation des foyers d'émission de tweets géolocalisés en temps réel ainsi que leur variabilité sémantique. Diffusés auprès du grand public internaute, les usages cartographiques de ces plateformes s'inscrivent dans la continuité de la logique exprimée dans le chapitre précédent. En général, elles affichent un contenu web géolocalisé, soit sous forme de points, de clusters agrégeant des points ou sous forme de carte de chaleur. L'application *#onemilliontweetmap*[29] (cf. figure 2.12) propose ainsi une cartographie mondiale des tweets collectés via l'*API Streaming* sur une période de vingt-quatre heures, que l'internaute peut afficher sous forme de clusters ou de carte de chaleur. Celui-ci peut, en outre, filtrer les tweets affichés sur la carte en fonction de mots-clés ou de hashtags, sélectionner des tweets en fonction de pas de temps variés (tweets émis pendant les 5, 15, 30 minutes ou 4 dernières heures, etc.) et visualiser, sous forme de graphique en

[29] https://onemilliontweetmap.com/

barres horizontales, les pays qui concentrent les flux de tweets les plus conséquents, les hahstags les plus récurrents, les langues les plus fréquentes et les villes les plus actives sur le réseau. La carte étant zoomable, ces statistiques d'utilisation se mettent automatiquement à jour en fonction de l'emprise de l'espace géographique visualisé par l'internaute. La figure 2.12 affiche ainsi la carte des foyers de tweets agrégés en clusters sur le planisphère global, accompagnée d'un graphique indiquant les métropoles les plus actives sur le réseau, ainsi qu'un zoom sur les Etats-Unis, présentant les mêmes données, actualisées au niveau de l'emprise spatiale de la carte :

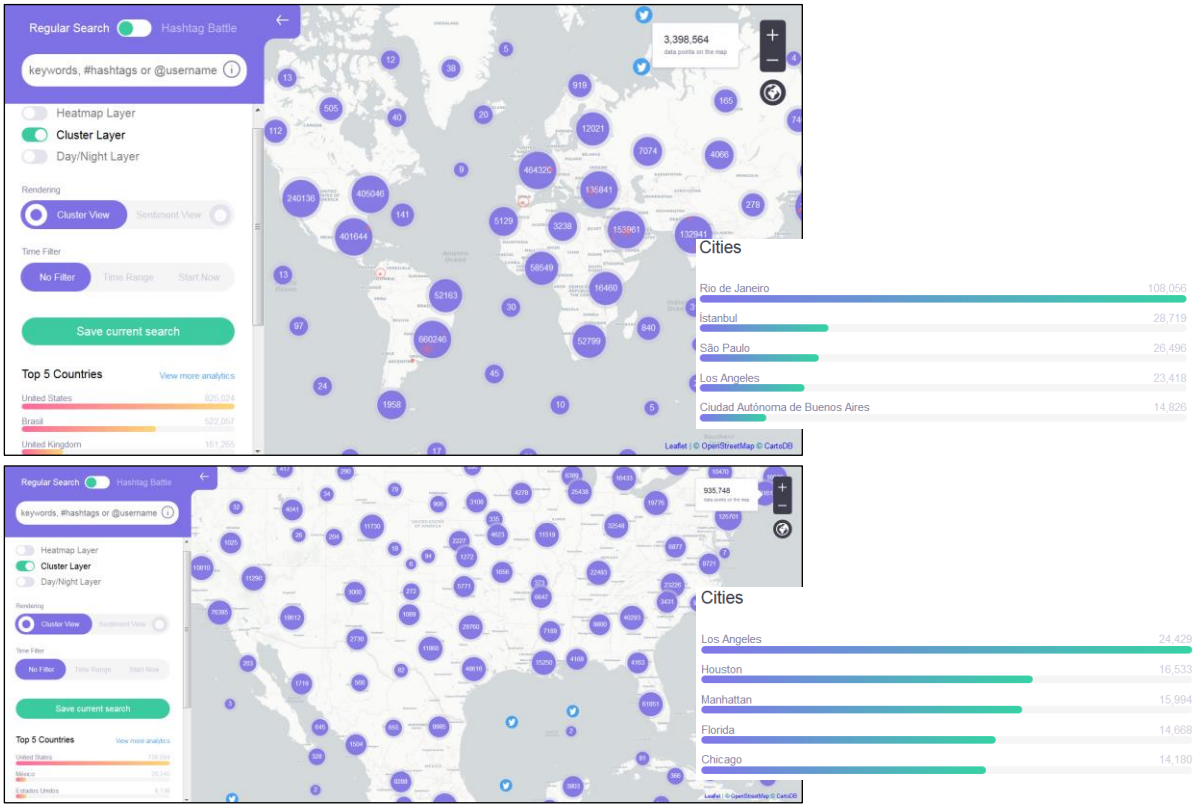


Figure 2.12 : Interface cartographique de visualisation de tweets géolocalisés agrégés en clusters et collectés en temps réel : OneMillionTweetMap

De la même manière, l'application *Omnisci Tweetmap*[30] (cf. figure 2.13) affiche un planisphère sur lequel est représenté un semis de points, chaque point figurant un tweet géolocalisé par ses coordonnées GPS ; les tweets sont colorés en fonction de leurs langues d'émission respectives, lesquelles peuvent être consultées sous la carte et sont hiérarchisées en fonction du nombre de tweets émis dans chaque langue. Sur cette interface cartographique, on trouve également un graphique représentant les variations temporelles des rythmes d'émission des tweets, agrégés sur une période de six heures, et visibles pour chaque journée de la durée de collecte des tweets affichés sur la carte (au 18 mars 2019 sont

[30] https://www.omnisci.com/demos/tweetmap/

ainsi affichés les tweets collectés depuis le 14 décembre 2018). L'application trie et restitue les hashtags les plus diffusés ; hashtags et statistiques d'utilisation se mettent également à jour en fonction du niveau de zoom choisi par l'internaute. La figure 2.13 représente ainsi les tweets géolocalisés sur le planisphère et colorés en fonction de la langue d'émission (les trois langues les plus inscrites sur le réseau restent l'anglais, le portugais et l'espagnol). Le volet situé à droite de la carte indique les hashtags les plus fréquents sur le réseau géolocalisé (à l'échelle mondiale, le premier hashtag identifié est ainsi *#twitterbestfandom*). La seconde carte représente un zoom actualisé sur la métropole de Houston au Texas : la langue la plus représentée sur le réseau reste donc l'anglais mais le hashtag le plus fréquent s'avère être *#houston* :

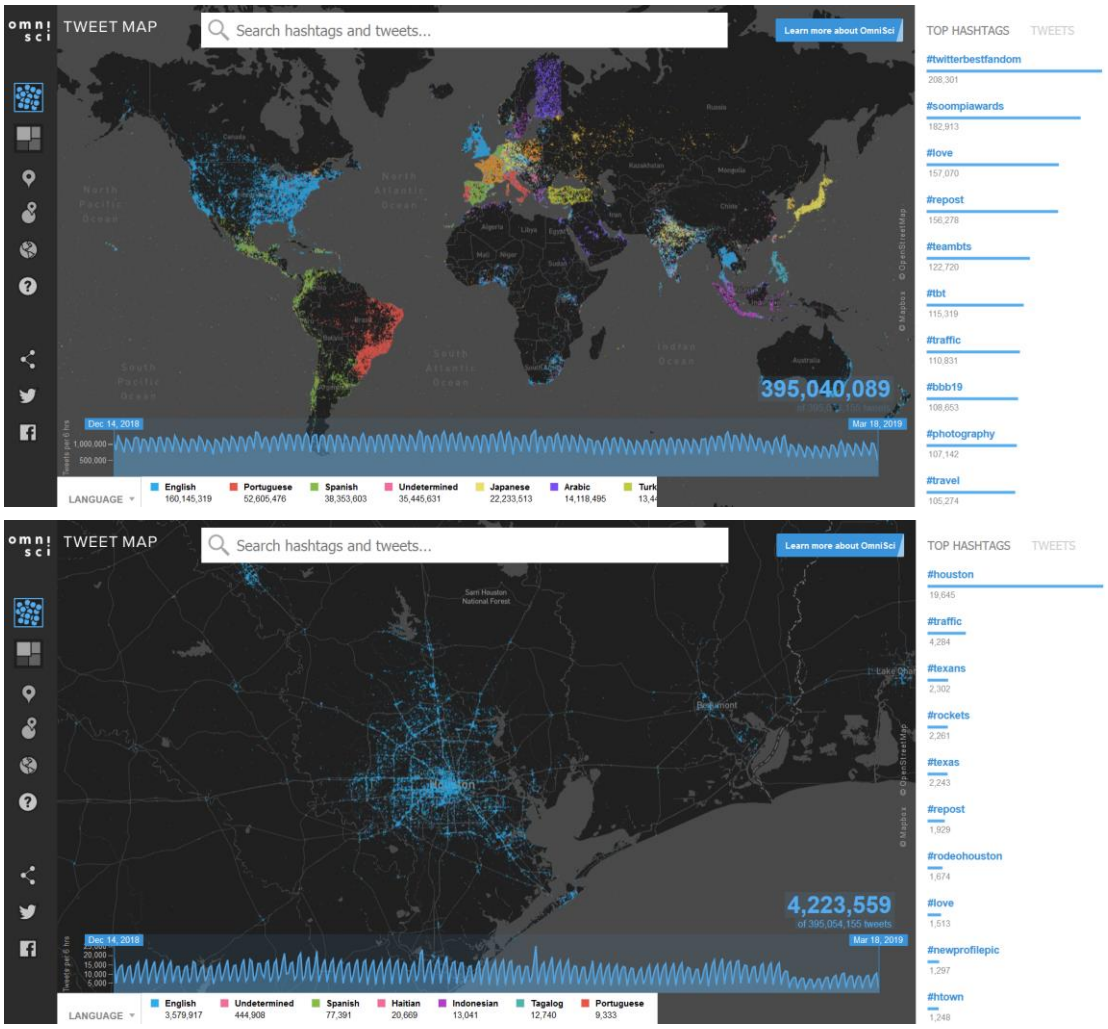


Figure 2.13 : Interface cartographique de visualisation de tweets géolocalisés sous forme de points en fonction de leur langue d'écriture, Omnisci TweetMap

Ces plateformes cartographiques en accès libre peuvent constituer un outil alternatif de veille et d'exploration des tendances ; en outre, puisqu'elles sont basées sur l'affichage de tweets géolocalisés uniquement, ces différentes tendances sont spatialisées : on peut ainsi

rapidement repérer les espaces actifs, les espaces peu visibles sur le réseau ainsi que les rythmes temporels de diffusion de l'information. Quelles utilités géographiques peut-on alors attribuer à ces plateformes ? La *Omnisci Tweetmap* de la figure précédente est présentée, sur le site de la plateforme, comme une base de données agrémentée d'une interface visuelle permettant de reconnecter les données et le *data analyst*[31]; le point essentiel mis en avant dans la présentation de la plateforme reste sa rapidité d'exécution ainsi que le volume de données affichées. Au-delà de l'aspect de stockage, d'interrogation et de visualisation des tweets géolocalisés, quel est le potentiel géographique d'une telle plateforme ?

En premier lieu, on peut identifier des tendances spatiales certaines comme la correspondance entre les foyers de peuplement mondiaux et les foyers d'activité virtuelle globale (à l'exception des régions moins connectées et des Etats interdisant l'accès à Twitter). Lorsqu'on zoome à l'échelle d'une métropole, on pourra nettement distinguer le réseau de la ville ainsi que les lieux qui polarisent l'activité virtuelle. Par ailleurs, si l'on assure une veille régulière sur la plateforme, on pourra sans doute observer des variations saisonnières des flux de tweets (en fonction de l'attractivité touristique des lieux par exemple). Sur l'aspect cartographique, *Omnisci Tweetmap* permet de valoriser la trace numérique géolocalisée en donnée géographique : l'interface peut en effet générer une carte "*choroplèthe*" indiquant le nombre de tweets géolocalisés inventoriés par pays pendant la période de collecte considérée ; c'est ce que montre la figure 2.14 pour un ensemble de tweets géolocalisés collectés entre le 14 juin 2019 et le 17 septembre 2019.

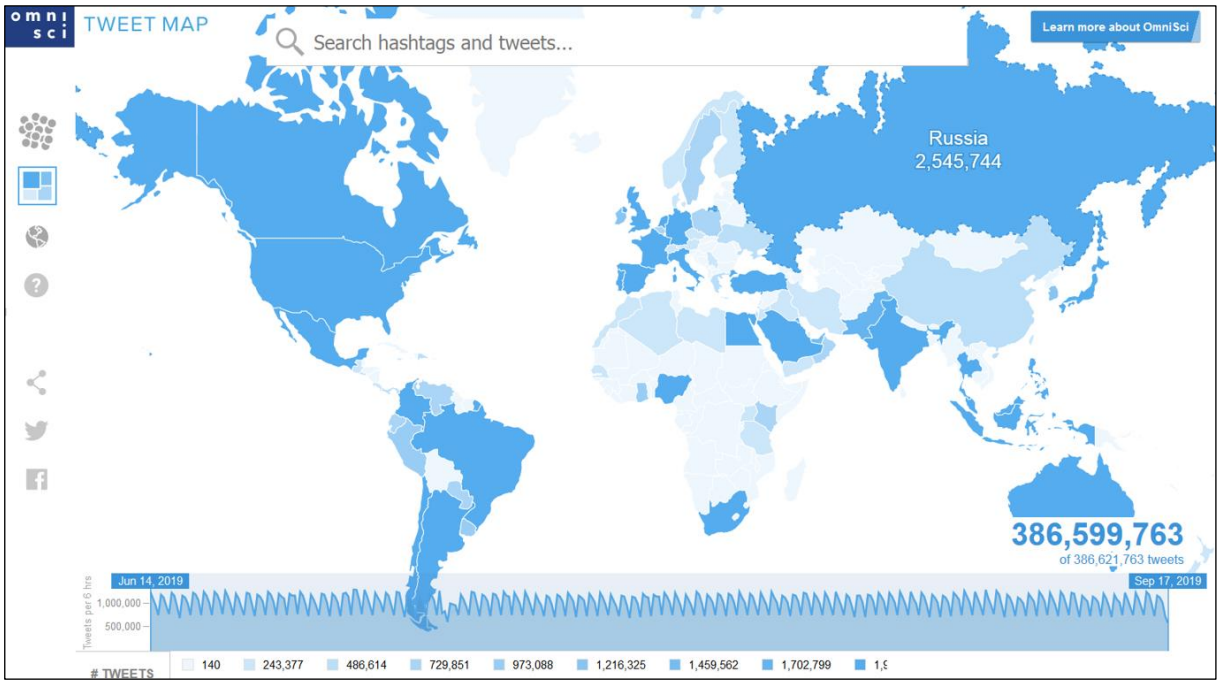


Figure 2.14 : Carte choroplèthe générée sur l'interface Omnisci Tweetmap, indiquant le nombre de tweets géolocalisés émis par pays entre le 14/06/2019 et le 17/09/2019

[31] Dans l'approche traditionnelle, le *data analyst*, contrairement au géographe, ne voit pas directement les données sur lesquelles il travaille, qui sont d'abord traitées par les algorithmes.

Pour autant, cette cartographie présente les défauts fréquents des cartes du Géoweb non réalisées par des géographes :

- en dépit de l'appellation de carte choroplèthe, annoncée par la plateforme[32], la représentation des données est fausse. Pour comparer l'activité virtuelle inscrite dans les différents Etats, il convenait par exemple de rapporter le nombre de tweets géolocalisés émis au nombre d'habitants. Telle que la carte apparaît à l'écran, nous pouvons constater que des pays apparaissant dans la même couleur (et dont on peut alors penser qu'ils présentent un profil d'activité virtuelle analogue) se révèlent en fait très différents : si l'on s'intéresse au rapport entre le nombre de tweets et le nombre d'habitants pour la Russie et les Etats-Unis, on calcule respectivement des ratios de 0,17 et de 0,33 (il y a donc presque deux fois plus de tweets par habitant aux Etats-Unis qu'en Russie, tendance qui ne transparaît pas sur la carte de la figure 2.14) ;
- on ne sait pas comment la discrétisation a été effectuée, de même que la carte n'offre pas de légende explicite (malgré les valeurs présentes dans la partie inférieure de la fenêtre mais dont les couleurs semblent plus nombreuses que celles de la carte des pays). Si l'utilisateur souhaite connaître le nombre de tweets par pays, il doit glisser la souris sur le territoire concerné (l'information apparaît alors sous la forme d'une étiquette).

Cette plateforme propose donc un exemple de transformation simple de la trace en donnée géographique mais dont la représentation cartographique est fausse par rapport aux normes et attentes des géographes.

2.2.3.2. Outils cartographiques thématiques

Certaines interfaces cartographiques affichant des tweets sont focalisées sur des thématiques plus précises ; l'application *#onemilliontweetmap*, présentée précédemment, propose par exemple de visualiser la localisation des tweets selon qu'ils sont *positifs* ou *négatifs* : autrement dit, l'application dispose d'un algorithme d'apprentissage d'analyse de sentiments qui classifie les tweets en fonction du vocabulaire positif (*amazing*, *handsome*, *celebrated*, *healthier*) ou péjoratif (*accident*, *rainy day*, *gloomy*, *tropical storm*) détecté dans la chaîne de caractères[33]. La figure 2.14 ci-dessous affiche une vue de la carte focalisée sur le centre de Houston ; les tweets restent agrégés en clusters dont la représentation en diagramme circulaire indique le poids de chaque type de tweets dans le cluster (les tweets non classés par l'algorithme sont étiquetés comme *neutres*).

[32] Le bouton sur lequel on clique pour faire apparaître cette carte s'intitule en effet *Choropleth*.
[33] L'algorithme présente néanmoins ses limites : des tweets diffusant des offres d'emploi sont classés dans les scores positifs alors que le tweets suivant "*I'm at Chick-fil-A* [chaîne de restauration rapide] *in Clear Lake Shores, TX*" est classé comme négatif alors qu'il pourrait simplement être neutre.

Dans cette même perspective - la localisation de tweets soumis à un traitement et non la simple localisation des tweets bruts collectés par l'API – on trouve la cartographie *Locals & Tourists* développée par Eric Fischer : elle se présente sous la forme d'un semis de points bleus et rouges représentant respectivement l'activité virtuelle supposée liée aux individus résidant dans un territoire (les utilisateurs ayant tweeté dans une ville pendant au moins un mois) et cette même activité supposée liée aux touristes (les utilisateurs ayant tweeté pendant moins d'un mois dans la ville en question ou enregistrés comme résidents d'une autre ville). Malgré les biais méthodologiques qu'on peut supposer (tout utilisateur résidant dans une ville n'est pas nécessairement un contributeur régulier au réseau), des tendances se dessinent : si l'on zoome sur Genève (figure 2.15), on distingue nettement un semis de points rouges sur l'aéroport ainsi que des noyaux regroupés dans le centre de la ville et au niveau du Palais des Nations Unies. Au contraire, les points bleus associés à l'activité des locaux se répartissent de manière homogène sur l'ensemble du territoire, dessinant ainsi le réseau de la ville.

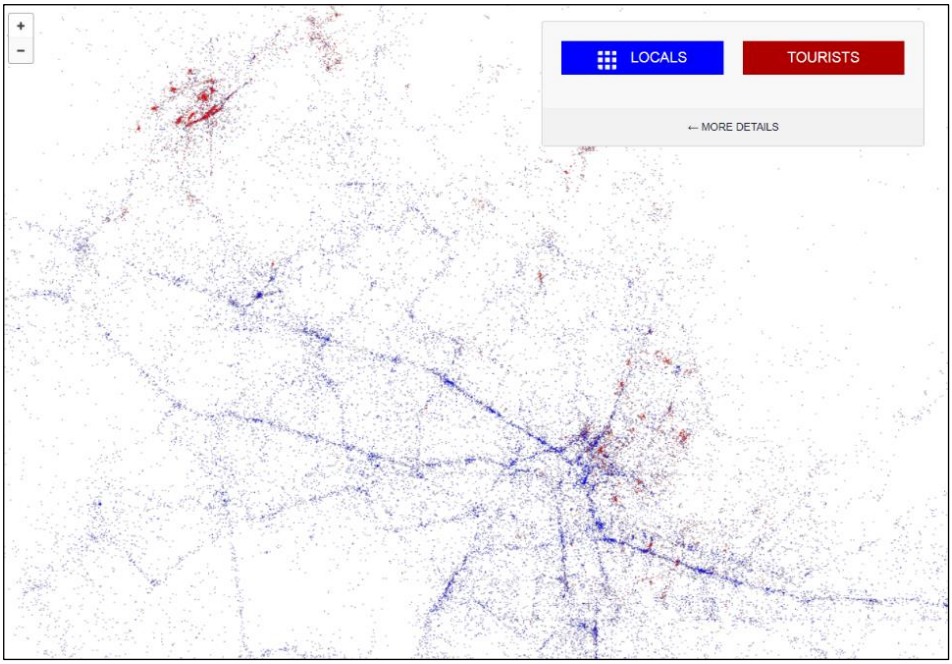


Figure 2.15 : Tweets géolocalisés sur la ville de Genève, classés en fonction du profil – local ou touriste - de l'utilisateur (Eric Fischer, Locals & Tourists)

De telles représentations cartographiques associées aux contenus issus des réseaux et médias sociaux numériques sont souvent appréhendées comme de nouvelles portes d'entrée dans la compréhension des liens tissés entre les usagers des territoires et les lieux à partir desquels les contenus web sont émis (Mericskay *et al.*, 2018). Mais ces représentations cartographiques soulèvent deux écueils :

- elles sont souvent associées aux travaux de *data artists* (Mericskay *et al.*, 2018) : or, si le cartographe de la Renaissance entremêlait performance artistique à rigueur scientifique, il n'en est pas toujours le cas dans les nouvelles cartographies du Géoweb. La carte de la figure

2.15 en témoigne : aucun fond référentiel ni aucune échelle ne figurent ; en conséquence, l'internaute qui navigue sur cette carte ne dispose que du réseau dessiné par les tweets pour se repérer dans l'espace ;

- le second problème, déjà mis en évidence, reste le défaut de représentativité des médias sociaux géolocalisés : diagnostiquer l'identité des lieux par les contenus web géolocalisés, c'est s'appuyer sur les pratiques et représentations d'une minorité d'individus dont il est difficile de caractériser les profils et dont on ne sait pas s'ils sont représentatifs de l'ensemble des populations qui interagissent avec chaque lieu en question.

2.3. Positionnement du tweet géolocalisé comme source potentielle d'information géographique

Pour étayer les propos exposés dans les paragraphes précédents, cette dernière section vise à positionner le matériau tweet géolocalisé comme source d'information géographique par rapport aux réflexions amorcées sur les différentes formes de données 2.0 dans le premier chapitre. Elle s'articule autour des caractéristiques propres aux tweets et des risques encourus pour le chercheur qui souhaite résoudre une question géographique à l'aide d'un matériau qui, s'il dispose d'une référence spatiale, constitue une information de base biaisée et imparfaite.

2.3.1. Les caractéristiques propres aux contenus géolocalisés des médias sociaux

2.3.1.1. Des dynamiques d'émission étroitement liées aux comportements des utilisateurs

Contrairement aux données des instituts qui sont produites périodiquement, comme les données de recensement, ou en réponse au besoin d'une étude particulière, les lieux et dynamiques spatio-temporelles des émissions de tweets géolocalisés témoignent de logiques propres aux comportements des utilisateurs du réseau. Nous considérons ainsi la participation régulière d'un individu à la création de contenus géolocalisés comme un phénomène social reposant sur deux critères influencés par une série de facteurs :

- l'*adhésion* aux technologies 2.0 et à l'esprit du Web social - l'individu s'inscrit et participe à la création de contenus sur la plateforme : dispose-t-il de moyens techniques suffisants et satisfaisants pour lui assurer une connexion à la plateforme en n'importe quel lieu, à n'importe quel moment de la journée ? Trouve-t-il un intérêt quelconque à faire partie de cette communauté virtuelle particulière ? Cet intérêt peut varier en fonction des sources

de médias vers lesquelles l'individu préfère s'orienter pour s'informer, des moyens de communications qu'il privilégie, de son appartenance à d'autres communautés virtuelles comme Facebook, etc.

- la *motivation* qui incite l'individu membre à tweeter : quels sont les *facteurs déclencheurs* de l'activité de création de contenu sur le réseau ? Est-ce un besoin impérieux de faire parler de soi ? Un besoin de partager des opinions ou des connaissances ? Un simple usage récréatif (conversation entre un groupe d'amis) ponctué (ou pas) de commentaires sur des événements inhabituels ?

En conséquence, le flux de contenus créés au quotidien peut se subdiviser en information de routine et en information de réponse à un fait réel. En effet, comme le signalent les tendances quotidiennes, cette *motivation* à tweeter est fréquemment influencée par des faits réels qui marquent l'actualité, qu'elle soit politique, environnementale, sociale, culturelle ou sportive. Nous proposons donc d'introduire deux facteurs psychologiques supplémentaires qui caractérisent les rythmes d'émission liés aux motivations à tweeter :

- la *réactivité* : nous considérons la réactivité comme un facteur d'échelle individuelle qui caractérise l'implication immédiate de l'individu à tweeter en réponse à la constatation d'un fait du monde réel. Ce paramètre reste subjectif et peut être biaisé par les caractéristiques sociales, culturelles et émotives de l'individu mais également par son expérience (le fait constaté vaut-il la peine d'être tweeté et donc enregistré sur le réseau ou non ?) : un individu familier d'un territoire ne prendra sans doute pas la peine de tweeter à propos d'une crue annuelle alors qu'un nouvel arrivant, assistant pour la première fois au phénomène, prendra le temps de le capturer et de diffuser l'information sur le réseau.

- la *sensibilité* aux événements : nous appréhendons la sensibilité comme un phénomène palpable à l'échelle collective, résultant de la réactivité individuelle : un fait constaté dans le monde réel a-t-il suscité une réactivité massive ou restreinte à un petit cercle d'individus membres ? Un attentat est localisé en un lieu précis et à une échelle locale mais va susciter une réactivité massive sur le réseau et résonnera à l'échelle du globe ; *a contrario*, un train en panne, bien qu'également un fait d'échelle locale, ne bénéficiera pas d'un tel retentissement car il ne suscite que la réactivité d'un groupe restreint aux usagers du réseau ferroviaire.

La création de contenus géolocalisés obéit ainsi à une logique routinière ou événementielle (celle qui nous intéresse dans ce travail). (Lucchini *et al.*, 2016) proposaient alors de définir l'événement comme un fait se distinguant du quotidien de par son caractère inattendu ou exceptionnel, tout en mettant en garde sur l'émotivité des utilisateurs qui auraient tendance à conférer ce caractère événementiel à tout fait du monde réel. Les événements peuvent être officiels (rencontre politique, sportive, etc.) ou aléatoires (attentat, accident). En ce qui concerne les risques et catastrophes naturels, nous considérons l'activité capturée sur Twitter comme un événement virtuel qui répond à un phénomène naturel déclenchant, par ses manifestations multiples, un événement dans le monde réel (routes

inondées, habitations détruites, etc.). Cet événement du monde réel est officiel : on peut prévoir l'arrivée d'une perturbation d'origine météorologique et anticiper ses effets grâce aux politiques de prévention des catastrophes naturelles. Pour autant, l'événement englobe une part d'aléatoire : quels sont les différents lieux qui seront frappés et à quelle intensité ? Les individus vont-ils suivre les comportements appropriés ou se mettre en danger ?

2.3.1.2. La représentativité des territoires et de leurs population en question

Qui tweete ?

Tout tweet est avant tout un contenu numérique porteur de biais sociaux : adhésion à la plateforme et motivation à tweeter constituent deux facteurs qui peuvent être étroitement liés à l'âge, à l'origine et aux conditions sociales de l'individu, à son accessibilité au réseau (autant financière que technique), à son niveau d'éducation, etc., soit autant de variables dont les effets se répercutent sur l'activité globale d'émission de tweets (qu'ils soient géolocalisés ou non) et sur l'activité en réponse à un événement réel.

Comme indiqué dans la section 2.1.2.2 (*Les enjeux de la géolocalisation des tweets*), cette activité, et en particulier lorsqu'il s'agit de contenus géolocalisés, reste difficile à associer à des profils particuliers ; en outre, ces profils d'individus qui utilisent la géolocalisation ne semblent pas représentatifs de l'ensemble de la communauté Twitter. A ces observations, on pourra ajouter l'existence d'une variabilité de ces profils en fonction des territoires considérés. (Li *et al.*, 2013) étudiaient le cas des contenus géolocalisés publiés via Twitter et Flickr (plateforme de partage de photographies) sur l'Etat de Californie. Ils parvenaient à mettre en évidence une corrélation forte entre haut niveau d'études, catégories professionnelles supérieures et activité géolocalisée sur Twitter[34]. (Cebeillac *et al.*, 2016) cherchaient à identifier les populations créatrices de contenus géolocalisés en un lieu précis : un centre commercial de luxe en Inde, fréquenté par la classe moyenne supérieure. Dans ce contexte, il apparaissait que la création de contenus géolocalisés s'associait davantage aux revenus moyens de cette classe, soit à des individus fréquentant le lieu dans une logique de promotion de soi par un environnement luxueux mais n'ayant pas nécessairement les moyens financiers de consommer en ce lieu. En revanche, les véritables clients du centre commercial se démarquaient de cette logique ostentatoire et s'avéraient peu actifs sur les réseaux sociaux.

Pour autant, rien ne garantit au chercheur que les résultats observés sur la population californienne, indienne ou britannique ne s'appliquent à d'autres régions du monde : contrairement aux données issues d'une enquête ou d'un sondage, un jeu de tweets n'est pas un échantillon aléatoire représentatif d'une population mère (Goodchild, 2013). Les résultats d'une étude fondée sur les tweets ne peuvent être généralisés dans un monde géographique

[34] La plus forte activité géolocalisée ayant alors été détectée sur la Silicon Valley, l'étude ne prend néanmoins pas en compte ce biais : ce territoire est en effet le lieu de concentration des multinationales développant les plateformes du web 2.0. L'activité sur le réseau peut donc être liée à l'activité professionnelle.

complexe, étant donné que le tweet géolocalisé ne représente pas une population mère. En conséquence, nous pouvons supposer que les tweets géolocalisés traduisent les préoccupations des utilisateurs mais nous ne savons pas si ces préoccupations sont partagées par l'ensemble des personnes qui vivent et parcourent un même espace.

Sur quels territoires ?

La seconde conséquence des modalités pratiques du facteur d'adhésion à la plateforme Twitter est la question de la représentation spatiale équitable des territoires. La distribution des contenus géolocalisés est-elle représentative du nombre d'individus recensés sur le territoire ? Comme indiqué dans l'introduction, le numérique tend à s'intégrer dans des espaces de prédilection pour en délaisser d'autres. En 2007, Goodchild introduisait d'ores-et-déjà le concept de *digital divide* (la fracture numérique) dans la production de la VGI, le phénomène étant essentiellement un fait urbain. En 2016, Capineri confirmait que la VGI et l'ensemble des contenus générés sur les réseaux sociaux numériques se concentraient dans les villes, celles-ci cumulant deux avantages : une meilleure connectivité aux réseaux et une concentration d'individus aux profils variés (résidents, touristes, étudiants, etc.). En outre, l'auteur s'appuyait sur l'étude de (Hecht & Stephens, 2014), sur la variabilité spatiale de l'activité tweeting dans les différents comtés des Etats-Unis. Celle-ci mettait alors en évidence une sur-représentation des utilisateurs en milieu urbain : pour un utilisateur d'un comté rural, on trouve 3,5 utilisateurs dans un comté urbain. Par ailleurs, signalons que les milieux urbains constituent l'objet de recherche privilégié dans les études à grande échelle géographique (Andrienko *et al*., 2013 ; Kounadi *et al*., 2015 ; Lucchini *et al*., 2016 ; Cebeillac *et al*., 2017).

Qu'en est-il des tweets géolocalisés du Texas ? A petite échelle géographique, une carte de la distribution de ces contenus reflète indubitablement la carte de la répartition des populations et met effectivement en exergue les milieux urbains et surtout métropolitains comme foyers de concentration des émissions de tweets géolocalisés. La figure 2.16 illustre ce comportement : si l'on calcule les ratios respectifs du nombre d'habitants recensés par la superficie des comtés, puis du nombre de tweets géolocalisés émis pendant un mois par la superficie des comtés, on constate la correspondance entre comtés métropolitains densément peuplés et comtés concentrant les émissions de twets géolocalisés (comtés des métropoles de Dallas, Houston, San Antonio et Austin, soit les quatre plus grandes agglomérations de l'Etat).

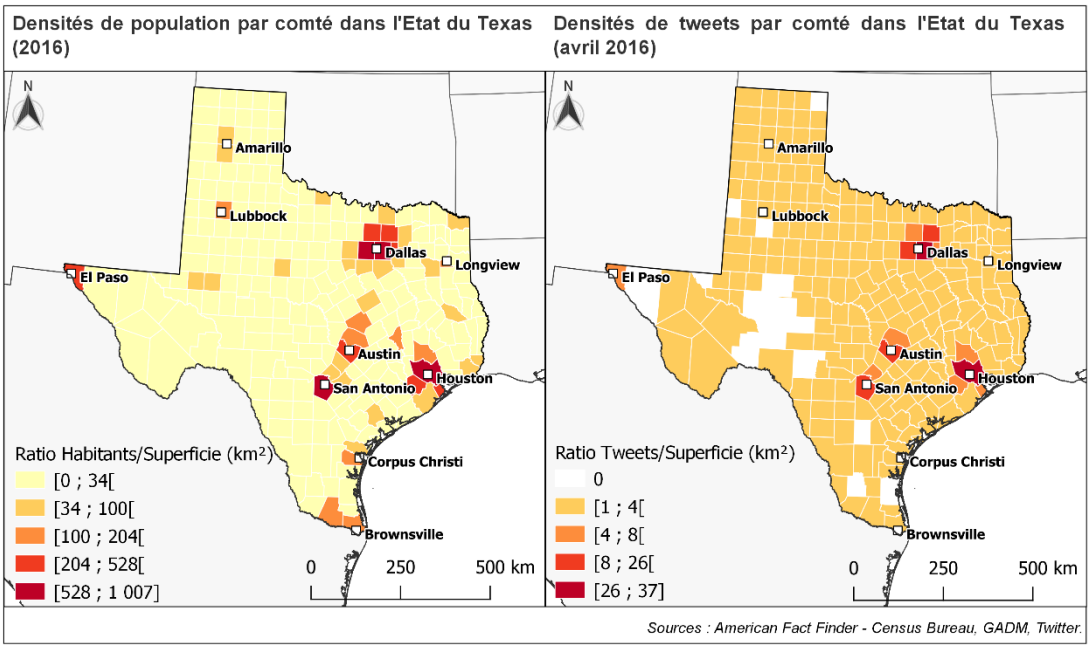


Figure 2.16 : Comparaison des densités de tweets géolocalisés et d'habitants recensés dans les comtés du Texas (C.Cavalière)

A plus petite échelle géographique, et précisément en milieu métropolitain, la répartition spatiale des tweets se démarque de cette logique de densités de population. En fait, si l'on considère que le tweet géolocalisé est capturé en situation de mobilité dans le cadre de l'exercice des activités quotidiennes, alors que le recensement de la population par unité spatiale est acquis au domicile, la répartition de ces contenus va mettre en évidence, d'une part, l'existence des espaces urbains attractifs qui concentrent les émissions et d'autre part, les espaces invisibles du réseau dans lesquels aucun contenu n'est créé. Ce comportement s'avère palpable sur la figure 2.17 qui indique le ratio entre nombre d'habitants recensés par la superficie des *census tracts* de l'aire métropolitaine de Houston d'une part, et d'autre part, le ratio entre le nombre de tweets géolocalisés émis en avril 2016 par la superficie de ces mêmes unités spatiales.

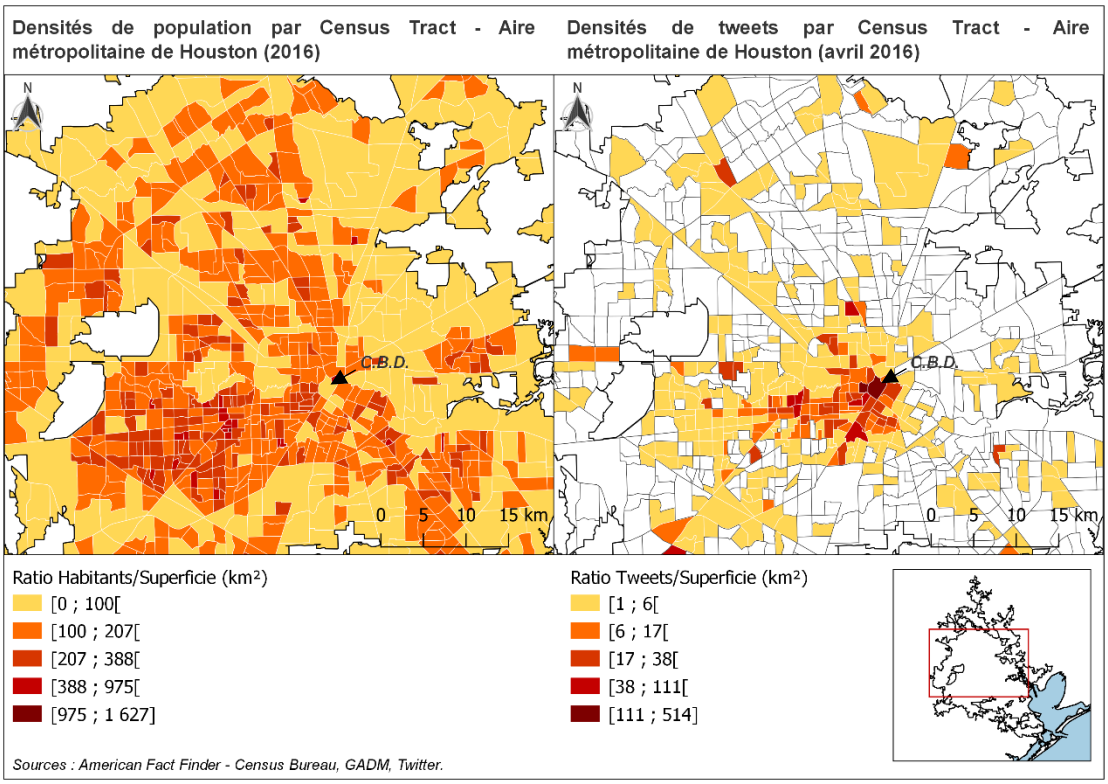


Figure 2.17 : Comparaison des densités de tweets géolocalisés et d'habitants recensés dans les Census Tracts de l'aire métropolitaine de Houston (C.Cavalière)

Les quartiers résidentiels visibles sur la carte de gauche (et en particulier dans le nord, le sud et l'extrême ouest) apparaissent peu actifs sur le réseau virtuel voire vides de tweets géolocalisés. A l'inverse, les unités proches du centre ainsi que le *CBD* lui-même présentent les plus fortes concentrations de tweets géolocalisés.

A l'échelle d'une métropole comme Houston et dans la problématique de gestion des risques naturels, cette tension entre espaces pleins/vides soumet un certain nombre d'enjeux (dont éthiques) : quelles sont les caractéristiques des populations qui tweetent ? Certains individus/territoires sont-ils exclus du numérique ? Les espaces enregistrant les concentrations d'émissions de tweets pendant une crise sont-ils représentatifs des espaces affectés et de l'intensité des phénomènes naturels ?

2.3.1.3. Biais quantitatifs à l'échelle individuelle

La représentation cartographique des tweets géolocalisés met en évidence une répartition spatiale inégale des contenus, qui introduit une hiérarchisation des territoires : les lieux urbains constituant les foyers d'émission de contenus géolocalisés, ils forment alors les objets de recherche polarisant les attentions au détriment de territoires aux populations moins actives sur le réseau. Pour autant, quand bien même on considère les individus

parcourant un territoire urbain comme producteurs de grandes quantités de traces, une nouvelle série de facteurs liés à l'utilisation individuelle des réseaux risquent de limiter la quantité de contenus produits.

En premier lieu, l'individu membre de la communauté ne tweete pas de manière équitable dans tous les lieux qu'il fréquente (Quesnot, 2016), ni à toute heure : les pics d'activité enregistrés sur le réseau témoignent de périodicités (Andrienko *et al*., 2013). On sait ainsi que les individus sont plus actifs à des moments précis de la journée : le matin, pendant la pause méridienne et le soir. De la même manière, à l'échelle de la semaine, les flux les plus conséquents sont drainés le weekend. En ce qui concerne l'activité de création de contenu géolocalisé d'un utilisateur en fonction du lieu fréquenté, la difficulté à appréhender les espaces de l'activité tweeting et les espaces invisibles provient de nouveau du caractère a-contextuel des traces : c'est ce que propose la figure 2.18. Elle représente l'activité d'un utilisateur en situation de mobilité. Dans cet exemple, l'utilisateur quitte son domicile sans tweeter, monte dans un tramway pour se rendre en ville, tramway dans lequel il émet un premier tweet. Il retrouve des amis dans une brasserie ; tous déjeunent ensemble mais notre utilisateur ne tweete pas. Ils se rendent ensuite, à pied, au cinéma. Pendant ce trajet, il n'émet pas mais alors qu'il visionne le film, il ne peut pas s'empêcher de tweeter pour commenter le film projeté. Ce type de comportement pourrait alors constituer une explication des vides et pleins constatés dans les milieux urbains qui concentrent des lieux d'intérêt propices à motiver l'activité virtuelle. Mais il démontre également le besoin impérieux d'une prise de recul vis-à-vis de ces contenus géolocalisés : tels qu'ils sont émis, ils ne peuvent pas capturer les pratiques spatiales des individus dans leur globalité (Quesnot, 2016).

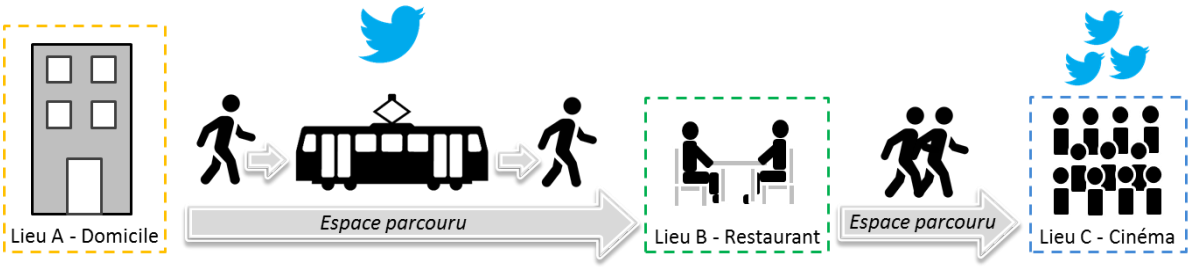


Figure 2.18 : Participation à l'activité virtuelle en situation de mobilité quotidienne

Pour le cas particulier des émissions créées en réponse à un phénomène d'origine naturelle, nous pouvons identifier deux facteurs d'échelle individuelle susceptibles d'influencer la création de contenus géolocalisés :

- une brève revue des travaux bibliographiques démontre que la distribution spatiale et temporelle des tweets est irrégulière et dépend de la distribution spatio-temporelle de la population (Li *et al*., 2013 ; Hawelka *et al*., 2013 ; Lucchini *et al*., 2016). Les effets vides/pleins constatés à différentes échelles géographiques ainsi que les rythmes d'émission de tweets se

répercutent alors directement sur la *visibilité* des événements du monde réel sur le réseau : des événements qui surviennent dans des espaces peu peuplés peuvent être sous-représentés ou quasiment absents du réseau. Par ailleurs, à l'échelle locale, les densités de population varient en fonction du temps : si un événement survient en pleine journée dans un quartier résidentiel, pendant une période où cet espace est peu occupé, il est fort probable qu'il ne soit pas enregistré sur le réseau. *A contrario*, les lieux concentrant les populations pendant la journée seront sur-représentés.

- l'*expérience vécue* d'un territoire ou d'un événement est étroitement liée à la *réactivité* de l'utilisateur local qui vit l'événement : un individu non familier d'un environnement perturbé peut davantage être enclin à réagir sur le réseau qu'un individu ayant déjà vécu un événement aux caractéristiques identiques. En revanche, si l'événement en question s'avère plus intense que ceux connus dans l'histoire personnelle de cet individu, alors celui-ci prendra sans doute l'initiative de l'enregistrer sur les réseaux sociaux.

- pour finir, la *sensibilité* des individus varie en fonction d'événements aux caractéristiques physiques différentes : ceux-ci n'ont ainsi pas le même impact de diffusion sur le réseau. Certains événements majeurs bénéficient d'un très fort retentissement médiatique et ont la capacité de générer des flux considérables de tweets : en général, l'enveloppe spatiale des tweets qui mentionnent ce type d'événement dépasse largement la zone physiquement et réellement affectée. Au contraire, des événements d'échelle locale, moins médiatisés, n'aboutissent pas à une telle mobilisation des utilisateurs sur le réseau : l'enveloppe spatiale des tweets ne dépasse pas ou dans une distance restreinte celle de l'événement physique réel (figure 2.19).

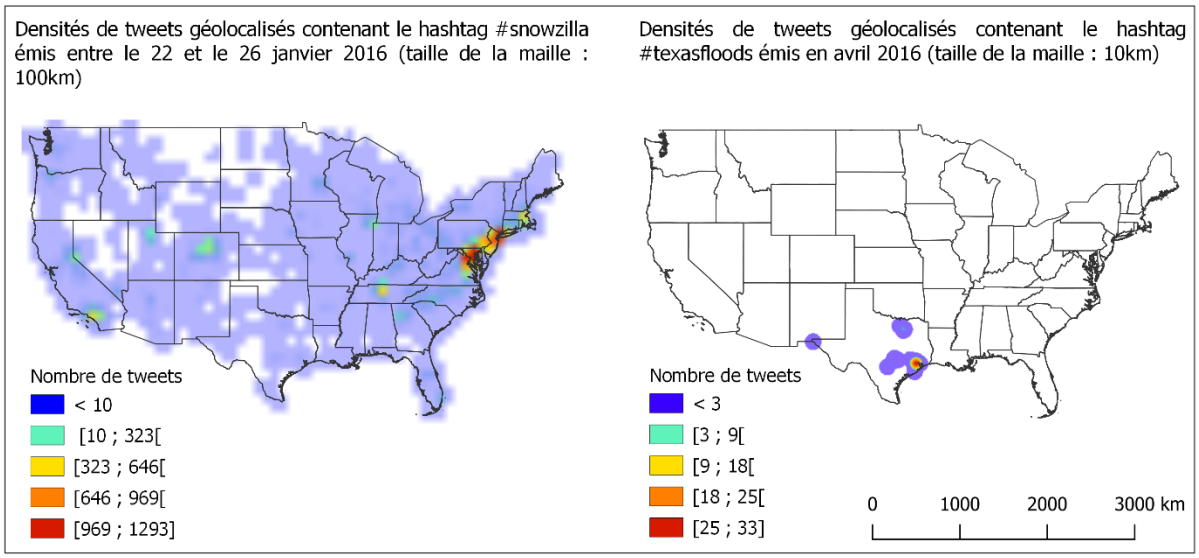


Figure 2.19 : Variabilité de la sensibilité en fonction de deux types de phénomènes d'origine naturelle (C.Cavalière)

Comme le soulignait la figure précédente, l'événement virtuel consécutif au phénomène météorologique extrême (tempête de blizzard Jonas ayant frappé la côte est des Etats-Unis en janvier 2016, identifiée par le hashtag *#snowzilla*) crée une activité tweeting géolocalisée qui se diffuse à l'ensemble du pays. A l'inverse, l'utilisation du hashtag *#texasfloods* (en référence aux phénomènes saisonniers associant pluies intenses et inondations) ne dépasse pas l'Etat physiquement affecté du Texas.

2.3.2. Positionnement des tweets géolocalisés dans les données géographiques

2.3.2.1. Une donnée géographique valorisable ?

En tant que contenu web géolocalisé issu d'une plateforme du Web social, le tweet géolocalisé hérite des propriétés inhérentes aux données géographiques traditionnelles. Tout tweet géolocalisé correspond ainsi à un contenu tridimentionnel :

- il dispose d'une composante géométrique puisque localisé par ses coordonnées dans l'espace terrestre et de forme ponctuelle (ou polygonale si capturé par une *bounding box*) ;
- d'une composante attributaire regroupant l'ensemble des variables décrivant chaque tweet : son numéro identifiant, le numéro identifiant de l'utilisateur, le texte, la langue du tweet, le nombre de retweets, etc. ;
- d'une composante temporelle de résolution précise puisque capturée sous format *année-mois-jour heure:minutes:secondes*.

A ce titre, il peut être considéré comme un contenu spatio-temporel cartographiable de la même manière que toute donnée produite par les instituts et acteurs officiels des données géographiques. En outre, comme l'ont montré les cartographies interactives citées précédemment, le tweet peut être valorisé d'un matériau brut vers des données transformées selon des approches quantitatives (agrégation en fonction d'entités spatiales, clustérisation spatiale) permettant d'identifier, à diverses échelles spatiales, les lieux ou régions de participation à la création de tweets géolocalisés, mais également selon des approches qualitatives (notamment au travers de l'analyse des sentiments) : celles-ci peuvent alors favoriser l'identification des activités pratiquées en des lieux précis ou révéler quels objets du territoire suscitent la réactivité des utilisateurs[35].

[35] A l'image de ce tweet posté le 18/09/2019 par un touriste en voyage à Chamonix, renvoyant à une photographie diffusée sur le réseau Instagram : l'objet du territoire sélectionné par l'utilisateur est la Mer de Glace et le choix des hashtags est révélateur des sentiments suggérés à l'utilisateur par l'apparence du glacier : "*#merdeglace #glacier #alps #france #chamonixmontblanc #beautufulbutscary #globalwarming*"

2.3.2.2. Les propriétés des tweets géolocalisés

Des Soft Data.

Pour autant, le tweet géolocalisé acquiert les caractéristiques propres de certains types de contenus du Web social, se démarquant ainsi des données géographiques produites par les instituts. Le tableau 2.2 propose une comparaison des caractéristiques des jeux de données officiels et d'un jeu de contenu web géolocalisé de type tweet.

Tableau 2.2 : Comparaison des caractéristiques des données traditionnelles et des tweets géolocalisés

	Jeux de données produits par les organismes officiels	**Contenus géolocalisés issus de Twitter**
Structuration / Formalisme	Géométrie homogène (données organisées en couches de points, lignes ou polygones) ; Attributs structurés et cohérence sémantique (tout champ est renseigné par des modalités définies en fonction de la variable : nom de commune, type d'occupation du sol, etc.)	Géométrie hétérogène : pour un même objet, le tweet, on peut distinguer deux types de géométrie : le point (tweet géolocalisé par coordonnées GPS) ou le polygone (tweet localisé par une *bounding box*). Données associées aux géométries structurées en attributs (identifiant, horodatage, texte, etc.) mais aux modalités non définies et par conséquent hétérogènes : la variable « texte » d'un tweet n'est limitée que par sa taille, soit 280 caractères maximum.
Modalités de capture	Travail à mener en amont : collecte sur le terrain (enquêtes, recensement), traitement de données (télédétection à partir d'images aériennes ou satellites, etc.)	Production réalisée sur le terrain, immédiate et d'expérience vécue par l'utilisateur/producteur équipé d'un smartphone.
Modalités de contrôle	Normes à respecter dans la formalisation des données afin de garantir la qualité finale des jeux de données pour l'utilisateur et renseignement de métadonnées pour la description et les conditions d'utilisation des données produites et téléchargées.	Les métadonnées d'un tweet correspondent à des données présentes sous forme de champs (numéros identifiant, texte, horodatage, gps, etc.). Les contraintes qui encadrent la production des tweets sont techniques (par exemple, la limite des 280 caractères).
Diffusion et mise à jour	Mise à jour cyclique – Production différée	Production en temps réel, disponibilité immédiate des contenus, mise à jour fréquente puisque l'activité tweeting est répétitive.
Coût	Coût pour le producteur et le chercheur pour certaines gammes de produits	Logique d'entreprise : collecte en partie gratuite mais développement de partenariats entre entreprises privées et ventes des données ; dépendance aux choix des entreprises propriétaires des traces numériques.

Comme le souligne le tableau 2.2, les différences fondamentales entre tweets géolocalisés et données institutionnelles se déclinent en fonction de deux volets principaux :

- les modalités d'*acquisition* des contenus : capturé sans la moindre norme[36], le contenu géolocalisé correspond ainsi à une acquisition spontanée (l'utilisateur est témoin/participe à un événement du monde réel et l'enregistre immédiatement sur le réseau : accident, rencontre sportive, manifestation, etc.) et régulièrement mise à jour : un événement

[36] La seule forme de "régulation" de production qu'on peut identifier est la conséquence des constats évoqués dans les paragraphes précédents sur les territoires du numériques : les individus doivent disposer des moyens techniques suffisants pour être acteurs des réseaux sociaux et trouver un intérêt à devenir acteur de ces réseaux.

inhabituel comme l'incendie de la charpente de la cathédrale de Notre-Dame de Paris rassemble une foule de témoins dans le réel qui alimentent les réseaux sociaux pendant toute la durée de l'événement ; d'un autre côté, certains lieux s'inscrivent dans une activité virtuelle régulière et sont ainsi tweetés selon des rythmes identifiables (Andrienko *et al*., 2013). En outre, le contenu est capturé en temps réel : il n'y a donc qu'un délai très bref (Severo & Romele, 2015) entre le temps où l'utilisateur assiste à l'événement et crée le contenu qui est ensuite mis à disposition des internautes sur le réseau et des développeurs via les API.

- la *structuration* et la *forme* des contenus : bien qu'étant un objet d'une seule nature, le tweet contenant une information de localisation prend des formes géographiques multiples : s'il est acquis par géolocalisation directe avec un smartphone, il figure sous la forme d'un point. S'il est acquis par ajout d'un lieu général au tweet, il figure sous la forme d'un polygone représentant la *bounding box* délimitant grossièrement le lieu ajouté par l'utilisateur. Au-delà de cette variabilité de forme géométrique, le contenu du tweet soulève la question de son emprise spatiale : en effet, ce contenu peut aussi bien se rapporter à l'objet du territoire où à l'activité directe accomplie par l'utilisateur (dans un rayon de quelques mètres), qu'à un environnement beaucoup plus large (comme c'était le cas avec ce tweet mentionné précédemment "*#merdeglace #glacier #alps #france #chamonixmontblanc #beautufulbutscary #globalwarming*" et dont la photographie incluse du panorama suggérait que la dimension spatiale du tweet dépassait le simple emplacement de l'utilisateur). Dans un tel cas, la localisation du tweet sur la carte sous la forme d'un point n'a plus beaucoup de sens puisque l'emprise spatiale du tweet dépasse de plusieurs centaines de mètres la localisation de son auteur. Dans un second temps, si l'on considère la composante sémantique, tout utilisateur peut tweeter en fonction d'intentions multiples (cf. typologie proposée dans le tableau 2.1). Ce comportement a pour conséquence directe l'hétérogénéité de la composante sémantique du tweet, dont les exemples de contenus cités dans cette même typologie en indiquent les effets : alors qu'une majorité de tweets sont *égocentrés* et sémantiquement pauvres (cf. "*c'est qui le chanceux qui doit aller aux courses alors qui vient de rentrer de soirée ?*"), d'autres peuvent présenter un intérêt pour l'ensemble de la communauté (cf. "*Ligne B manifestation régionale gilets jaunes Depuis 8h00 – durée indéterminée La ligne ne circule pas entre les stations Sainte-Claire et Gares.*")

En conséquence, les contenus géolocalisés émis via les plateformes réseau comme Twitter peuvent être qualifiés de *soft data*, expression qualifiant des données produites sans protocole rigoureux ni aucun contrôle administratif généralement diffusées sur le Web social (Severo & Romele, 2015).

Quid de l'autosuffisance de ces soft data pour les problématiques socio-spatiales ?

A ce stade, il s'avère périlleux de définir, à la manière des données institutionnelles, des critères attestant de la qualité des contenus géolocalisés du Web social : par exemple, le fait d'avoir une paire de coordonnées géographiques est déjà un gage de précision spatiale par rapport à une *bounding box* ; la *bounding box* resterait en revanche plus pertinente dans le

cas particulier souligné ci-dessus où le contenu du tweet s'étend au-delà de la localisation précise de l'utilisateur. En revanche, la conséquence directe de l'hétérogénéité sémantique et des modalités d'acquisition du tweet (spontanéité et limitation à un nombre précis de caractères) s'illustre par la diversité des informations contenues dans la composante sémantique des tweets, diversité qu'on a cherchée à mettre en évidence dans la typologie proposée. Dans la plupart des cas, le tweet est un contenu personnel, créé par un utilisateur humain qui n'a pas l'intention de diffuser une information riche et précise destinée à être réutilisée dans un dessein scientifique. Il existe ainsi une dichotomie entre les pratiques des utilisateurs et les besoins rigoureux du chercheur. Le tableau 2.3 ci-dessous soumet un nouvel exemple. Les tweets (1) et (2)[37] correspondent à des contenus centrés sur l'individu, postés dans une logique horizontale. Le tweet (1) ne contient qu'une seule information environnementale relative à la survenue d'une averse ; la sémantique du tweet (2) est plus riche, elle contient trois informations relatives aux conditions environnementales, à la diffusion d'une information officielle ainsi qu'au report d'un comportement individuel face aux deux premières informations. Le tweet (3) interpelle un acteur officiel, le *VISOV*, et contient un hashtag (#MSGU[38]) permettant de rattacher le contenu créé à la gestion de crise : il s'agit donc d'un tweet de logique *bottom-up* qui contient deux types d'information : la survenue d'une perturbation environnementale en cours et ses conséquences sur les infrastructures électriques.

Tableau 2.3 : Tweets géolocalisés liés à des phénomènes météorologiques mais dont la richesse sémantique est variable

N° Tweet	Texte	Type de tweet	Nombre d'informations
(1)	"Il pleut"	Sphère privée – Destinataire non ciblé – Logique horizontale	1 (information environnementale)
(2)	"Il fait nuit, il pleut, c'est l'alerte orange mais je fais les magasins"	Sphère privée – Destinataire non ciblé – Logique horizontale	3 (information environnementale, information officielle, comportement individuel)
(3)	"Gros orage sur #Grabels, 34. eclairs en continu. quartier prédimau-montalet complètement disjoncté, élec coupée @VISOV1 #MSGU"	Informer d'un phénomène – Destinatire ciblé – Logique bottom-up	2 (information environnementale, information relative aux dommages)

[37] Ces tweets sont issus d'un jeu relatif aux épisodes cévenols survenus dans le sud de la France pendant l'automne 2014.

[38] VISOV (Volontaires Internationaux en Soutien Opérationnel Virtuel) et MSGU (Médias Sociaux en Gestion d'Urgence) seront présentés dans le chapitre suivant.

Par ailleurs, et quel que soit le nombre d'informations de nature différente contenues dans cette composante sémantique, il nous paraît essentiel de préciser que tout auteur d'un tweet retranscrit avec son propre vocabulaire et en fonction de sa connaissance du territoire ses observations de l'environnement : il s'agit donc d'une *réalité perçue*. Entre alors en jeu un processus cognitif d'interprétation des conditions environnementales du moment. Celui-ci résulte dans l'enregistrement des observations par des termes subjectifs évoquant les représentations propres à chaque individu : ainsi, dans le tableau précédent, on peut se demander ce que signifie, pour l'auteur du tweet (3) un *Gros orage* ; celui-ci nous donne une précision avec la mention d'*eclairs en continu* ; mais qu'en est-il de la pluie, du vent et a-t-il entendu de proches impacts de foudre ?

Enfin, les comportements mêmes des utilisateurs et l'effet réseau sont susceptibles de nuire à la qualité de l'information. Dans un premier temps, si les individus peuvent réagir et transmettre des observations locales en temps réel, ils peuvent également, et tout particulièrement dans le cas d'une mobilisation massive en réponse à un événement, discuter d'événements distants, auxquels ils n'ont pas personnellement assisté. Miller (2007) qualifie ce phénomène de *téléprésence asynchrone* pour décrire toute information postée par l'utilisateur situé en dehors du lieu de l'événement physique et éventuellement décalée temporellement par rapport à sa survenue. Dans un second temps, l'effet réseau présenté au paragraphe 2.1.1.2 peut s'avérer délétère pour la fiabilité du contenu sémantique des tweets géolocalisés. Précédemment, nous avons évoqué la panique provoquée dans plusieurs grandes métropoles américaines en février 2018 par des utilisateurs du réseau ayant reçu un message d'alerte test au tsunami, lu trop rapidement. La figure 2.20 propose un exemple fictif de diffusion d'information sémantique inexacte, liée à l'interprétation trop rapide d'un événement de l'environnement d'utilisateurs du réseau. Ce schéma propose une situation simple : un accident impliquant deux véhicules à un carrefour situé sur la place X : la voiture A brûle le feu rouge et percute la voiture B, il n'y a qu'un conducteur légèrement blessé. Des passants s'activent sur Twitter : l'individu 1 a été témoin direct de l'événement et rapporte la situation exacte. L'individu 1bis arrive deux minutes après l'accident, voit des véhicules de secours et pense que des personnes sont gravement blessées ; il va diffuser auprès de ses abonnés une information *bruitée*, c'est-à-dire partiellement fausse, qui est reprise par son réseau (individu 2.4). Malgré tout, les abonnés de l'individu 1, qui disposent de l'information exacte sur les conditions de l'accident, peuvent eux aussi biaiser l'information sémantique d'origine : l'individu 2.1 ne mentionne pas les détails sur les conditions de survenue de l'accident (on perd de l'information) ; l'individu 2.3 conserve ces informations mais inverse la responsabilité des conducteurs dans la survenue de l'accident : il va donc rediffuser une information bruitée, de la même manière que les utilisateurs 1bis et 2.4.

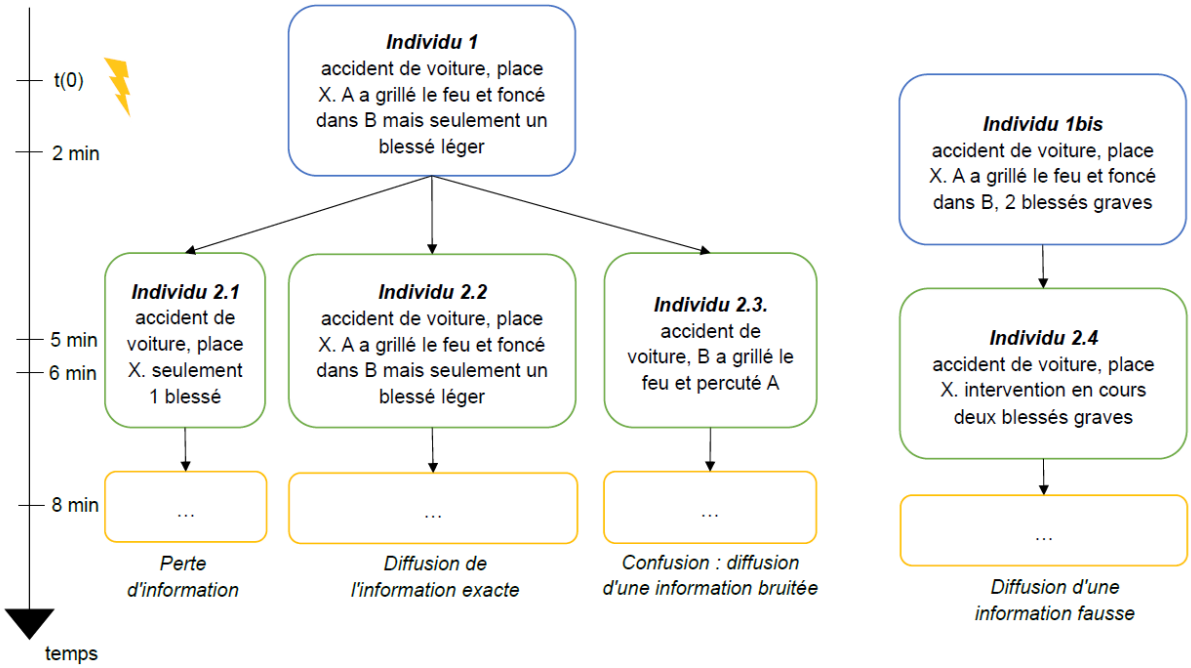


Figure 2.20 : Schématisation de la diffusion d'une information émise en réponse à un événement et du risque d'introduction de bruit

Comme l'indiquait alors (Goodchild, 2013), les tweets géolocalisés, en tant que *soft data*, s'avèrent porteurs de biais de multiples natures : ils ne pourraient ainsi faire sens qu'une fois associés à des jeux de données officiels traditionnels permettant de recontextualiser de manière objective les conditions environnementales de leur émission.

2.3.2.3. Positionnement des tweets géolocalisés dans le Web social spatialisé

Une trace numérique géolocalisée détachée de la logique volontaire.

En écho aux définitions annoncées dans le chapitre précédent, nous envisageons le tweet géolocalisé comme une trace numérique à composante spatiale dans la mesure où il témoigne de la présence d'un individu à un endroit précis à un instant connu et peut alors renseigner les pratiques spatiales de l'individu en question (Severo et Romele, 2015 ; Lucchini *et al*., 2016). Par ailleurs, nous considérons cette trace comme contributive car sa composante spatiale ne peut être acquise sans le consentement de l'utilisateur. En revanche, nous la séparons de la logique volontaire propre à la VGI : (Severo et Romele, 2015) résumaient l'enjeu central des contenus géolocalisés issus des réseaux sociaux en ces termes : "*Ces données participatives*[39] *peuvent également être utilisées à des fins non prévues par leur créateur pour créer une information sur mesure utile au décideur public".* En effet, l'intention associée à la

[39] En revanche, comme indiqué, nous n'employons pas dans cette recherche, le terme de données pour qualifier des tweets géolocalisés, ni le terme de participatif (qui est remplacé par contributif, pour les raisons évoquées dans le chapitre 1 : la production des traces numériques s'affranchit de tout cadre institutionnel).

plupart des tweets, géolocalisés ou non, ne semble pas résider pas dans une logique de co-construction d'un outil développé par les contributions des usagers de la plateforme (à l'inverse de la cartographie *OpenStreetMap* qui s'inscrit dans cette logique de construction collective d'un outil libre). En outre, nous ne connaissons pas le degré de conscience des membres de la plateforme vis-à-vis des diverses réutilisations de leurs traces par des acteurs tiers, qu'il s'agisse d'entreprives privées ou de chercheurs. Néanmoins, il faut souligner l'existence de poignées de tweets qui se démarquent de l'information destinée à faire parler de soi et qui révèlent cette intention volontaire de communiquer des informations localisées, pertinentes et utiles à d'autres acteurs de la plateforme : c'était le cas du tweet (3) présenté dans le tableau 2.3 ("*Gros orage sur #Grabels, 34. eclairs en continu. quartier prédimau-montalet complètement disjoncté, élec coupée @VISOV1 #MSGU*") qui faisait état des conditions environnementales en cours et des perturbations consécutives sur les infrastructures ; de plus, il s'adressait à un acteur officiel de la gestion de crise en utilisant un hashtag permettant d'identifier rapidement le tweet comme signal de la survenue d'une perturbation localisée.

La dimension spatiale du tweet géolocalisé est-elle géographique ?

En ce qui concerne la pertinence de la dimension spatiale des traces, les avis de la communauté ont évolué : si les premières recherches engagées analysaient la trace numérique géolocalisée sans aborder les biais liés aux conditions de sa production, les avis se révèlent désormais plus partagés. (Severo et Romele, 2015) présentaient la géolocalisation comme une porte d'entrée dans la compréhension des dynamiques spatiales collectives à condition que celle-ci soit recoupée aux autres dimensions des traces numériques, et notamment à leur composante sémantique, trop souvent négligée (Steiger *et al*., 2015).

En revanche, (Quesnot, 2016) présentait les traces numériques géolocalisées comme des produits commerciaux dont la dimension géographique n'était que *miroir aux alouettes* en raison des nombreux biais sociaux, spatiaux, comportementaux et finalement statistiques qu'elles portent. Il le soulignait alors en ces termes : le caractère géographique d'une infomation englobe un espace physique (renseigné effectivement par les traces géolocalisées) mais également les interactions entre cet espace et l'humain; et sur ce point, la trace numérique est prise en défaut par l'absence d'éléments de contexte indiquant les raisons pour lesquelles on tweete dans un lieu précis à un moment donné plutôt qu'en un autre lieu, à un moment différent. A cette argumentation, on pourra reprocher le fait qu'elle ne tienne pas compte des traces numériques géolocalisées dont le discours est en rapport direct avec un objet perçu du territoire : on peut considérer le tweet déjà mentionné précédemment ("*#merdeglace #glacier #alps #france #chamonixmontblanc #beautufulbutscary #globalwarming*") comme un contenu géographique : le tweet contient une géolocalisation par GPS, une photographie qui capture l'objet en temps réel ainsi que le ressenti de l'individu face à l'objet du territoire. En revanche, il est très probable que l'activité virtuelle liée à ce glacier soit sur-représentée par rapport à d'autres glaciers tout autant menacés dans le massif

du Mont-Blanc, mais qui ne sont pas accessibles par le train ou le téléphérique (et donc non fréquentés par les touristes). Dans un tel cas, même si les tweets peuvent alors arborer une dimension géographique, le chercheur se retrouve en présence d'un biais quantitatif lié à la fréquentation d'un objet accessible à tous et ancré comme témoin dans l'imaginaire collectif, au dépit des objets environnants.

Face à ces difficultés et avis divergents, le travail de recherche entrepris se focalisera également sur la question la pertinence de la dimension spatiale de ces traces dans un contexte d'événement naturel. Il s'agit en effet de savoir si les utilisateurs producteurs de tweets géolocalisés pendant les crises d'origine naturelle émettent des contenus relatifs à leur pratiques ou représentations du territoire en crise, et de quelles manières valoriser ces contenus en données géographiques en réduisant ces biais de représentativité.

Positionnement des tweets géolocalisés dans les Big Data.

Enfin, le tweet géolocalisé s'inscrit dans la lignée des propriétés des *Big Data* : 300 milliards de tweets ont été émis depuis le lancement de la plateforme le 21 mars 2006, le trafic quotidien est estimé à 500 millions[40] de tweets et si l'on considère le taux de tweets géolocalisés à 0,85% de ce flux, le nombre de contenus géolocalisés créés chaque jour serait de 4,25 millions. En outre, leur exploitation soulève des interrogations quant aux postures épistémologiques adoptées, ainsi qu'aux outils et méthodes employés pour *faire parler les données*[41] (Miller & Goodchild, 2015). Pour autant, les tweets géolocalisés utilisés dans la recherche géographique se démarquent des propriétés des *Big Data* sur deux points :

- si les *Big Data* sont parfois associées à l'exhaustivité (Anderson, 2008), les géographes ne peuvent prétendre à cette même propriété (Quesnot, 2016) en raison des biais d'acquisition et de représentation évoqués dans les paragraphes précédents.

- la plupart des travaux en sciences humaines et sociales ne travaillent qu'avec une infime partie de ce flux qui ne représente au final que quelques gigaoctets de volume : à titre d'exemple, les travaux de (Sloan et Morgan, 2015) cités précédemment sont basés sur un jeu contenant un peu plus de 7 millions de tweets géolocalisés collectés au Royaume-Uni en avril 2015. Dans cette recherche, la table de notre base de données contenant l'ensemble de tweets bruts collectés au Texas entre avril et juin 2016 contient 2,25 millions de tweets géolocalisés et pèse 440 mégaoctets : on est donc loin des téra voire pétaoctets de données. De plus, lorsque les tweets sont filtrés pour répondre à des thématiques particulières (et il en est le cas dans cette recherche), ces volumes diminuent encore : l'étude de (Dashti *et al.*, 2013) sur les inondations au Colorado entre le 11 et le 20 septembre 2013 est bâtie à partir d'un jeu de 212 672 tweets géolocalisés.

[40] Source : http://www.blogdumoderateur.com/chiffres-twitter/ (Consulté pour la dernière fois le 19/03/2019)
[41] Postures et méthodes sont introduites dans le prochain chapitre.

Si la plupart des travaux scientifiques thématiques ne travaillent donc qu'avec des volumes réduits de traces numériques, les difficultés rencontrées dans la valorisation de ces matériaux nous paraissent davantage liées à leur nature flexible et non structurée qu'à leur volume : en cela, s'il est aujourd'hui possible de transformer un jeu de tweets bruts en un contenu plus élaboré, nous interrogeons la possibilité de construire une information géographique recoupant plusieurs sources de données dont les tweets géolocalisés, à la manière des cartes de vigilance météorologique ou de zonage des risques naturels présentées dans le chapitre 1.

Conclusion du chapitre 2

A l'issue de ce deuxième chapitre, nous pouvons dresser le constat suivant : le tweet géolocalisé se présente sous la forme d'une trace numérique à composante spatiale au caractère contributif mais généralement involontaire, faisant partie des *soft data*. A ce titre, l'assertion redondante de la littérature scientifique selon laquelle ces *soft data* ouvriraient de nouvelles perspectives pour les sciences humaines dans la mesure où elles capturent les intérêts des individus qui les produisent au lieu de répondre spécifiquement aux intérêts énoncés par les chercheurs ou acteurs du territoire (dans le cadre des enquêtes traditionnelles), s'est finalement imposée comme un *credo.* Et pour cause, les tweets géolocalisés, en tant que traces enregistrées par smartphone, offrent l'avantage d'assurer le suivi des activités et pensées de l'utilisateur en situation de mobilité ; en d'autres termes, ces tweets capturent les interactions entre l'individu et l'espace vécu.

Toutes ces considérations restent pour autant très théoriques et l'enthousiasme des chercheurs qui s'annonçaient plutôt optimistes tend à se tarir (Severo & Romele, 2015), notamment en raison des multiples biais liés aux conditions actuelles d'adhésion à la plateforme et de contribution à la production de contenus géolocalisés. Ainsi, si les tweets géolocalisés ne peuvent capturer les pratiques spatiales des individus dans leur globalité, ils ne constituent pas non plus un échantillon statistiquement représentatif :

- d'une population mère : seule une partie des individus est consommatrice régulière des réseaux sociaux, et parmi ces individus, un faible pourcentage recourt à la géolocalisation. En outre, les profils de ces individus tendent à varier en fonction des territoires considérés et des activités pratiquées.
- d'un espace : les lieux qui concentrent les émissions de tweets sont régulés en fonction de facteurs multiples difficiles à identifier dans leur globalité (Quesnot, 2016) : accessibilité des personnes aux TIC, densités de populations en fonction du temps, présence d'un objet suscitant une réactivité générale des utilisateurs de la plateforme, etc.. En conséquence, quelle que soit l'échelle géographique considérée, tous les territoires ne sont alors pas équitablement représentés.

Quelles conséquences pour les situations de catastrophes naturelles ? Après la survenue de l'ouragan Harvey en août 2017, le chercheur en géomatique Anthony Robinson émettait le tweet suivant : *"The tragedy in Texas will magnify the fact that the most vulnerable ppl are not going to tweet about it. Response can't be driven by tweets."* En effet, l'ensemble des pratiques et des biais présentés dans ce chapitre tendent à indiquer l'existence d'une sélection des territoires et des populations par le numérique, ce qui rejoint une question déjà évoquée dans l'introduction de ce manuscrit : pour qui va-t-on construire la connaissance des territoires par le numérique ?

Malgré leurs défauts actuels, les traces numériques géolocalisées constituent néanmoins une alternative assurant l'obtention d'informations acquises sur le terrain par les

individus lambdas confrontés au phénomène naturel et aux événements qu'il entraîne (et non par les secouristes ou par les observateurs formés à la capture d'informations destinées à des acteurs officiels). On peut résumer l'intérêt de cette source alternative selon deux points centraux :

- ces individus-capteurs forment un réseau connecté ayant la capacité d'enregistrer une information spontanée acquise en temps réel (à la différence des questionnaires d'enquêtes établis par des experts et distribués dans un temps différé par rapport à la survenue du phénomène) ;
- en conséquence, l'intérêt majeur pour le chercheur réside dans la captation d'une information de terrain dans un contexte particulier où le géographe manque de données relatives aux conditions environnementales et dynamiques de réponses locales (soit autant d'informations que les traces numériques géolocalisées générées par ceux qui vivent directement le phénomène et ses effets ont le potentiel de capturer).

Sur le plan de la cartographie, nous considérons que l'enjeu de l'analyse des tweets géolocalisés réside ici : les professionnels construisent de l'information géographique en recoupant, analysant, cartographiant différentes sources de données acquises par d'autres professionnels qui respectent des protocoles rigoureux. Ceux qui se saisissent des traces numériques géolocalisés involontaires (donc de matériaux qui ne correspondent pas aux exigences des disciplines) peuvent-ils alors, par leurs méthodes habituelles, valoriser ces traces jusqu'à l'obtention d'une information géographique interprétée et précisément délimitée sur la carte ?

3. Le tweet comme matériau de construction de connaissances sur la société numérique naissante

Dans leur revue de littérature dédiée aux travaux scientifiques d'analyses spatio-temporelles effectuées à partir du matériau tweet géolocalisé, (Steiger *et al.*, 2015) dressaient le constat suivant : *"[…] a constantly increasing amount of Twitter research articles have been published during the reviewed time period (01/01/2005 – 30/09/2013)."* Bien que les références citées dans cette revue aient désormais entre cinq à dix ans[1], elles témoignent de cet engouement qu'avait suscité, dans l'ensemble de la communauté scientifique, la disponibilité de ce nouveau contenu numérique capté par tout un chacun, en temps réel et sur le terrain : "*[…] the potential of Twitter has been increasingly recognized by numerous research domains over the last years*." (Steiger *et al.*, 2015).

Ce chapitre présente ainsi une revue des thèmes et questions abordées autour du matériau tweet, selon ses différentes composantes (spatiale, temporelle et sémantique), et en fonction des disciplines qui l'ont inclus dans leurs pratiques. Dans un premier temps, le chapitre expose les différents cas d'étude qui se sont succédés dans les premières années de la recherche, qu'il s'agisse de tweets disposant d'une composante spatiale ou non. La publication chronologique des papiers indique en effet que les tout premiers temps de la recherche scientifique se sont attelés à l'analyse des comportements des utilisateurs de Twitter, avant d'appréhender ces mêmes comportements dans l'espace et dans le temps, grâce à la mise en service de la fonctionnalité de géolocalisation. Dans un deuxième temps, l'exposé se focalise sur les pratiques de l'utilisation des tweets géolocalisés vis-à-vis de la question des risques naturels, dans laquelle on distingue généralement trois approches : la détection précoce de phénomènes physiques, la gestion de crise en temps réel et l'analyse virtuelle post-crise.

Enfin, nous aborderons la question des outils et méthodologies existants pour l'analyse et la cartographie du matériau tweet ; dans ce dernier point, il s'agira également de soulever les récents doutes émis par certains chercheurs quant à la possibilité d'approfondir les connaissances existantes sur les tweets, ainsi qu'à l'adéquation entre les propriétés de ce nouveau matériau et nos méthodologies et outils d'analyse traditionnels.

[1] Sur ce point, il faut d'ailleurs réfléchir à ce qui sera considéré comme *vieux* : alors que les tweets géolocalisés n'ont que dix ans, faut-il considérer qu'un article publié en 2012 est daté et que son contenu est obsolète, quand certaines références de plus de trente ans font toujours autorité dans les disciplines académiques ? D'un autre côté, il faut garder en mémoire que les pratiques numériques sont susceptibles d'évoluer rapidement : un fait constaté de manière répétée sur Twitter il y a sept ans est-il toujours d'actualité aujourd'hui, en 2019 ?

3.1. Thèmes récurrents d'utilisation des tweets dans la recherche académique générale

En dehors de la géolocalisation et de la dimension spatiale des tweets, une première série de publications s'est focalisée sur les comportements des utilisateurs à travers l'étude des rythmes des émissions virtuelles articulées au monde réel. La question de recherche centrale est donc la suivante : dans quels contextes les utilisateurs tweetent-ils et comment ?

3.1.1. Analyse des comportements par la réactivité et la sensibilité virtuelles

Au travers des premières publications académiques, on a en effet rapidement mis en exergue qu'outre les messages constituant la routine quotidienne du réseau virtuel, qui diffusent un contenu généralement rattaché à la sphère privée de l'individu et de ses activités, il existait une articulation entre les événements réels perturbant le quotidien de l'individu et les sujets évoqués dans les tweets. En fait, l'activité tweeting se positionne comme écho des phénomènes inhabituels ou générateurs de perturbations du quotidien dans le monde réel. (Lee Hughes et Palen, 2009) font partie des premiers auteurs qui se sont intéressés à l'usage de Twitter dans le contexte particulier de la survenue, à petite échelle géographique, de phénomènes rompant cette routine, et affectant des millions d'individus. Les phénomènes en question, survenus en 2008, correspondaient à deux événements politiques, les congrès nationaux des partis démocrate et républicain avant la présidentielle des Etats-Unis, et à deux catastrophes naturelles, les ouragans Gustav et Ike. Leur étude mettait ainsi en évidence d'une part, l'existence d'une variabilité quotidienne des émissions mais un pic atteint le jour même de la survenue du phénomène (lorsque l'ouragan touchait terre ou que le congrès avait lieu), et d'autre part, en ce qui concerne les ouragans, l'existence d'une corrélation entre intensité des dégâts et quantité de tweets émis. L'approche focalisée sur les utilisateurs indiquait que 95% des utilisateurs actifs pendant l'un des phénomènes étudiés envoyaient moins de sept tweets. Au cours du temps, le nombre d'utilisateurs engagés sur le phénomène diminuait alors que le nombre de tweets émis augmentait : certains utilisateurs diffusaient donc massivement les informations alors que la majorité participait de manière ponctuelle. En cela, les comportements des utilisateurs semblent se plier à la *règle du Web du 1%*[2].

Ce premier constat illustrant d'une part, l'existence d'une sensibilité des usagers de Twitter face aux phénomènes du monde réel et d'autre part, l'existence de rythmes temporels d'émission marquant la réactivité des utilisateurs, se positionne finalement comme une loi,

[2] La règle du 1% ou le principe du 90-9-1 désigne les comportements de contribution des utilisateurs des plateformes en ligne : moins de 1% contribue de manière active, 9% contribuent de manière occasionnelle et 90% des utilisateurs inscrits restent passifs et ne produisent pas de contenu. Source : https://fr.wikipedia.org/wiki/R%C3%A8gle_du_1_%25 (Consulté pour la dernière fois le 15/07/2019)

soit un comportement inéluctable des réseaux sociaux numériques. (Lin *et al*., 2013) ont ainsi approfondi la question comportementale liée à l'activité tweeting en se focalisant sur les dynamiques de diffusion et la durée de vie des hashtags. Leur étude est alors axée sur le cycle de vie de 256 hashtags créés pendant la campagne présidentielle américaine de 2012, et notamment apparus sur le réseau en réponse à des maladresses diffusées en direct via les médias traditionnels. Dans un premier temps, il s'est avéré que la réactivité des utilisateurs sur Twitter, et par conséquent la création de hashtags, étaient une conséquence immédiate de la consultation d'un média traditionnel (par exemple, une maladresse exprimée lors d'un débat télévisé) : la création du hashtag arrive en réponse immédiate à la *gaffe* politique du candidat évoquant un sujet sensible pour les spectateurs et peut s'affirmer comme tendance populaire du réseau dans les quinze minutes suivant sa création. Ce phénomène virtuel s'illustre, d'après les auteurs, notamment par les retweets : en fait, une poignée de tweets font le *buzz* et sont massivement rediffusés par l'ensemble de la communauté Twitter. Mais surtout, les auteurs ont identifié l'existence de rythmes temporels de diffusion des hashtags, en représentant le nombre de tweets émis par minute, contenant le hashtag d'analyse (figure 3.1).

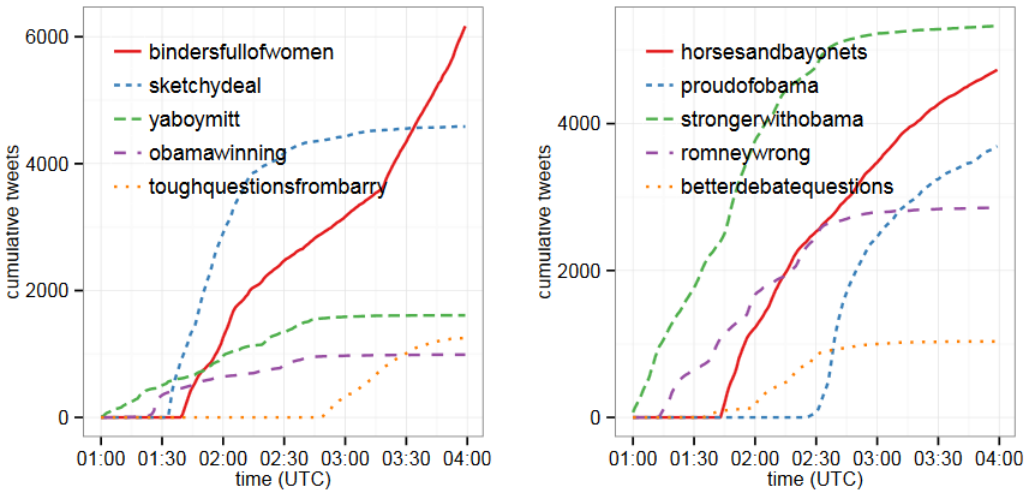


Figure 3.1 : Rythmes temporels d'émissions de hashtags créés en réponse aux débats de la campagne présidentielle américaine de 2012 (Source : Lin et al., 2013)

Certains hashtags manifestent ainsi une croissance rapide (généralement en moins d'une heure) et atteignent une saturation rapide (la saturation étant définie comme le moment où 99% des tweets contenant le hashtag en question ont été émis). Cette saturation est visible sur les courbes présentant un point d'inflexion marquant un pallier : par exemple, sur la figure 3.1, on pourra observer cette situation sur la courbe du hashtag *#sketchydeal* (graphique de gauche), dont la saturation s'amorce environ trente minutes après l'apparition du hashtag, ou encore sur la courbe du hashtag *#romneywrong* (graphique de droite) dont la saturation apparaît plus tardive (environ 1h30 après l'enregistrement du hashtag). A l'inverse, d'autres hashtags témoignent d'une croissance rapide et d'un saturation lente (la courbe ne

présente pas de point d'inflexion, indiquant ainsi que le hashtag ne s'essouffle pas sur le réseau) : toujours sur la figure 3.1, c'est par exemple le cas du hashtag *#bindersfullofwomen* sur le graphique de gauche ou encore du hashtag *#horsesandbayonets* du graphique de droite, tous deux continuant de s'accroître plus de trois heures après leur première émission.

Enfin, les études de cas consultées montrent que cette sensibilité collective des utilisateurs, qui s'illustre dans l'inscription des phénomènes réels sur le réseau virtuel, est également valide dans le cas de phénomènes sociaux ayant lieu à une échelle géographique locale : la criminalité en milieu urbain fait ainsi partie de ces thèmes explorés entre 2009 et 2015. (Kounadi *et al.*, 2015) se sont intéressés aux variations temporelles de la réactivité des utilisateurs de Twitter vis-à-vis des homicides commis à Londres en 2012. Leurs résultats affichent des tendances identiques à celles des auteurs précédemment cités (figure 3.2) : 80% des tweets sont émis dans le mois suivant l'homicide. En revanche, les rythmes d'émission s'avèrent moins rapides que dans le cas des hashtags générés par la réactivité des utilisateurs face aux maladresses politiques : ici, il faut compter une semaine pour atteindre 80% des émissions évoquant l'événement considéré.

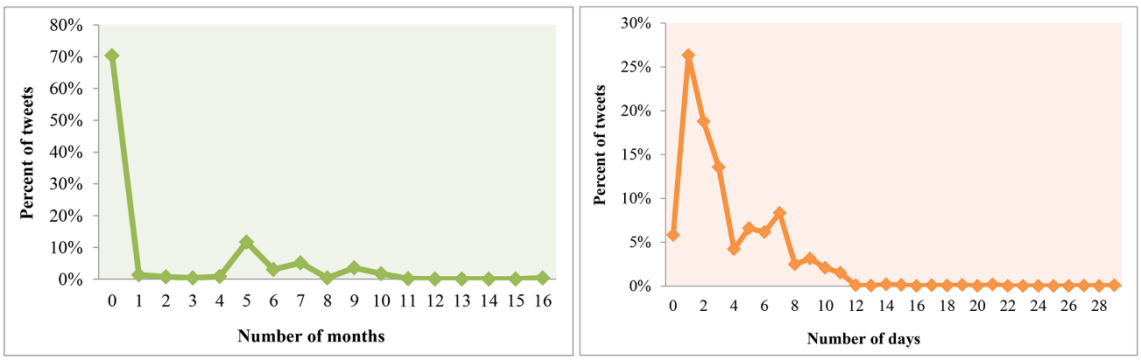


Figure 3.2 : Distribution temporelles des réponses Twitter aux homicides, par mois et par jour (Source : Kounadi et al., 2015)

Par ailleurs, les auteurs soulignaient un second phénomène remarquable dans les rythmes d'émission : la résurgence, quelques mois après l'homicide, de l'évocation de l'événement en question sur le réseau virtuel, qui se révélait liée à un événement consécutif du crime (procès, aveux, etc.)[3]. Au niveau de la sensibilité des utilisateurs, les auteurs étaient parvenus à identifier des facteurs influençant les quantités de tweets émis en réponse à des événements, et notamment la nationalité et l'âge de la victime[4].

[3] Ce phénomène de résurgence se constate de manière récurrente : par exemple, à chaque date anniversaire d'un événement traumatisant, comme un attentat, on va observer la présence de hashtags commémorant l'événement dans les tendances Twitter.

[4] Un crime impliquant une victime jeune ou de nationalité britannique drainait ainsi une réponse plus volumineuse qu'une victime étrangère, plus âgée ou appartenant à un gang.

A côté des événements extraordinaires générateurs de dynamiques d'émission particulières, d'autres recherches focalisent leur objet d'étude sur des pans qui constituent l'information virtuelle routinière. Et l'un de ces pans routiniers, ancré dans des enjeux sociaux majeurs actuels, qui a attiré les chercheurs, reste l'articulation entre santé, nutrition et environnement (Widener et Wenwen, 2014 ; Chen et Yang, 2014 ; Ghosh et Guha, 2013 ; Gomide *et al*., 2011) .

3.1.2. Le tweet comme capteur des problématiques sociales contemporaines en santé, nutrition et environnement

Certains géographes, principalement aux Etats-Unis, se sont appliqués à l'analyse des discours des usagers de Twitter, relatifs à une problématique d'ampleur dans les sociétés occidentales de nos jours : l'accès à une nourriture saine et non transformée. (Widener et Wenwen, 2014) se positionnent comme les premiers géographes à tenter de croiser des tweets géolocalisés avec des données émanant des institutions (dans la problématique de la santé et de l'accès à une nourriture de qualité), soit une méthode préconisée par les conclusions de la revue de littérature de (Steiger *et al*., 2015) : leur objectif consiste alors à vérifier si les discours portant sur les *unhealthy foods* (aliments ultra-transformés gras/sucrés) sont plus ancrés dans les territoires dont l'accès à une nourriture saine et non transformée est difficile (territoires aux revenus médians bas et marqués par la présence de déserts alimentaires[5]). Leurs résultats indiquent en effet que dans ces déserts alimentaires, les tweets mentionnant les comportements alimentaires apparentés à la malbouffe apparaissent plus nombreux que les tweets traduisant des habitudes alimentaires saines ; à l'inverse, dans les lieux où l'accès à une nourriture de meilleure qualité nutritive est garanti, on trouve davantage de tweets exprimant des comportements alimentaires sains.

L'étude complémentaire et localisée à l'échelle d'une métropole publiée par (Chen et Yang, 2014) permet néanmoins de nuancer des constats qui tendent à établir des liens de cause à effet systématiques entre les profils socio-économiques des populations d'un territoire et leurs comportements alimentaires. Cette étude se focalise sur l'influence directe de l'environnement voisin de l'usager de Twitter, dans une situation de mobilité quotidienne, sur ses choix alimentaires ; en d'autres termes, un utilisateur entouré de *fast foods* va-t-il se restaurer dans ce type d'établissement ou préfère-t-il aller chez le vendeur de produits frais ? Et en effet, leurs résultats indiquent que les choix alimentaires ne correspondent pas

[5] Aux Etats-Unis, l'*USDA* (United States Department of Agriculture) définit les déserts alimentaires (*food deserts*) comme les territoires d'un comté dans lesquels il est impossible d'acheter des produits alimentaires frais (non transformés). En milieu urbain, on considère un désert alimentaire lorsque 33% des résidents vivent à plus de 1 mile de distance d'un magasin où ils peuvent trouver ces produits frais (soit 1,6 km) ; en milieu rural, ce seuil augmente à 10 miles (soit 16 km). Source : http://americannutritionassociation.org/newsletter/usda-defines-food-deserts (Consulté pour la dernière fois le 16/07/2019)

forcément à des stéréotypes qu'on peut escompter : il s'avère que l'environnement local[6] de l'utilisateur qui tweete n'explique que partiellement les choix alimentaires. Il apparaît que la proximité d'un environnement alimentaire marqué par la malbouffe n'est pas corrélée avec de mauvaises habitudes alimentaires : si la présence d'établissements vendeurs de produits frais peut inspirer la création des tweets voisins affichant une volonté de recherche de nourriture saine, d'autres utilisateurs pourtant moins entourés de *fast foods*, manifestent clairement leur préférence pour les aliments de la malbouffe.

Malgré ces observations, il reste deux écueils à souligner dans ces approches utilisant les tweets pour des questions socio-spatiales : le premier, souligné par (Ghosh et Guha, 2013) rejoint le constat établi dans le chapitre précédent : nous ne disposons que de peu de conclusions pour caractériser les populations productrices de tweets géolocalisés. Leur étude soulignait par ailleurs que les individus évoquant le problèmes de l'obésité et de la malbouffe dans les tweets correspondaient en fait à des individus présentant un niveau d'éducation élevé et parmi lesquels le problème de l'obésité restait marginal. Le second écueil était évoqué par (Steiger *et al*., 2015) et concernait l'ensemble des articles relus dans leur revue de littérature : les Etats-Unis constituaient le terrain privilégié des analyses[7]. En ce qui concerne l'articulation entre santé, nutrition et environnement, on peut faire le même constat : en consultant d'autres références (Gore *et al*., 2015 ; Huang *et al*., 2019), on peut rapidement s'apercevoir que les Etats-Unis restent le terrain prédominant alors que la problématique concerne l'ensemble des sociétés occidentales ainsi que les pays émergents.

Pour autant, la problématique santé/environnement a pu être abordée dans des terrains différents et concernant également un enjeu sanitaire d'échelle mondiale : la propagation des virus transmis par le moustique tigre. (Gomide *et al*., 2011) proposaient d'analyser les tweets mentionnant les cas de dengue survenus au Brésil entre 2009 et 2011. Leur approche ne consistait pas à juxtaposer, par la cartographie, traces numériques issues des réseaux sociaux et jeux de données officiels afin de recontextualiser l'environnement dans lequel les tweets sont émis, mais à comparer données officielles et tweets afin d'évaluer leur pertinence et leur fiabilité comme source d'information à distance (en d'autres termes, le recours aux données officielles intervient après l'analyse des traces numériques, afin de vérifier la validité des informations extraites depuis les tweets). Le signal lancé par le nombre de tweets émis contenant le mot-clé *dengue* apparaît fiable pour lancer une alerte, à l'échelle d'une ville, dès que les cas de dengue se multiplient : en comparant les effectifs de tweets émis entre deux villes de taille identique (Manaus et Belo Horizonte), les auteurs constatent que l'agitation virtuelle autour de la dengue coïncide temporellement avec la variabilité

[6] Le *local* est mesuré par construction d'une zome tampon de 0,5 mile (0,8 km) et de 1 mile (1,6 km) autour de chaque tweet.

[7] Cela peut également s'expliquer en regardant les statistiques d'utilisation de Twitter. Pour rappel, on pouvait constater un tournant amorcé à partir de 2015-2016 ; l'utilisation actuelle du réseau tend à stagner dans les pays occidentaux alors qu'elle connaît une croissance rapide dans les pays de l'Asie du Sud, du Golfe et d'Amérique du Sud.

épidémique saisonnière enregistrée dans les données officielles. En outre, les auteurs observent une très forte corrélation positive entre le nombre de tweets émis dans les villes et le nombre de cas de dengue recensés.

Les cas d'étude présentés ici, qu'il s'agisse des problématiques épidémiologiques liées aux pratiques alimentaires ou à des agents infectieux, mettent en évidence un fait incontestable : c'est dans les territoires urbanisés qu'on trouve une majorité de tweets géolocalisés. Ces territoires urbains se sont donc rapidement imposés comme les territoires de prédilection des études de cas.

3.1.3. Appréhender l'espace vécu par les tweets géolocalisés

(Andrienko *et al*., 2013) publiaient l'une des premières études à appréhender le territoire métropolitain dans sa globalité en analysant conjointement les composantes spatiale, temporelle et sémantique des tweets géolocalisés. En analysant les lieux et rythmes temporels de l'activité tweeting de Seattle, ils présentaient le tweet géolocalisé comme une approche alternative aux données traditionnelles pour détecter et caractériser des comportements individuels ou collectifs variant dans le temps et dans l'espace (en se basant sur l'hypothèse que les comportements numériques reflètent les modes de vie des utilisateurs de la plateforme). Les auteurs s'intéressaient alors au discours véhiculé par les tweets en fonction des lieux et de la variabilité temporelle, ainsi qu'à la répétitivité des structures identifiées. Après avoir classé les tweets collectés en fonction de thèmes détectés par analyse sémantique (maison, transports, éducation, nourriture, musique, jeux, amis, santé, etc.), l'analyse exploratoire mettait en évidence l'existence d'une cohérence spatiale et temporelle entre le thème des tweets et les lieux et heures d'émission : les thèmes de l'éducation et des sports sont majoritairement présents dans les complexes sportifs et le quartier universitaire ; concernant les boissons consommées, le thème du café[8] est omniprésent dans le centre de la métropole (*CBD*), sauf dans le quartier asiatique où le thé domine. Côté temporel, le thème des transports témoigne de pics de tweets du lundi au vendredi entre 6h et 9h puis entre 15h et 18h, soit lorsque les individus sont en situation de mobilité pour rejoindre ou quitter leur lieu d'étude/travail. Les tests effectués dans cette publication témoignent également d'une autre propriété du tweet géolocalisé : il est *témoin direct* des activités de l'individu (c'est-à-dire que l'utilisateur tweete au moment où il fréquente le lieu et accomplit telle activité, et non en différé).

La question des lieux de l'activité tweeting et des rythmes d'émission est également abordée par (Lucchini *et al*., 2016) sous l'angle de la réactivité spatio-temporelle des utilisateurs face à la survenue d'événements perturbateurs, et non détectables par les données traditionnelles. Selon une approche qui se veut quantitative, le tweet géolocalisé est

[8] Les auteurs rappelaient à ce propos que Seattle reste la ville dont les habitants consomment le plus de café aux Etats-Unis.

envisagé comme un marqueur des *pulsations urbaines*, c'est-à-dire qu'il a la capacité d'enregistrer la variabilité des lieux d'attractivité des métropoles (en l'occurrence, il s'agit de Paris). Les auteurs mettent alors en exergue l'existence de la juxtaposition de plusieurs activités tweeting liées à des événements variés dont le sens dépend des lieux et des profils d'utilisateurs : les émissions de tweets liées au tourisme représentent un événement individuel qui se traduit par un flux localisé constant en des lieux précis. En ce qui concerne l'activité des locaux, il est possible de détecter des événements collectifs ponctuels programmés ou spontanés (concerts / regroupement du mouvement Nuit Debout sur la place de la République), ou encore de détecter une activité virtuelle anormale : au lendemain des attentats du 13/11/2015, la ville apparaît en sous-activité virtuelle globale, à l'exception de deux lieux : la Tour Eiffel et l'hôtel de ville (alors lieu d'information pour rechercher des proches). Les auteurs proposent ainsi une typologie des événements d'origine sociale qu'on peut détecter au travers de l'examen des variations quantitatives des émissions de tweets géolocalisés, en fonction de leur localisation (ponctuelle, multi-sites ou diffuse) ainsi que de leurs temporalités d'exercice (cyclique, répétée ou spontanée).

Qu'observe-t-on dans les métropoles de l'Asie (soit dans les pays où l'adhésion à la plateforme Twitter s'accroît régulièrement depuis quelques années) ? (Cebeillac *et al*., 2017) proposent de caractériser les discontinuités spatiales de la métropole de Bangkok à travers les activités et mobilités enregistrées sur le réseau virtuel, qui dessinent des vides, des pleins ainsi que des espaces de transition. L'étude des variations temporelles de la densité d'utilisateurs met alors en évidence l'existence d'un modèle d'activité tweeting fondé sur une logique centre-périphérie répétée, mais soulignant une variabilité spatio-temporelle des lieux centraux générateurs d'activité virtuelle. Par ailleurs, l'analyse des flux origine/destination permet de mesurer les dynamiques d'attractivité des lieux de la métropole, mettant en évidence d'une part, le centre comme capteur essentiel des flux d'origine domiciliaire ainsi que l'émergence de pôles concurrents dans les couronnes périurbaines, et d'autre part, l'existence de quartiers en marge des dynamiques virtuelles.

Rythmes spatio-temporels et facteurs d'émission ont également été étudiés à l'échelle du globe ; (Hawelka *et al.*, 2013) examinent les configurations des mobilités internationales liées aux flux migratoires ou touristiques qu'on peut détecter à travers les émissions de tweets géolocalisés sur l'année 2012. L'analyse reste toujours animée par les motivations des auteurs précédemment cités : explorer le potentiel des traces numériques afin de les positionner comme source alternative aux données traditionnelles, éparses ou chères (enquêtes de déplacement, statistiques de voyageurs ou données du trafic aérien international). Dans l'analyse, l'utilisateur mobile est considéré comme celui qui est sorti du pays de résidence qui lui est assigné. La quantification des flux temporels entrants/sortants par pays met en évidence l'existence de tendances universelles (la mobilité reste assujettie à des contraintes géographiques : les utilisateurs insulaires ou résidant dans des pays en développement sont moins mobiles à l'échelle du globe ; la mobilité augmente pendant l'été et à la fin du mois de décembre), qui contrastent avec des tendances régionales s'expliquant par des facteurs

culturels (figure 3.3). Dans les pays occidentaux, les pics de mobilité s'observent en été, en décembre ainsi qu'en mai. Dans les pays du Golfe, les utilisateurs restent peu mobiles en été et pendant le Ramadan. Au contraire, le pic de mobilité est enregistré pendant le pèlerinage de La Mecque.

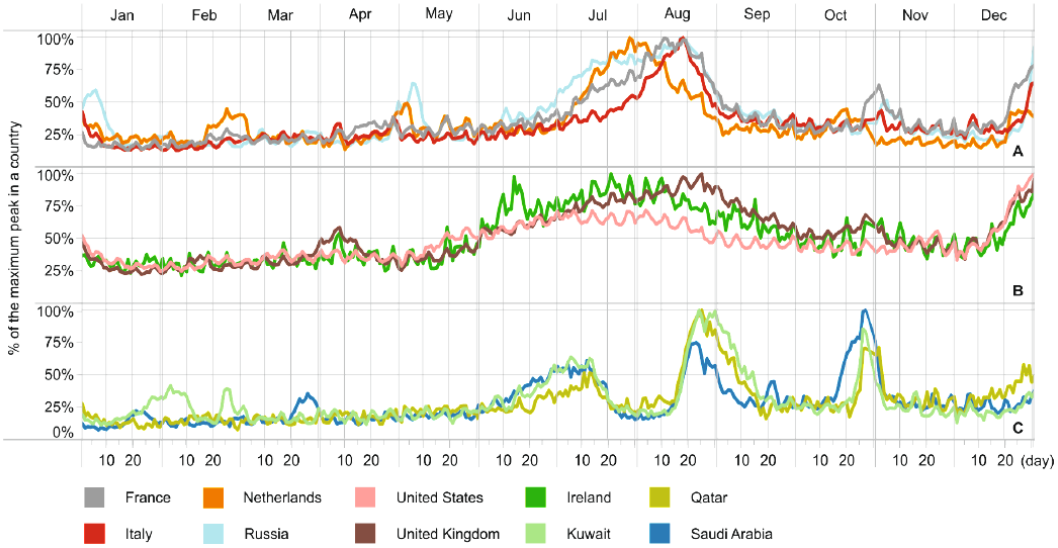


Figure 3.3 : Flux d'utilisateurs sortant de leur pays de résidence, exprimés en pourcentages (Source : Hawelka et al., 2013)

Par ailleurs, l'analyse des flux en fonction des pays d'origine et des pays de destination révèle une tendance à la cohésion régionale et ce, à diverses échelles : les flux intra-américains et les flux intra-européens s'avèrent plus forts que les flux transatlantiques ; de même, les flux entre pays d'Europe de l'ouest sont plus importants que les flux entre l'ouest et l'est. Finalement, dans le souci d'examiner la cohérence de ces résultats avec le terrain, les auteurs ont identifié l'existence d'une forte corrélation linéaire entre les flux entrants par pays et les données officielles enregistrant le nombre de touristes ainsi que les recettes générées par le tourisme.

(Salas Olmedo *et al.*, 2017) ont justement analysé les comportements spatiaux liés à l'activité virtuelle émanant des touristes à l'échelle de la métropole de Madrid (ceux-ci générant des quantités importantes de traces numériques dans les lieux visités, pourtant peu étudiées). L'originalité de leur approche tient à ce que ces auteurs ont croisé trois sources de traces numériques et testé leur capacité à s'autocompléter afin de capter les territorialités du tourisme. Les trois sources de traces sont ainsi Twitter, Foursquare (média social permettant d'indiquer à l'utilisateur l'endroit où il se trouve via un *check-in* qui le géolocalise, Foursquare est généralement utilisé dans les lieux de consommation) et Panoramio (site de photographies géoréférencées : fermé sous son nom de domaine en 2017, il ne reste accessible qu'aux dépositaires d'un compte Google+). Ces trois sources ont révélé à la fois des redondances et des complémentarités : les plus fortes densités d'utilisateurs des trois médias sociaux se

trouvent dans le centre historique de la ville. Les utilisateurs de Twitter s'étendent néanmoins le long des principaux axes de la ville (le long desquels se situent les locations et les hébergements touristiques) alors que ceux de Panoramio gagnent le nord (stade et quartier des affaires). Les modèle de régression et coefficients de détermination calculés[9] confirment cette tendance à la complémentarité des données en ce qui concerne l'identification des lieux fréquentés par les touristes : les coefficients sont positifs mais faibles entre Twitter et Panoramio, ainsi qu'entre Foursquare et Twitter. Et en effet, par l'analyse des résidus des régressions, il s'avère que Foursquare présente un surplus d'utilisateurs dans le centre historique de la ville ; à l'inverse, c'est Twitter qui présente plus d'activité dans les lieux périphériques de la ville par rapport aux deux autres sources. Comment identifier alors les activités des touristes en fonction des lieux ? Les auteurs proposent une clustérisation des unités administratives en fonction du profil dégagé (le profil correspondant à l'activité majoritaire émergente). Sur la figure 3.4, les couleurs rouge et marron correspondent à une activité majoritaire sur Panoramio ; les couleurs verte et violette mettent en évidence les lieux de consommation ; le bleu indique les lieux de l'activité Twitter majoritaire et le jaune révèle les lieux en marge de la dynamique touristique virtuelle.

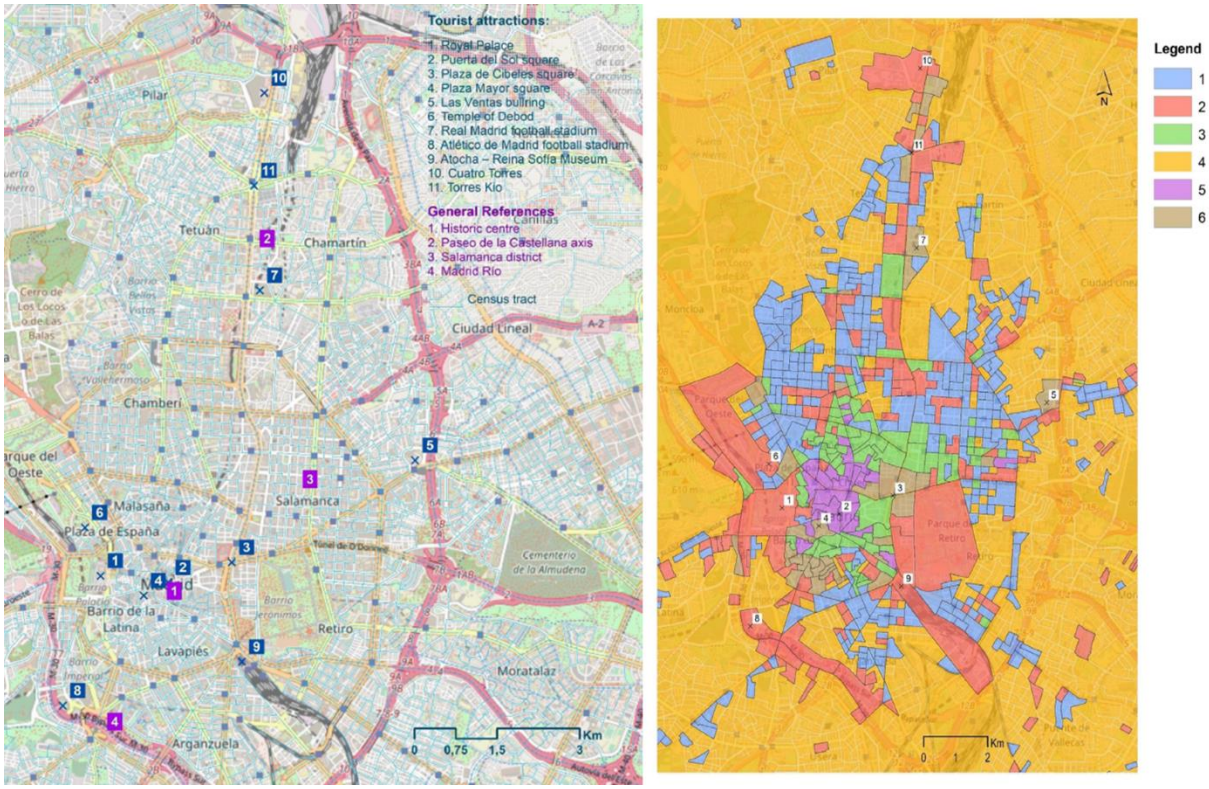


Figure 3.4 : Lieux d'intérêts de Madrid et modalités de l'activité touristique virtuelle à Madrid (Source : Salas Olmedo et al., 2017)

[9] Ces modèles sont construits à partir de la densité de touristes calculée, pour chaque média social considéré, dans les unités administratives de la métropole.

Cette section a donc livré un premier aperçu, fondé sur les publications les plus représentatives, des thèmes et questions de recherche fréquemment abordés dans la littérature focalisée sur les dimensions des tweets, en dehors de la question des risques et catastrophes naturels. Que peut-on en retenir ? On constate un résultat général qui valide le potentiel des tweets comme source de renseignements humains à dimensions spatiale et temporelle, et alternative aux données traditionnelles. Les résultats sont statistiquement encourageants : les coefficients de corrélation linéaire entre variables sont élevés donc les comportements semblent prédictibles (Gomide *et al*., 2011 ; Widener et Wenwen, 2014), les indices d'autocorrélation spatiale témoignent d'une homogénéité des espaces voisins (Salas Olmedo *et al*., 2017) et on peut détecter des structures à l'activité inhabituelle par l'utilisation d'indices (Lucchini *et al*., 2016). Qu'est-ce qui fait défaut ? Certains thèmes d'étude, comme la santé et la nutrition, méritent des analyses plus approfondies sur la question de la représentativité des populations utilisatrices des réseaux sociaux. Même si ce problème est parfois évoqué (Ghosh et Guha, 2013), il reste globalement peu pris en compte dans les résultats d'analyse et les discussions qui s'ensuivent.

3.2. Le tweet géolocalisé dans la problématique des risques et catastrophes naturels

Ainsi que l'ont indiqué (Steiger *et al*., 2015) dans la revue de littérature, la question du potentiel des tweets dans la problématique des risques et catastrophes naturels s'est rapidement imposée comme l'un des thèmes d'étude majeurs. On peut le subdiviser en trois questions : la détection d'événements (*event detection*) et l'alerte précoce (*early warning*) ; la gestion de crise en situation d'urgence ; l'analyse géographique des traces en situation post-crise.

3.2.1. Détection d'événements émergents et dispositifs d'alerte précoce

La communauté scientifique considère comme fait acquis que l'individu-capteur est en capacité, de par la réactivité que nous avons décrite au paragraphe 3.1.1., de rapporter la survenue d'un phénomène perturbateur émergeant bien avant que l'information n'arrive aux autorités, scientifiques et médias traditionnels. L'individu connecté devient en quelque sorte le primo-capteur et primo-vecteur de capture et de diffusion rapides de l'information de terrain, ce que résume l'image ci-après (figure 3.5), en rapport avec le séisme survenu à La Rochelle le 28 avril 2016. Enregistré à 8h46, il est détectable sur Twitter dans la minute suivante alors que la première annonce officielle est transmise à 9h08 via les réseaux de

capteurs sismiques traditionnels[10] (soit 21 minutes de retard par rapport à Twitter). C'est en cela que les signaux émis par les individus connectés et actifs sur les réseaux sociaux permettraient de les appréhender comme des lanceurs d'alerte précoce, par rapport à la longueur des protocoles traditionnels soumis à la contrainte de la hiérarchie (Douvinet *et al.*, 2017).

Figure 3.5 : Illustration de la réactivité des utilisateurs de Twitter face aux phénomènes perturbateurs d'origine physique (Source : Eric Appéré, BRGM – MEEM / https://www.brgm.fr/sites/default/files/2016-11_obs-citoyen2.jpg)

Dès 2010, des chercheurs se sont en effet penchés sur la question de l'usage des réseaux sociaux, notamment pour détecter des phénomènes naturels émergents et potentiellement dangereux via l'agitation du réseau. (Sakaki *et al.*, 2010) avaient ainsi développé l'un des premiers prototypes mêlant détection de phénomènes via Twitter et émissions de messages d'alerte au Japon. Ce prototype s'appuyait sur une architecture recoupant :

- l'analyse sémantique afin de détecter, en temps réel et par un algorithme d'apprentissage, les tweets contenant des mots-clés comme *earthquake* ou *shake* employés dans le contexte des séismes ;
- un second algorithme estimant la localisation de l'épicentre du phénomène à partir de la géolocalisation des tweets collectés et triés.

Cela dit, le système d'alerte précoce devait encore prendre en compte, dans le cas particulier des séismes, la vitesse de diffusion des ondes sismiques à la surface de la Terre, estimée par les auteurs entre 3 et 7 km/s. Une primo-analyse avait permis de détecter

[10] Sources : https://www.brgm.fr/projet/plateforme-suricate-nat-sur-risques-naturels-collecter-informer-prevenir-grace-reseaux (Consulté pour la dernière fois le 19/07/2019)

l'existence d'une réactivité virtuelle temporelle exponentielle, qui s'engageait sur Twitter en moins d'une minute suivant la survenue du phénomène. Le temps moyen de capture et de traitement avant émission des messages d'alerte était ainsi estimé entre 20 secondes et une minute (soit le délai entre le ressenti d'un séisme, capturé dans un tweet, le traitement des tweets par le prototype et l'émission de courriels ou de SMS diffusant l'alerte). En définitive, le prototype se montrait plus rapide que les médias traditionnels pour diffuser une alerte auprès de populations à risques.

Pour contraster avec cette première approche et mesurer l'effet *écho* de Twitter, (Chatfield et Brajawidagda, 2012) avaient analysé la réactivité temporelle des utilisateurs de Twitter en Indonésie, non pas en fonction de la perception d'un phénomène physique en temps réel mais en réponse à la diffusion d'une alerte tsunami par un organisme officiel, via son compte Twitter : la majorité des retweets du message d'alerte officiel par les abonnés au compte ont ainsi eu lieu dans les cinq premières minutes suivant son émission (à la vitesse moyenne de 13 retweets par seconde). Par ailleurs, conséquence directe de l'effet d'une communication virtuelle réticulaire, les auteurs estimaient à quatre millions le nombre potentiel d'utilisateurs informés dans les quinze minutes suivant l'émission du message d'alerte officielle. En outre, c'est dans la première minute suivant la diffusion du message d'alerte que le nombre de personnes informées s'accroissait au maximum (et il s'agit en fait de l'effet du retweet du message d'alerte original par le compte d'une chaîne de télévision disposant de plus de deux millions d'abonnés).

Pour autant, il faut souligner l'existence d'un cetain nombre d'écueils. De tels systèmes fondés sur les réseaux d'individus-capteurs présentent un fonctionnement qui dépend de l'émission de tweets lanceurs d'alerte ; et cette faculté à lancer l'alerte dépend des moyens, de l'accessibilité, des activités et des pratiques numériques des individus. Or, celles-ci sont étroitement liées aux milieux et aux variables socio-démographiques : (Sakaki *et al*., 2010) reconnaissaient ainsi la moindre fiabilité de leur prototype dans les territoires reculés du Japon, où la population témoigne d'une empreinte numérique plus faible que dans les métropoles. Ce constat ne se limite pas au Japon ni aux pratiques numériques du début des années 2010 : (Douvinet *et al*., 2017) indiquaient, en s'appuyant sur des enquêtes menées dans des communes rurales du département du Vaucluse en mars 2016, que leurs habitants étaient moins adeptes des réseaux sociaux et des technologies associées aux smartphones. De la même manière, même si un phénomène de diffusion rapide d'une alerte s'observait chez (Chatfield et Brajawidagda, 2012), on peut s'interroger sur sa portée sur le réseau global des utilisateurs en Indonésie : le message d'alerte d'origine émis par l'organisme chargé de la surveillance et de la détection des séismes et tsunamis en Indonésie n'a été retweeté que dix fois, malgré ses quasi 11 000 abonnés (soit 1 retweet pour 1 100 abonnés) ; la chaîne de télévision ayant retweeté ce message auprès de ses deux millions d'abonnés ne les a finalement guère plus mobilisés (1 retweet pour 1 142 abonnés). Même si la dynamique de

diffusion des messages s'engage rapidement, l'audience d'un compte n'est pas la garantie d'une mobilisation massive de ses abonnés pour le retweet des messages[11].

Quels outils et stratégies existent alors dans le cadre de l'*early warning* (système d'alerte précoce) et quels rapports entretiennent-ils avec les réseaux sociaux et les individus-capteurs qui les alimentent ? (Douvinet *et al*., 2017) indiquent une situation assez disparate en fonction des pays, notamment en ce qui concerne la prise en compte officielle de l'individu-capteur comme lanceur d'alerte officielle, dans une posture verticale *bottom-up*. Dans le cas de la France, même si l'individu connecté peut participer à la création d'informations sur les réseaux sociaux via Twitter et que les services municipaux et/ou de secours disposent généralement d'un compte Twitter, la diffusion de l'alerte reste ancrée dans les protocoles institutionnels, c'est-à-dire dans une approche *top-down* des services de l'Etat vers le territoire local et ses habitants (Douvinet *et al*., 2017). Si l'on s'intéresse aux outils numériques vecteurs d'alerte, on peut rapidement établir deux constats :

- un retrait paradoxal de l'Etat : l'application nomade SAIP (Système d'Alerte et d'Informations aux Populations), lancée par l'Etat français à la suite des attentats terroristes de 2015, a été officiellement fermée par le Ministère de l'Intérieur le 1er juin 2018, en raison de dysfonctionnements répétés et d'un manque d'audience[12] ;
- une avancée des acteurs privés dans le développement d'applications (Douvinet *et al*., 2017) ou dans l'envoi de SMS diffusant des alertes et des recommandations comportementales, à l'image de ce SMS relatif à une période de canicule, émis par une assurance complémentaire santé : "*CANICULE : Hydratez-vous régulièrement, restez au frais, évitez les efforts physique. Pour en savoir plus, rdv sur xxxxxx[13].fr/canicule-conseils*".

Aux Etats-Unis, les stratégies de diffusion *top-down* d'une vigilance (*watch*) ou d'une alerte (*alert*) s'appuyant sur les technologies et outils mobiles prennent diverses formes. En premier lieu, le pays a adopté la technique du *cell broadcasting* (diffusion cellulaire), qui consiste à diffuser massivement auprès des abonnés de l'Internet mobile, des messages d'alerte concernant un territoire précis dans un temps donné. L'utilisateur du smartphone n'a pas besoin d'installer une quelconque application : la réception des messages d'alerte est automatique (à moins que l'utilisateur ait volontairement désactivé la fonctionnalité). Depuis 2012, le *Wireless Emergency Alerts* (*WEA*) émet, dans des territoires ciblés et sans surcoût pour l'abonné, trois types d'alerte dont celles qui représentent une menace immédiate pour la sécurité ou la vie des personnes[14]. Le *National Weather Service* a intégré ce système *WEA*

[11] Et il faut préciser que même si les auteurs estimaient à quatre millions le nombre potentiel d'utilisateurs ayant eu connaissance du message d'alerte dans les quinze minutes suivant son émission officielle, l'île de Sumatra comptait, en 2010, plus de cinquante millions d'habitants. Source : https://fr.wikipedia.org/wiki/Sumatra (Consulté pour la dernière fois le 22/07/2019)

[12] Source : https://www.interieur.gouv.fr/Actualites/Information-de-la-population-en-cas-de-danger/Fin-de-l-application-SAIP (Consulté pour la dernière fois le 22/07/2019)

[13] Dans l'optique de ne pas faire la promotion d'une mutuelle privée, celle-ci est anonymisée.

[14] Source : https://www.fcc.gov/consumers/guides/wireless-emergency-alerts-wea (Consulté pour la dernière fois le 23/07/2019).

via le *Commercial Mobile Alert System* (Ngo et Wijesekera, 2012) ; néanmoins, l'institution n'envoie que des alertes relatives aux phénomènes extrêmes représentant un danger immédiat : tornades, crues éclair, cyclones, tempêtes de sable, vents violents et tsunamis. Par ailleurs, l'organisme inscrit également sa présence sur les réseaux sociaux : les différents établissements locaux du pays disposent de comptes sur Twitter et Facebook, affichant généralement une large audience[15], et utilisés comme outils de dialogue entre l'institution et les utilisateurs de Twitter[16] :

- dans une logique descendante, les établissements locaux diffusent quotidiennement des bulletins météorologiques et assurent leur mission de prévention et d'éducation aux comportements à adopter en cas de survenue d'un phénomène dommageable ; en cas de crise imminente, les réseaux sociaux sont utilisés comme un vecteur supplémentaire de diffusion massive de messages d'alerte ;
- dans une logique ascendante, les utilisateurs peuvent faire remonter leurs observations locales afin d'aider à détecter un phénomène émergent dommageable d'une part, et d'autre part, en cas de crise, de cibler les interventions des secouristes.

En Indonésie, l'intégration des outils numériques ne s'est, de même, pas réduite à une diffusion unilatérale positionnant l'habitant comme un individu passif. Le projet *Peta Jakarta*[17] est directement né du besoin, pour les autorités municipales de la mégalopole indonésienne de plus de 30 millions d'habitants, fréquemment soumise aux inondations, de disposer d'informations acquises en temps réel sur le terrain, afin de détecter et de cibler les lieux inondés puis d'organiser l'aide aux populations[18]. *Peta Jakarta* est une plateforme de collecte de l'information créée par les individus-capteurs qui prend la forme d'un double dialogue *bottom-up* / *top-down* entre les habitants connectés et les autorités municipales. Les émissions de tweets occupent une place centrale dans le fonctionnement de la plateforme : la proportion d'utilisateurs de Twitter est plus importante que dans les autres capitales d'Etats et l'Indonésie est le cinquième pays au monde en termes d'effectifs d'utilisateurs actifs. Les utilisateurs reçoivent des consignes précises afin de créer une information directement exploitable pour les autorités (utiliser des *geohashtags* afin de spatialiser l'information en fonction des lieux observés, estimer les hauteurs d'eau). A chaque inondation, ces traces sont cartographiées et recoupées avec des jeux de données topographiques, météorologiques, démographiques et socio-économiques, afin de vérifier la validité des informations issues du public participant et d'étudier l'impact des phénomènes dans différents espaces et sur différentes catégories de population (Meier, 2015). Une cartographie interactive, accessible

[15] Le compte Twitter de Dallas *(@NWSFortWorth*) dispose de 110 000 abonnés ; celui de Norman (*@NWSNorman*), 98 400 abonnés et celui de Houston (*@NWSHouston*), de 62 000 abonnés.
[16] Source : https://www.weather.gov/wrn/summer-article-how-the-NWS-leverages-social-media (Consulté pour la dernière fois le 23/07/2019).
[17] Renommé récemment en Peta Bencana : https://petabencana.id/ , et disponible pour trois autres villes de l'île de Java.
[18] Source : https://perspectives.eiu.com/infrastructure-cities/surging-cities/case-study/case-study/mapping-flood-new-data (Consulté pour la dernière fois le 23/07/2019)

via le nouveau site *Peta Bencana* assure ainsi le suivi en temps réel des phénomènes détectés et de leur intensité, tout en permettant aux internautes d'alimenter l'information relative à ces phénomènes par leurs contenus générés sur les réseaux sociaux. La figure 3.6 montre les différents niveaux d'alerte en cours dans divers points de la ville de Jakarta, en fonction des hauteurs d'eau. Un utilisateur présent dans ces lieux peut contribuer à la collecte d'information pour renseigner le phénomène en cours via les réseaux sociaux, en suivant la procédure indiquée par le guide.

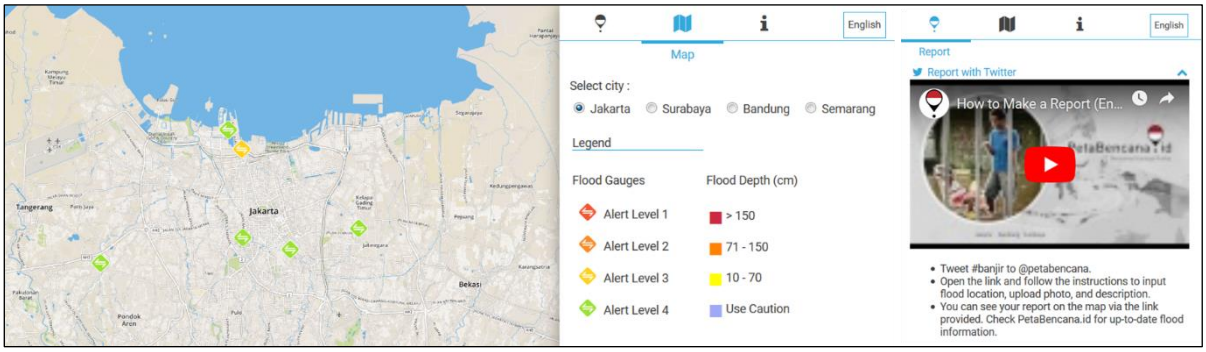


Figure 3.6 : Peta Bencana le 22/07/2019 : carte interactive des niveaux d'alerte inondation en cours et procédure à suivre pour documenter un phénomène en passant par Twitter

Plus récemment et dans la problématique des séismes, le système *Lastquake* (Bossu *et al.*, 2018) s'inscrit dans la continuité de la logique de *Peta Bencana* : il s'appuie sur un ensemble de composants nomades et de bureau (des sites Web, une application mobile, un système d'envoi de messages instantanés comme Twitter et Telegram) fonctionnant dans la double logique verticale ascendante et descendante. Ainsi, lorsqu'un utilisateur de l'application mobile ressent une secousse sismique, il peut la rapporter en créant un témoignage agrémenté d'une photo géolocalisée. Tout témoignage est directement publié sur le site Web du système ; l'utilisateur de l'application mobile n'est notifié que lorsque que les témoignages de terrain sont vérifiés et que la localisation du foyer sismique est déterminée. En outre, à chaque notification *top-down*, l'utilisateur de l'application est invité à faire remonter un témoignage de terrain (via un questionnaire en ligne ou la description de dégâts sous la forme d'images à sélectionner en fonction des degrés de dommages observés).

Les nouveaux outils et nouvelles approches numériques vis-à-vis des risques naturels intègrent les individus connectés mais à des degrés variés, de la simple réception de l'information dans une logique de communication *top-down* à l'*échange* vertical d'information de terrain entre les témoins directs et les gestionnaires distants, dans une posture contributive et volontaire[19]. Mais l'inconvénient majeur de ces systèmes de veille et d'alerte demeure identique : ils risquent d'introduire des clivages sociaux et territoriaux. D'un côté, les marges

[19] L'utilisateur alimentant le contenu de *Peta Bencana* ou encore de *Lastquake* a la conscience et l'intention de fournir une information pertinente et exploitable dans une démarche de construction collective.

numériques bénéficiant d'une couverture Internet mobile faible (voire inexistante) ne peuvent pas s'intégrer à cette dynamique (Douvinet *et al*., 2017) ; de l'autre côté, on ne peut pas contraindre l'ensemble des habitants de territoires à risques à disposer d'un smartphone connecté, à télécharger et/ou s'incrire à des applications lanceuses d'alerte en cas de danger[20] (Sakaki *et al*., 2010 ; Douvinet *et al*., 2017).

3.2.2. Stratégies d'utilisation du tweet pendant la gestion de crise

La gestion de crise en temps réel correspond en fait à l'une des premières manifestations de l'intérêt porté aux réseaux sociaux lorsque survient un phénomène physique perturbateur du quotidien et générateur de danger, notamment lorsque les réseaux de télécommunication traditionnels sont saturés ou en panne (McDougall, 2012 ; Hecker, 2014). La gestion de crise en temps réel consiste alors à collecter et à vérifier les témoignages transmettant des observations ou informations de terrain (dégâts, victimes, etc.) émises par les internautes via diverses sources (tweets, SMS, courriels), puis à les afficher sous forme d'une cartographie interactive publique de crise, nommée *crisis map* (McDougall, 2012 ; Meier, 2015). Dans cette phase de gestion de crise, le traitement des données est *opérationnel* et s'applique comme indiqué dans la figure 3.7 :

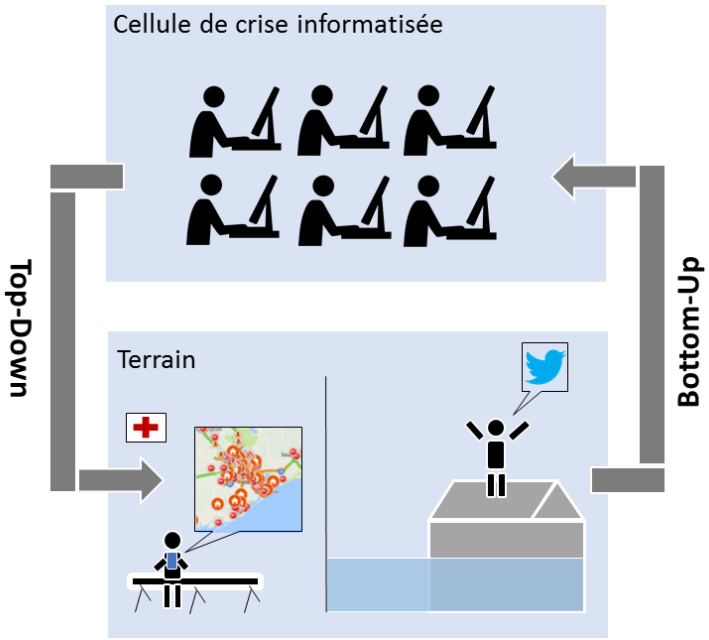


Figure 3.7 : Schéma illustrant les principes de la cartographie collaborative de crise

[20] La diffusion cellulaire présente les mêmes inconvénients : la fonctionnalité n'est pas disponible sur les téléphones portables autres que smartphones, de même que certains opérateurs et pays ne l'ont pas adoptée.

Dans un sens vertical ascendant, les individus témoins des conditions environnementales du terrain affecté par un phénomène diffusent une information localisée (observations d'aléas, appels au secours, etc.) ; dans la cellule de crise, les informations capturées par les témoins sont collectées, analysées et recoupées avec des données physiques et sociales afin de vérifier la validité des témoignages reçus et de cibler les lieux dans lesquels les secours doivent être prioritaires. Dans un sens vertical descendant, les gestionnaires de la cellule de crise transmettent et mettent à jour, via la cartographie collaborative de crise, les informations *vitales* auprès du public (McDougall, 2012 ; Meier, 2015) : lieux de refuges, lieux où trouver du ravitaillement ou des soins, dégâts et routes fermées, lieux à éviter, etc.

L'origine de cette approche remonte à l'année 2008 pendant une période d'émeutes post-électorales ayant affecté le Kenya. Ces troubles alors peu médiatisés, une activiste locale a l'idée de recueillir les témoignages *directs* des habitants – par les SMS, tweets, flux RSS – et de les cartographier afin de permettre à la population de visualiser les espaces de conflits violents à éviter (Hecker, 2014). A la suite de cet événement social est née l'organisation Ushahidi qui a développé le site et le logiciel de cartographie de crise éponymes, rapidement adoptés à l'échelle globale[21]. La figure 3.8 en soumet quelques exemples à travers deux crises d'origine naturelle : les séismes de Haïti en 2010 et du Népal en 2015.

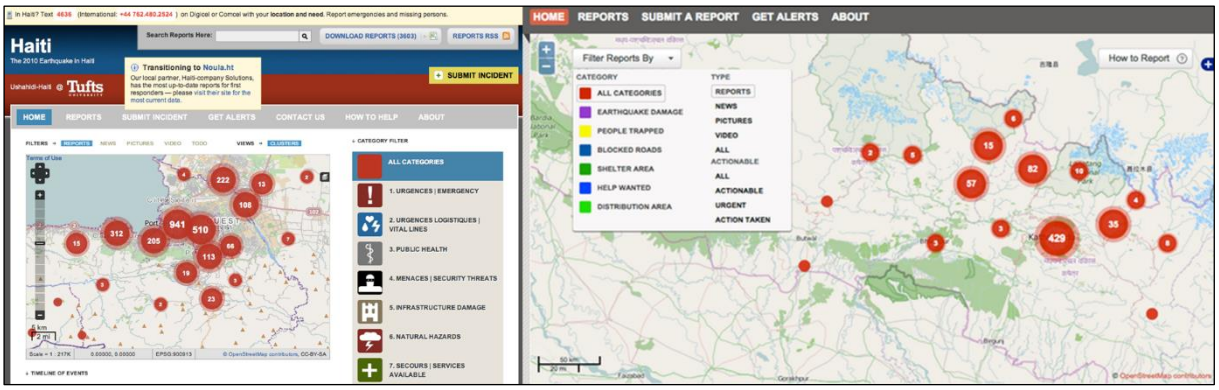


Figure 3.8 : Cartographies collaboratives de crise des séismes ayant frappé Haïti en décembre 2010 (à gauche) et le Népal en avril 2015 (à droite), (Source : Ushahidi.com)

Les clusters rouges représentent le nombre de témoignages transmis et validés sur un espace (ces témoignages pouvant être des tweets, des messages Facebook, des SMS ou des courriels). L'internaute qui consulte la carte peut filtrer l'affichage des données en fonction de deux critères :

[21] En 2015, on estimait à 60 000 le nombre de cartographies collaboratives de crise créées à partir de Ushahidi. Source : http://www.lemonde.fr/afrique/article/2015/11/10/ushahidi-une-technologie-africaine-qui-a-conquis-la-planete_4806913_3212.html

- *Category* permet de visualiser les témoignages en fonction de leur thème : refuges, demandes d'aides, lieux de ravitaillement, individus bloqués, etc. ;

- *Type* permet de filtrer les témoignages en fonction de leur nature (texte, image, vidéo, etc.)

L'usage de la cartographie collaborative de crise et des réseaux sociaux s'est développé aussi bien dans les pays du Nord que du Sud. L'une des premières applications test en situation de catastrophe naturelle correspond au séisme de janvier 2010 à Haïti : messages transmis par SMS ou émis via les plateformes sociales Twitter et Facebook arrivent en appui à l'organisation des équipes de secours qui n'ont pas la connaissance préalable du terrain et qui se retrouvent confrontées à des difficultés de localisation des victimes (De Blomac, 2012 ; Hecker, 2014). Cette catastrophe naturelle marque également une nouvelle approche en termes de gestion dans la mesure où l'ensemble des informations de terrain émises par les témoins capteurs directs n'a pas été traitée sur place mais par des équipes d'étudiants du Massachusets, épaulées par des Haïtiens expatriés ; entre 40 000 et 60 000 messages ont ainsi été centralisés, traduits, géocodés et restitués sur une carte destinée aux équipes de secours et ONG sur place (De Blomac, 2012). Cette catastrophe naturelle marque la mise en place d'équipes bénévoles localisées à distance des lieux affectés mais mobilisables (les *digital humanitarians*[22] de Patrick Meier) qui soutiennent les équipes de secours, primo-intervenants et gestionnaires de crise, notamment par leur présence et actions numériques. Le *VISOV* (Volontaires Internationaux en Soutien Opérationnel Virtuel) en constitue un exemple : cette communauté de bénévoles assure une veille des réseaux sociaux sur lesquels sont émises, dans les premières minutes après la survenue d'un phénomène, les observations et photographies dont les services de secours ont besoin afin de localiser le phénomène et d'estimer sa gravité. Cette veille se manifeste notamment par la fouille sémantique des contenus émis sur les réseaux sociaux[23].

En 2012, lors du passage de l'ouragan Sandy sur la côte Est des Etats-Unis, Twitter s'est imposé comme le porte-parole des autorités locales ou fédérales dans la diffusion de messages d'alertes et de consignes auprès des populations locales. Mises en cause dans leur gestion de l'ouragan Katrina en 2005, les autorités ont développé une véritable stratégie de communication ancrée sur les réseau sociaux et plus particulièrement sur Facebook et Twitter (Hecker, 2014) qui permettent une communication multilatérale entre autorités et habitants sans être confrontés à l'effet de saturation des numéros d'urgence traditionnels (Chatfield *et al.*, 2014). Plus de 130 000 tweets ont ainsi été échangés, quelle que soit la phase du phénomène (préparation, gestion de crise, résilience), entre les autorités et les habitants.

[22] L'expression est employée pour désigner les volontaires ou professionnels qui se mobilisent rapidement en cas de catastrophe naturelle afin de traiter les informations émises par SMS, sur les réseaux sociaux ou encore les images satellites afin d'appuyer l'action des équipes de secours présentes sur place. Source : http://www.digital-humanitarians.com/ (Consulté le 24/07/2019)

[23] Source : https://www.francebleu.fr/infos/societe/le-visov-des-pompiers-volontaires-du-web-pour-aider-les-secours-en-vaucluse-1499937149 (Consulté pour la dernière fois le 24/07/2019)

Cette pratique reste inscrite dans la gestion des secours pendant les catastrophes naturelles : pendant le passage de l'ouragan Harvey sur Houston en août 2017, de nombreux résidents piégés par les inondations ont lancé un appel à l'aide, non pas par le numéro téléphonique de secours général 911, mais par des tweets contenant leur adresse ainsi que des hashtags caractéristiques : *#HoustonSOS* ou *#HoustonRescue*[24]. Mais aux Etats-Unis, force est de constater que ce sont les entreprises privées qui s'inscrivent dans le développement des cartographies de crise : le 29 octobre 2012, pendant l'ouragan Sandy, Google avait lancé la *Superstorm Sandy crisis map.* L'initiative est répétée pendant le passage de l'ouragan Harvey, avec la *Hurricane Harvey crisis map* qui permet de localiser les routes fermées, les accidents ainsi que les refuges ouverts et leur capacité à accueillir des sinistrés.

Malgré ces initiatives et le fait indéniable qu'elles permettent de sauver des vies et assurent un gain de temps considérable pour les équipes de secours, on n'a que peu de retours sur l'efficacité réelle et globale de ces pratiques[25] (De Blomac, 2012). Dans des tweets publiés pendant le passage de l'ouragan Harvey sur les côtes texanes, le chercheur en géomatique Anthony Robinson manifestait son opposition au recours systématique des tweets comme succédanés du numéro d'appel 911 pour diriger les secours aux populations. Son argument était le suivant : parmi les individus les plus vulnérables, on trouve des personnes invisibles sur les réseaux numériques car non connectées ; la gestion de crise doit donc également se focaliser sur les territoires et populations qui ne sont pas visibles sur le réseau (figure 3.9) car non consommatrices de technologies numériques.

Figure 3.9 : Tweets émis le 27 août 2017 par Anthony Robinson, en rapport à la gestion des secours par les réseaux sociaux

Et en effet, on peut envisager un certain nombre de risques et de biais dont certains ont déjà été présentés dans l'introduction de ce manuscrit, ainsi que dans le chapitre 2 : par les contraintes sociales et spatiales d'accessibilité aux réseaux de l'Internet, fixe ou mobile, les

[24] Source : https://www.foxnews.com/tech/tropical-storm-harvey-is-twitter-becoming-the-new-911 (Consulté pour la dernière fois le 24/07/2019)
[25] Et notamment lors de la catastrophe de Fukushima en 2011.

dispositifs numériques risquent de hiérarchiser territoires et populations en fonction de leur visibilité virtuelle ; c'est justement ce processus tendancieux de sélection qui risque de s'introduire dès lors qu'on fait appel aux traces numériques pour résoudre un problème intégrant une variabilité socio-spatiale. Par ailleurs, certains de ces biais sont déjà connus et rapportés : l'activité sur les réseaux sociaux ne va pas s'inscrire comme priorité pour les personnes, certes connectées, mais dans une situation d'urgence immédiate. (Bossu *et al*., 2018) décrivaient par exemple l'*effet doughnut* perceptible lors des violents séismes (dont la magnitude est supérieure à 5). Cet effet se manifeste ainsi : la collecte de messages depuis les zones endommagées proches de l'épicentre est rare et retardée de dix à vingt minutes alors que les territoires éloignés sont immédiatement actifs. Dans les premières minutes suivant un séisme violent, on observe ainsi un rayon proche de l'épicentre mais vide d'activité virtuelle alors que des territoires situés dans un rayon plus éloigné de l'épicentre sont actifs virtuellement. De la même manière, les travaux de (Chatfield et Brajawidagda, 2012) relatifs à la vitesse de diffusion des alertes tsunamis en Indonésie montraient finalement, en s'intéressant à la localisation des retweets, que seul 1% des messages était localisé à Banda Aceh (au nord de Sumatra, c'est la ville qui a subi le plus de dégâts en 2004 et qui restait la plus menacée en 2012). Les villes de la côte ouest de Sumatra, directement exposées au risque de tsunami, ne comptaient que 7% des retweets totaux. Au contraire, la capitale Jakarta, sur l'île de Java, rassemblait 46% des retweets émis alors qu'elle se trouve relativement abritée face au risque de tsunami. Il existait donc un décalage spatial et quantitatif entre populations et lieux menacés et populations et lieux actifs sur les réseaux numériques.

3.2.3. Exploration du potentiel des tweets géolocalisés en situation post-crise

Si la plupart des papiers publiés sont focalisés sur les approches développées ci-avant (Steiger *et al*., 2015), d'autres s'appuient sur l'analyse spatio-temporelle des tweets géolocalisés émis en réponse à un phénomène naturel, dans une situation *post-crise* (Dashti *et al.*, 2014), autrement dit après la fin du phénomène et des événements engendrés. Cette approche est fondée sur l'analyse spatio-temporelle des tweets collectés et les objectifs sous-jacents annoncés visent à évaluer le potentiel des tweets vis-à-vis de la connaissance des dynamiques, individuelles ou collectives, engagées en réponse aux phénomènes naturels générateurs de dangers. Ce potentiel est généralement présenté comme suit :

- les tweets peuvent être appréhendés comme un soutien complémentaire afin d'évaluer les dégâts en fonction de différents lieux affectés ; en conséquence, l'analyse des tweets peut avoir pour finalité une révision de la planification existante, et notamment en ce qui concerne la vulnérabilité des lieux et des populations (De Longueville *et al*., 2009 ; Brovelli *et al*., 2014).

- De par leurs composantes triples (spatiale, temporelle et sémantique) et leur caractère spontané, les tweets offrent une perspective inédite pour évaluer les dispositifs d'alerte et de résilience (Blanford *et al*., 2014) ;

- Les tweets offrent également une porte d'entrée dans l'étude de la *situational awareness*, c'est à-dire la perception des éléments de l'environnement dans un temps et dans l'espace directs, la compréhension de leur sens et l'anticipation de leur évolution dans un futur proche (Hertfort *et al*., 2014).

- En conséquence, les tweets sont indispensables pour améliorer cette *situational awareness*, et, en envisageant une utilisation optimale des médias sociaux par l'ensemble des acteurs impliqués et des individus affectés, la résilience des populations (Dashti *et al.*, 2014).

Ainsi, l'un des premiers travaux focalisés uniquement sur les dynamiques spatio-temporelles d'émission de tweets en situation post-phénomène s'est attaché à l'exploration de l'événement virtuel consécutif à un feu de forêt, survenu en juillet 2009 aux alentours de Marseille (De Longueville *et al*., 2009). Les travaux mettaient en évidence l'existence, comme dans le cas des événements perturbateurs d'origine sociale, d'une dynamique temporelle se manifestant par des pics et des creux d'émission ; le pic le plus important était alors enregistré au moment où le *ressenti* du phénomène par la population était le plus intense (hauteur des flammes, fumées et nuages de cendres). En revanche, l'émission du premier tweet rapportant le phénomène restait tardive, soit 1h30 après le début de l'incendie (tendance qui s'avère désormais inversée) et associée à un journal local. Par rapport à la date de l'étude (2009), trop peu d'utilisateurs avaient recours aux fonctionnalités spatiales de Twitter, remettant en question la pertinence de la plateforme comme source d'information géolocalisée. C'est en fait la composante sémantique des tweets qui contenait davantage d'indices quant à la perception des lieux par les utilisateurs : même si le seul nom de *Marseille* restait majoritairement mentionné par les tweets, on en identifiait une poignée évoquant le lieu de départ de l'incendie, les quartiers les plus exposés, les étangs d'approvisionnement des canadairs, etc.

Dans un deuxième temps, les chercheurs se sont focalisés sur la question de la possibilité de distinguer l'existence de différentes phases de réponse au phénomène réel dans l'événement virtuel d'une part, et d'évaluer la correspondance des dynamiques spatio-temporelles entre phénomène physique et événement virtuel d'autre part : ce nouveau volet d'exploration s'est ainsi fondé sur l'intégration de l'analyse sémantique à l'analyse spatio-temporelle. A travers trois phénomènes[26], (Roy Chowdhury *et al*., 2013) mettaient en évidence l'existence de l'emploi de différents temps (présent, passé) et d'un vocabulaire typique permettant de démarquer les différentes phases de l'événement virtuel : la phase pré-événement (*warning*, *alert*), la phase de l'événement en cours (*now*, *sweeps* dans le cas d'un typhon) et la phase post-événement (*aftermath*, *donate*). (Blanford *et al*., 2014) dressaient le

[26] Le séisme ayant frappé Haïti en janvier 2010, la tornade ayant dévasté la ville de Joplin dans le Missouri en mai 2011 ainsi que le typhon Nesat ayant affecté les Philippines en septembre 2011.

même constat d'après l'exploration de l'événement virtuel provoqué par le passage d'une tornade dans la ville de Moore, en Oklahoma : avant la tornade, les tweets contenaient des mots-clés descriptifs d'une menace (*storm*, *tornado*, *watches*, *warnings*), pendant le passage de la tornade, les tweets témoignaient davantage de l'engagement d'une dynamique de réponse au phénomène et de la mise en place de mesures d'urgence (*sirens*, *shelters*) ; enfin, lorsque le phénomène météorologique s'était dissipé, l'événement virtuel amorçait les constatations des dégâts. Par ailleurs, il s'avérait que le passage d'une phase à l'autre était rapide et directement lié aux événements de l'environnement perçu par les utilisateurs : par exemple, les mots *sirens* et *be safe* apparaissaient dans la minute même où les sirènes de Moore commençaient à retentir.

D'autres auteurs (Dashti *et al*., 2014 ; Hertfort *et al*., 2014 ; Saravanou *et al*., 2015) se sont intéressés à la question de la pertinence de la composante spatiale des tweets : les tweets géolocalisés sont-ils proches ou distants des lieux réellement affectés par le phénomène physique ? Les travaux publiés par (Dashti *et al*., 2014) ont testé la validité des informations relayées dans les tweets en proposant une représentation cartographique croisant, d'une part, sources officielles permettant d'identifier les lieux affectés (imagerie aérienne témoignant de l'ampleur de l'inondation de la ville de Boulder dans le Colorado en 2013), et d'autre part, l'information contenue dans un jeu de tweets géolocalisés (texte, photos). Il s'est avéré que la localisation et le contenu des tweets témoignant d'inondations et de dégâts correspondaient avec deux sources de données officielles : l'imagerie satellite et la cartographie de l'aléa inondation de la ville de Boulder. C'est ce que montre la figure 3.10 : les tweets dont le contenu photographique est affiché sur la carte et témoigne de dommages, sont bien localisés dans les zones à risque d'inondation élevé (en rouge sur la carte) ainsi que dans le lit d'inondation de fréquence quincentenale (en jaune sur la carte).

Figure 3.10 : Exemple de tweets géolocalisés pertinents superposés à la cartographie de l'aléa inondation de Boulder au Colorado (Source : Dashti et al., 2014)

(Hertfort *et al*., 2014) annonçaient également une correspondance spatiale entre la distance des tweets géolocalisés aux bassins versants inondés (de l'Elbe en Allemagne) et le rapport de ces tweets au phénomène ; en outre, après avoir classifié les tweets en fonction de leur contenu sémantique, il apparaissait que les catégories de messages mentionnant des hauteurs d'eau et des actions d'entraide étaient les plus *proches* des bassins versants affectés (distance moyenne des tweets géolocalisés au centre de gravité des bassins versants affectés de 39 km). (Saravanou *et al*., 2015) introduisaient une nouvelle piste : l'estimation de l'intensité des phénomènes (comment identifier les territoires durement frappés). Leur conclusion indiquait que le nombre d'utilisateurs actifs n'était pas un indicateur pertinent et qu'il fallait mieux privilégier le nombre de tweets émis dans un territoire voire même la création d'un indice ayant l'avantage d'éviter les biais liés aux densités variables de la population.

Qu'en est-il de la connaissance des réponses aux phénomènes extrêmes d'origine hydrométéorologique ? Après les phénomènes extrêmes dévastateurs survenus dans l'Atlantique ouest pendant la saison cyclonique 2017, de nouvelles publications s'attellent à d'autres pistes de recherche. (Samuels *et al*., 2018) adoptent toujours une approche quantitative, mais qui prend en compte l'une des problématiques pointées par Anthony Robinson dans ses tweets, à savoir la question de la prise en compte des lieux invisibles. Ces auteurs analysent ainsi, non pas les lieux en fonction de l'émergence de l'activité tweeting en réponse à la crise, mais de la disparition de l'activité tweeting en fonction du temps. Leur objectif consiste en effet à tester l'hypothèse suivante : la décroissance ou la disparition de l'activité tweeting peut témoigner d'une impossibilité matérielle à tweeter en raison d'une situation urgente locale. Les paramètres statistiques testés dans l'étude, (écart-type et écart moyen calculés pour chaque jour dans des mailles fines) indiquaient en effet l'existence d'une corrélation positive entre déclin de l'activité tweeting quotidienne et nombre de propriétés endommagées. D'autres recherches se sont appuyées sur l'analyse qualitative, à partir de la sémantique des tweets : (Nguyen *et al*., 2019) ont développé un système destiné à diagnostiquer, à partir de la sémantique des tweets, les besoins des individus confrontés à une situation de crise, ainsi que l'évolution de leurs besoins dans le temps. En comparant les tweets classés en différentes catégories exprimant des besoins (*food*, *help*, *rescue*, etc.) identifiés dans trois phénomènes (ouragans Harvey, Irma et Sandy), ils parvenaient à souligner la particularité de l'ouragan Harvey en termes d'effets sociaux : les besoins majoritaires exprimés dans les tweets évoquaient alors le ravitaillement de première nécessité (*food*, *power*, *gas*) ainsi que les secours et l'entraide (*rescue*, *help*, *victim*, *volunteer*, *donations*, *relief*, *affect*). Pendant l'ouragan Irma, même si les besoins en ravitaillement restaient tangibles, les autres thèmes majoritaires concernaient en premier lieu les recommandations comportementales (*advisory*). En outre, le système révélait une transition claire, pour l'ouragan Harvey, entre les besoins exprimés dans les tweets : avant le 27 août 2017, ceux-ci concernaient la nourriture et le carburant ; le vocabulaire témoignant d'un besoin d'aide augmentait rapidement à partir du 27 août pour culminer à son paroxysme le 28 août 2017.

La décennie passée de travaux scientifiques argumente ainsi en faveur d'un potentiel certain des tweets géolocalisés en tant qu'indicateur spatial, temporel et sémantique. A l'échelle de l'individu, le tweet capture le comportement humain résultant de l'interaction entre l'utilisateur auteur du tweet et de son environnement local ; d'un point de vue collectif, il renseigne les rythmes des activités et des lieux pratiqués par différents profils d'utilisateurs. Lorsque survient un événement ou phénomène perturbateur, il a la capacité de le signaler immédiatement, de même qu'il renseigne avec cohérence les dynamiques des phénomènes et événements d'un territoire en crise. En dépit de ces conclusions positives, la question de la représentation virtuelle de l'ensemble des lieux et populations d'un territoire, reste en suspens et peu abordée dans la plupart des publications consultées[27]. La section suivante, et dernière de ce chapitre, présente alors les méthodologies d'analyse fréquemment employées qui fournissent les résultats des travaux présentés dans les deux premières sections.

3.3. Quelles méthodologies et outils d'analyse des tweets dans la recherche pluridisciplinaire ?

Cette dernière partie introduit les méthodologies d'analyse et les outils statistiques, cartographiques ou d'analyse spatiale identifiés dans la bibliographie. L'exploitation des tweets géolocalisés se trouve en effet confrontée à un certain nombre de verrous méthodologiques : en premier lieu, les questions de recherche précises, dont la problématique des risques et catastrophes naturels, nécessitent une identification préalable des tweets en rapport au problème d'étude. D'un point de vue cartographique, il s'agit de transformer l'amas de points que représentent les tweets géolocalisés émis par smartphone en donnée géographique valorisable et combinable à d'autres sources de données. Enfin, d'un point de vue exploratoire, le dernier verrou majeur qui persiste reste la question de l'adéquation entre ces traces aux propriétés particulières et les outils d'analyses statistiques et spatiales qui restent identiques à ceux utilisés traditionnellement pour l'analyse des données institutionnelles standardisées. Face à ce problème, il sera alors question d'introduire les propositions mises en lumière, à partir de 2015, par des chercheurs géographes/géomaticiens concernant un questionnement sur la posture épistémologique vis-à-vis de ces nouvelles sources d'information numérique.

[27] Comme signalé dans le chapitre 2, les travaux axés sur la question des profils d'utilisateurs activant les fonctionnalités de géolocalisation mettent en évidence des caractéristiques socio-démographiques ainsi que des motivations variées, en fonction des territoires d'étude. Dans la question des tweets géolocalisés émis en réponse aux phénomènes naturels dommageable, la prise en compte de ces biais reste secondaire. A notre connaissance, ce comportement a commencé à évoluer après la saison cyclonique de 2017 : par exemple, (Zou *et al.* 2018) ont publié un travail focalisé sur la caractérisation des disparités sociales liées à l'activité tweeting générée en réponse à l'ouragan Harvey (le travail et les réflexions de ces auteurs seront présentés dans le chapitre suivant).

3.3.1. Les méthodes d'analyse sémantique

Comme indiqué ci-dessus, un certain nombre de travaux thématiques se focalisent sur une partie ciblée de toute l'information capturée par le réseau. Pour ces travaux, la première étape indispensable constitue donc à identifier et à isoler les contenus rattachés au thème d'étude. Les méthodologies d'analyse sémantique se distinguent ainsi en deux phases : la première phase correspond à l'extraction d'un jeu de *tweets utiles* à l'étude de la problématique de recherche. Par l'expression *tweets utiles*, on désigne ici les tweets rattachés à une thématique particulière étudiée, et la méthode la plus communément employée, dans la littérature, pour extraire ces tweets, consiste à établir une liste de mots-clés (mots simples ou hashtags) relatifs à la thématique d'étude et à filtrer les tweets selon qu'ils contiennent au moins l'un des mots-clés inventoriés (Steiger *et al*., 2015). Dans leurs travaux relatifs aux effets de l'environnement direct d'un utilisateur sur ses choix alimentaires, (Chen et Yang, 2014) avaient proposé un lexique de mots-clés correspondant aux noms des enseignes de restauration rapide d'un côté, et des noms des enseignes vendant des produits non transformés de l'autre, implantées sur le terrain d'étude (Colombus, capitale de l'Ohio). Le lexique final contenait plus de seize noms d'enseigne. (Widener et Wenwen, 2014) soumettaient un glossaire composé de 158 mots-clés classés en fonction de noms d'enseigne, de noms de produits alimentaires non transformés et de noms de produits emblématiques de la malbouffe.

Si l'on retourne sur l'exemple des risques naturels, le contenu du lexique d'extraction de tweets utiles n'apparaît, dans la plupart des cas, pas autant volumineux : dans leurs travaux sur l'ouragan Sandy, (Chatfield *et al*., 2014) ne collectaient que les tweets contenant le hashtag *#sandy* ; de même, (Alam *et al*., 2018) ne définissaient que trois à quatre mots-clés d'extraction pour les ouragans de la saison cyclonique de 2017 à l'étude : *Hurricane Harvey*, *HurricaneHarvey*, *Harvey*, *HurricaneIrma*, *Irma Storm*, *Irma*, *Hurricane Maria*, *Hurricane Maria*, *HurricaneMaria*, *Maria Storm* et *Tropical Storm Maria*. Pour l'étude des inondations survenues dans le Colorado, (Dashti *et al*., 2014) mêlaient simples mots-clés et hashtags, introduisaient les noms de lieux à différentes échelles et les noms d'acteurs de la crise. En outre, ces mots-clés évoluaient en fonction du temps de la crise ; c'est ce que montre le tableau 3.1. Les 11 et 12 septembre 2013, on identifie des mots-clés liés aux phénomènes d'inondations en cours : ces mots-clés font référence à des entités géographiques (*boulderflood* et *waldoflood* pour les villes de Boulder et de Waldo, *jeffcoflood* pour le comté de Jefferson) ou à des acteurs (*nwsboulder* pour le centre local du *National Weather Service* de la ville de Boulder). Les 19 et 20 septembre 2013, les mots-clés inventoriés sont davantage ancrés dans les réponses sociales destinées à enrayer la crise (*gas*, *infrastructure*, *#cofloodrelief*) ou apporter un soutien moral *(#coloradostrong*).

Tableau 3.1 : Lexique de mots-clés utilisés pour l'extraction de tweets utiles (Source : Dashti et al., 2014)

Date	Keyword Set Terms
09/11/2013	boulderflood, cowx, nwsboulder
09/12/2013	coflood, cofloods, coflooding, cuboulder flood, #boulder, #cccf, jeffcoflood, waldoflood
09/15/2013	Boulderfloods
09/19/2013	flood gas, flood infrastructure, #boco_trails, cdot, #cofloodrelief
09/20/2013	#coloradostrong

Tout jeu de tweets bruts, collectés sur un territoire donné en un temps donné, peut ainsi êtres filtré sémantiquement à partir de le sélection des tweets contenant au moins l'un des mots-clés ciblés par le chercheur[28]. Pour autant, le fait qu'un tweet contienne l'un de ces mots-clés ne garantit pas pour autant son rattachement certain à la problématique étudiée : tout jeu de tweets ayant subi un filtrage sémantique peut encore contenir du *bruit résiduel*. Par l'expression bruit résiduel, on entend les tweets filtrés dans le jeu de tweets utiles car contenant l'un des mots-clés recherchés mais qui ne sont pas pertinents pour la résolution de la problématique d'étude :

- un tweet marqué comme bruit résiduel peut contenir un mot-clé recherché, employé dans un tout autre contexte. (Sakaki *et al*., 2010) avaient d'ores-et-déjà identifié ce problème, en citant ce tweet – "*Someone is shaking hands with my boss"* - qui contient l'un des mots-clés recherchés pour la détection précoce des séismes (*shaking*) mais dans un contexte sans rapport avec la survenue d'un phénomène sismique. Sur notre terrain, nous avons pu identifier de tels tweets via les multiples annonces d'emploi : par exemple, ce tweet "*Can you recommand anyone for this #job ? Driver CDL – Trench Safety - #Houston TX, #Transportation #Veterans*" est extrait par la recherche du mot-clé *safety* mais ne concerne pas la problématique des risques naturels[29].

- un tweet peut également être considéré comme bruit résiduel si, bien qu'il contienne l'un des mots-clés recherchés dans le contexte précis de l'étude, son sens reste ambivalent. (Sakaki *et al*., 2010) avaient également noté ce type de tweet, dont le discours est effectivement articulé autour des séismes mais qui ne s'accorde pas à la problématique de détection de phénomènes en temps réel, puisqu'il évoque un phénomène passé : "*The earthquake yesterday was scaring*". De la même manière, (Chen et Yang, 2014) excluaient l'ensemble des tweets filtrés par les mots-clés qui n'étaient pas en lien direct avec une activité présente et accomplie par l'utilisateur auteur du tweet. Dans notre cas, on peut également

[28] L'étape de filtrage est simplement exécutée par une requête SQL depuis la base de données stockant les tweets.

[29] Pour écarter ce type de tweets, il suffit alors d'exécuter une requête excluant les tweets contenant les caractères *#job* dans leur sémantique.

identifier quelques tweets commémorant une crise passée et/ou comparant un phénomène en cours avec un phénomène passé[30] : "*Today marks one year since the Memorial Day*[31] *weekend flood. Here's drone footage comparing last year to this year*", "@[utilisateur A] @[utilisateur B] @[utilisateur C] *the lightning are far back from where they were last year*".

Comment détecter et supprimer ce bruit résiduel ? Devant la complexité et l'hétérogénéité des discours des tweets, la relecture des tweets filtrés est fréquemment manuelle (Hertfort *et al*., 2014 ; Saravanou *et al*., 2015; Roberts, 2017), voire inexistante (Widener et Wenwen, 2014 ; Dashti *et al*., 2014 ; Alam *et al*., 2018). (Saravanou *et al*., 2015) avaient estimé la quantité de bruit résiduel contenu dans un jeu de tweets filtrés par mots-clés, à partir de la sélection aléatoire d'un échantillon de 1 000 tweets : après relecture manuelle, 11,5% des tweets contenant les mots-clés recherchés étaient employés dans un autre contexte que celui de l'étude.

Il existe néanmoins des méthodes qui permettent d'évaluer l'adéquation entre un tweet et le contexte de la problématique d'étude. Au niveau des outils automatisés, les algorithmes d'apprentissage automatique sont également utilisés pour valider les tweets utiles. En intelligence artificielle, l'algorithme d'apprentissage[32] désigne un outil fonctionnant sur une approche probabiliste permettant aux ordinateurs d'optimiser leurs performances dans l'exécution de tâches, et ce, en apprenant à partir de données similaires à celles qu'ils doivent traiter. Ces outils fonctionnent en deux phases : la phase d'entraînement (ou d'apprentissage) est accomplie par le recours à un jeu de données test à partir desquelles l'algorithme apprend à différencier les données (dans le cas de la suppression du bruit résiduel, il s'agit d'apprendre à distinguer les tweets utiles des tweets au contenu non approprié : la phase d'entraînement est alors fondée sur un jeu de tweets déjà étiquetés comme valides ou bruités). La seconde phase correspond à la mise en production de l'apprentissage : l'algorithme traite le jeu de données à classer. (Sakaki *et al*., 2010) avaient recours à un algorithme de type *SVM* (*Support Vector Machine*) pour sélectionner les tweets rapportant des phénomènes sismiques évoqués au présent. Les *SVM* sont généralement utilisés pour la résolution de problème de régression ou de discrimination des données (Joachims, 1998). A ce titre, ils sont fréquemment employés pour le traitement des tweets dans le cadre de la suppression du bruit ou de l'analyse des sentiments, celle-ci permettant de catégoriser des tweets selon qu'ils véhiculent un vocabulaire et une idée *positive*, *négative* ou *neutre* (Rani et Singh, 2017 ; Naz *et al*., 2018).

[30] Dans le cas de l'utilisation des tweets géolocalisés en situation post-crise, on peut néanmoins considérer ce type de tweets comme utile : l'utilisateur qui tweete régulièrement afin de comparer des phénomènes passés et connus à des phénomènes en cours peut être un indicateur local d'intensité et de ressenti de la crise.

[31] Aux Etats-Unis, le *Memorial Day* correspond à un jour ferié marqué par les hommages aux soldats morts au combat, tous conflits confondus. Il a lieu en général autour du 25 mai ; en 2015, cette journée a été marquée par l'une des inondations les plus catastrophiques ayant frappé le Texas : à Houston, environ 305 mm de précipitations ont été cumulées en une dizaine d'heures. Source : https://www.click2houston.com/weather/remembering-houstons-2015-memorial-day-flood (Consulté pour la dernière fois le 29/09/2019).

[32] Source : https://fr.wikipedia.org/wiki/Apprentissage_automatique (Consulté pour la dernière fois le 29/09/2019)

Dans le cas de l'identification des tweets pertinents pour rapporter la survenue d'un séisme dans le temps présent, (Sakaki *et al*., 2010) avaient testé l'algorithme selon trois jeux d'entraînement étiquetés en fonction de trois critères différents (afin de tester la méthode d'apprentissage qui fournirait les résultats les plus précis) : (1) le nombre de mots dans le tweet, (2) les mots différents dans le tweet, (3) les mots situés avant et après le mot-clé recherché. Après la constitution d'un jeu d'entraînement de 597 tweets étiquetés comme positifs (et donc valides par rapport à la problématique de recherche), l'algorithme de classification était exécuté sur le jeu de tweets filtrés par mots-clés ; au final, c'est le paramètre (1), soit le nombre de mots dans le tweet, qui fournissait les meilleurs résultats (F-score[33] de 0,74). Et en effet, dans le cas de la détection de phénomènes émergents, ces auteurs, ainsi que (Bossu *et al*., 2018) présentaient la longueur d'un tweet comme un indicateur fiable : les tweets courts, constitués d'un ou d'une poignée de mots-clés (comme le tweet *Earthquake!*) sont en général les plus fiables pour détecter la survenue d'un phénomène en temps réel et sur le lieu géographique affecté.

Le problème reste qu'une partie des algorithmes appliqués aux traces numériques sont développés sur une approche de fouille de données massives (*data mining*). L'approche *data mining* a été avancée en réponse à la situation actuelle où l'explosion de la production de données dépasse amplement les capacités humaines ou techniques traditionnelles de stockage et de traitement ; les outils de traitement habituels des bases de données tendent ainsi à être remplacés par ces algorithmes d'apprentissage (Miller et Goodchild, 2015). Les informaticiens résument en général les principes de la fouille de données comme suit : "*Find a small set of precious nuggets from a great deal of raw materials*" (Han *et al*., 2012) ; ces mêmes auteurs indiquent alors que la fouille de données a pour objectif principal la découverte de tendances descriptibles dans les masses de données traitées par les algorithmes, à travers la mise en évidence de corrélations, régressions, discriminations, etc.. Pour autant, l'approche étant fondée sur les données massives, les volumes en question sont bien plus importants que les quelques centaines à quelques milliers de tweets collectés localement dans le cas des catastrophes naturelles, et la précision de la plupart des algorithmes employés s'avère sensible au volume des données à traiter. Les tweets suivants constituent deux exemples de messages filtrés par la recherche de mots-clés et validés par un algorithme d'apprentissage. Ils font partie du jeu de 10 467 tweets collectés via la plateforme SURICAT-NAT[34], relatif au phénomène d'inondation ayant frappé l'Aude en octobre 2018 :

- "*Les lagons de Polynésie c'est le paradis* ***sous l'eau*** *http://tahiti-ses-iles-et-autres-bouts-du-mo.blogspot.com/2012/09/lagons-de-polynesie.html*"

[33] Le F-score correspond à un paramètre statistique combinant deux critères d'évaluation de l'algorithme : la précision (le nombre de documents -dans notre cas, un document correspond à un tweet- pertinents trouvés par l'algorithme par rapport au nombre total de documents inventoriés dans la base de données) et le rappel (le nombre de documents pertinents trouvés par l'algorithme par rapport au nombre total de documents pertinents enregistrés dans la base de données).

[34] CF. présentation de la plateforme dans le paragraphe suivant.

- "*Macron la théorie du **ruissellement** (de nos yeux) se confirme : en théorie suppression de l'ISF pour plus d'investissements en France et création d'emplois ... en réalité suppression de l'Isf donne plus de revenus disponibles aux plus riches qui n'investissent toujours pas*".

Même s'ils contiennent les mots-clés vraisemblablement recherchés (*ruissellement* et *sous l'eau*), ces deux tweets restent attachés à un territoire distant et à une question politique. Il existe néanmoins une solution alternative au recours systématique aux algorithmes d'apprentissage automatique pour l'accomplissement de tâches consistant à discriminer les tweets utiles du bruit résiduel : cette méthode ne repose non pas sur l'intelligence artificielle mais sur l'intelligence humaine collective, via les plateformes web qui intègrent des outils de *crowdsourcing* : depuis 2005, *Amazon Mechanical Turk* vend des services dématérialisés d'exécution de micro-tâches (accomplies par des employés) dont les traitement d'images ou traitements sémantiques (Borromeo *et al.*, 2017 ; Amer-Yahia, 2019). En décembre 2017, le *BRGM* a également lancé la plateforme SURICAT-NAT[35], qui assure une veille constante sur Twitter pour la collecte rapide d'informations en cas de survenue d'un phénomène naturel dommageable (séismes, inondations, mouvements de terrain). Les tweets utiles sont traités par un algorithme d'apprentissage mais la plateforme intègre également une approche contributive[36]: tout internaute qui s'y connecte peut effectuer un tri manuel de tweets en les marquant comme liés ou sans rapport à un phénomène indiqué (figure 3.11).

Figure 3.11 : Contribution manuelle de l'internaute à la validation de tweets en rapport à un phénomène sur la plateforme SURICAT NAT

[35] http://www.suricatenat.fr/Suricate-Nat/
[36] Source : https://www.brgm.fr/actualite/lancement-suricate-nat-vigie-citoyenne-risques-naturels (Consulté pour la dernière fois le 29/07/2019)

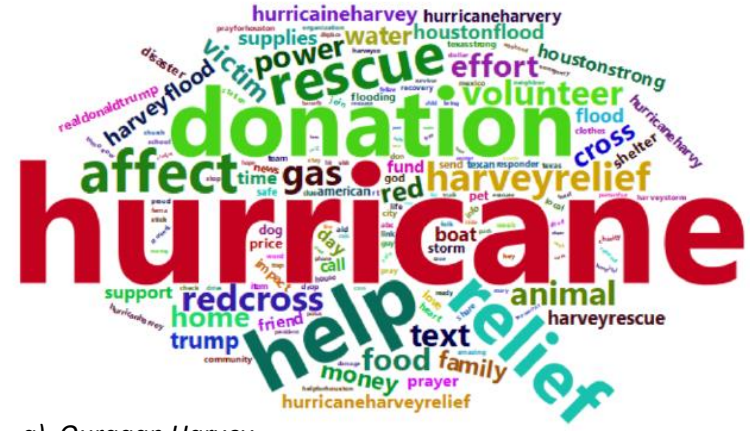
La seconde phase de l'analyse sémantique consiste à explorer la construction et le contenu du jeu de tweets utiles, ou encore à classifier les tweets en fonction de leur contenu lexical. Les tweets émis en réponse à un phénomène naturel peuvent se focaliser sur différents aspects de la crise : diffusion de l'alerte, observations de phénomènes physiques, opérations de secours, victimes et dégâts, *etc.*. (Hertfort *et al*., 2014), dont le jeu sur les inondations de l'Elbe n'était constitué que de 398 tweets filtrés et validés, proposaient de les inventorier en fonction du contenu sémantique, en quatre classes : *actions bénévoles* (tout tweet mentionnant une mesure destinée à diminuer l'impact des inondations), *média* (tweet relatif à la couverture médiatique du phénomène ou à toute action politique), *conditions de circulation* (tweet relatif à toute perturbation du trafic ferroviaire ou routier), *niveau de l'eau* (tweet relatif à une mesure quantitative ou une appréciation qualitative du niveau des eaux). Tous les tweets non inclus dans ces catégories étaient alors étiquetés comme *autres* (et ils représentaient la part la plus importante des tweets dans l'ensemble du jeu, soit 31,66%).

La classification des tweets peut encore s'opérer en fonction d'autres critères, et de manière automatisée. Comme indiqué précédemment, l'une des pratiques courantes dans l'analyse sémantique du contenu des tweets consiste à effectuer une *analyse de sentiments* ; (Widener et Wenwen, 2014) présentaient l'analyse de sentiments par les tweets comme une tâche ardue en raison de la brièveté et de l'hétérogénéité générales des contenus numériques créés par les internautes. Pour analyser l'opinion des utilisateurs sur différents aliments, ils commençaient néanmoins par collecter des mots assimilés à de bons indicateurs d'opinion, en général des adjectifs et des adverbes (constitution des données d'apprentissage). La classification des tweets s'est effectuée de nouveau par un algorithme d'apprentissage, entraîné à partir de données déjà classées et qui assignait un sentiment positif, négatif ou neutre en fonction de mots marqueurs de connotations positives ou négatives. Dans leurs travaux sur les ouragans, (Alam *et al*., 2018) positionnaient l'analyse de sentiments comme un indicateur permettant d'identifier les inquiétudes, la panique mais également des manques de la gestion de crise (en situation de retour d'expérience). Leur classification montrait ainsi, sans surprise, une large domination des tweets au sentiment négatif ayant pour cause vraisemblable, en début de crise, la présence d'un vocabulaire de plainte, de colère mais également vulgaire. Pendant la crise, la forte présence de sentiments négatifs s'expliquait par les individus constatant des dégâts ou critiquant la gestion par les autorités.

La classification des messages peut également s'effectuer en fonction de la phase pendant laquelle le tweet est émis. (Roy Chowdhury *et al*., 2013) ont développé un algorithme de classification des tweets par période du phénomène (*Tweet4act*), fondé sur deux paramètres lexicaux : l'établissement d'un glossaire inventoriant les mots de vocabulaire typiques des phases pré, pendant et post phénomène ainsi que l'analyse des temps des verbes. *Tweet4act* analyse ainsi les mots inclus dans chaque tweet, en se référant à trois règles attribuant un score au mot : (règle 1) si le mot est inclus dans le glossaire, ajouter 1 à la période correspondante et arrêter le processus ; (règle 2) si le mot est un auxiliaire, ajouter 1 à la période correspondante au temps de l'auxiliaire ; (règle 3) si le temps du verbe principal est

en accord avec le temps de l'auxiliaire, ajouter 0,5 à la période correspondante à l'auxiliaire. La somme des scores de chaque période est calculée pour chaque tweet et la période assignée est celle qui contient alors le score le plus fort. Evalué sur les trois phénomènes indiqués au paragraphe 3.2.3 (séisme de Haïti en janvier 2010, tornade de Joplin dans le Missouri en mai 2011, typhon Nesat aux Philippines en septembre 2011), sa précision est comparée aux résultats obtenus par d'autres algorithmes d'apprentissage semi-automatiques : *Tweet4act* se montre finalement plus précis que ces algorithmmes classiques dans le cas de la tornade et du typhon.

Enfin, l'analyse sémantique peut se focaliser sur les mots seuls, sans chercher à catégoriser les tweets qui les incluent. Le nuage de mots est l'une des représentations les plus courantes pour offrir un rapide aperçu du lexique inclus dans un jeu de tweets, en fonction de la récurrence des mots-clés identifiés (Godfrey *et al*., 2014 ; Allen et McAleer, 2018). Sa sémiologie fonctionne de la manière suivante : la taille de la police est proportionnelle à la fréquence des mots représentés. En conséquence, plus le mot est utilisé dans les tweets filtrés, plus sa taille dans le nuage est grande et inversement. La figure 3.12 (page suivante) affiche les nuages de mots constitués par (Nguyen *et al*., 2019) dans le cadre de la première exploration lexicale des jeux de tweets filtrés pour l'étude sémantique des ouragans Harvey et Irma, tous deux survenus en 2017.



a). Ouragan Harvey

b). Ouragan Irma

Figure 3.12 : Nuages de mots construits à partir des jeux de tweets filtrés liés aux ouragans Harvey et Irma (Source : Nguyen et al., 2019)

Le nuage de mots permet ainsi de comparer les tendances de l'activité *tweeting* dans différents lieux géographiques, différentes temporalités ou comme ici, à l'occasion de phénomènes de nature identique mais aux impacts variables : le discours lié à l'ouragan Irma témoigne d'une certaine hétérogénéité (on peut identifier des thèmes généraux comme les éléments physiques - *winds*, *storm* - ou encore du vocabulaire lié aux différentes phases de la crise - *forecast*, *warning*, *emergency*, *evacuation*, *shelters*, *devastation*, etc.) ; en revanche, l'ouragan Harvey se distingue par son lexique ciblé exclusivement sur les mesures destinées à enrayer la crise sociale et environnementale provoquée par le phénomène physique : on note la présence de mots-clés témoignant des besoins urgents de secours (*rescue*, *help*, *victim*, *relief*) mais également des besoins en termes de ravitaillement (*food*, *gas*) et d'appels aux dons (*fund*, *money*, *donation*).

(Jung, 2014) proposait une autre forme de nuages de mots, le *code cloud*, dont la construction n'est pas fondée uniquement sur l'aspect quantitatif (c'est-à-dire la fréquence

des mots dans l'ensemble du jeu de tweets), mais sur une approche qualitative destinée à cerner les idées-clés contenues dans un jeu de tweets spatialisés. En d'autres termes, là où le nuage de mots traditionnel privilégie la représentation d'un contenu brut en fonction de sa fréquence au détriment des *outliers* (soit les mots marginalement employés qui peuvent parfois apporter plus de sens), le *code cloud* identifie et représente des thèmes en fonction des espaces de l'activité tweeting. A partir d'un foyer de concentration identifié sur le campus universitaire de Seattle, (Jung, 2014) procède à la lecture des tweets et à leur annotation manuelle en fonction de l'idée générale qui se dégage de chaque tweet. Il identifie ainsi des thèmes généraux propres à l'environnement d'émission des tweets (*building*, *exam*, *professor*, *homework*, *food*, *drink*) mais également un vocabulaire émotionnel (*struggle*, *hate*, *excitement*, *complaint*). La figure 3.13 affiche le nuage de mots traditionnels et le *code cloud* représentant les thèmes annotés par l'auteur, dans le territoire exploré (rectangle violet) :

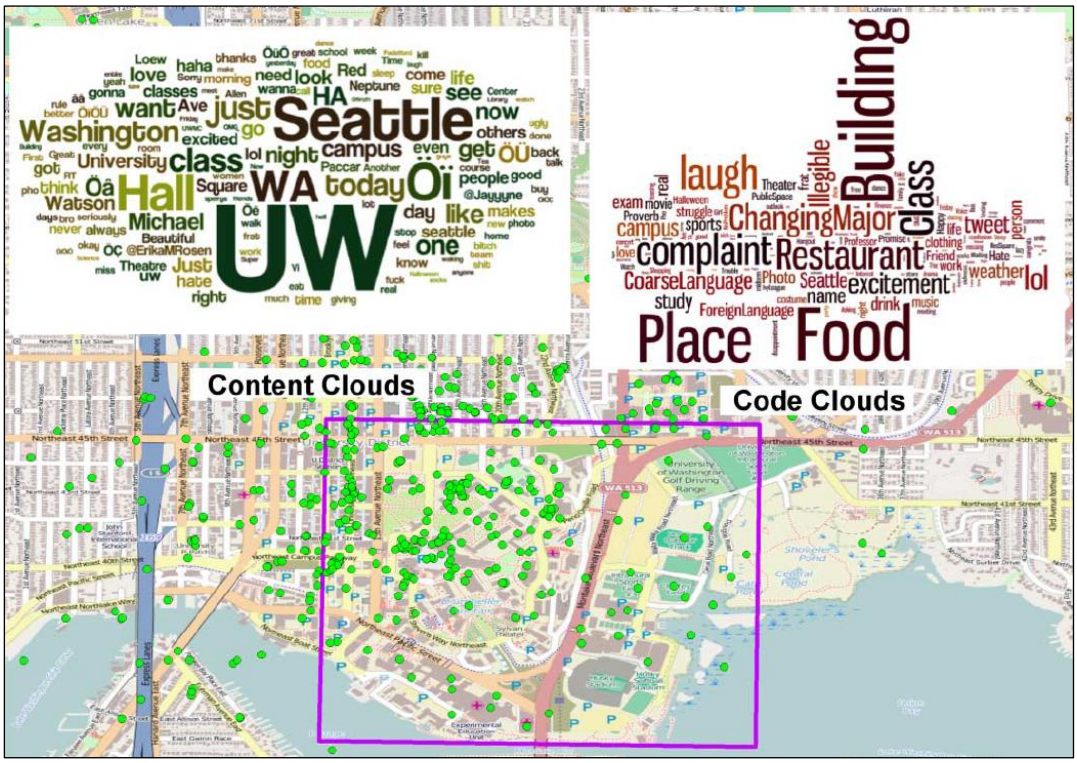


Figure 3.13 : Nuage de mots traditionnels et code cloud pour l'exploration sémantique des tweets géolocalisés émis sur le Campus de Seattle (Source : Jung, 2014)

Le *code cloud* est ainsi présenté comme un outil permettant de réduire l'hétérogénéité sémantique des tweets et ainsi de faire émerger plus facilement le sens des tweets émis en fonction des lieux. En effet, le *code cloud* présenté assure l'identification rapide des thèmes et sentiments majoritaires dans le corpus de tweets (alimentation, cours, plaintes, rires, etc.) alors que dans le nuage de mots traditionnels, ils apparaissent noyés dans une majorité de mots qui sont en rapport aux territoires (*WA*, *Washington*, *Seattle, UW*) mais sans apporter de sens supplémentaire quant à l'usage de ces lieux ou les sentiments suscités : par exemple,

le thème *laugh* affiché dans le *code cloud* de droite est visible dans le nuage de mots de gauche par la présence du *lol* (acronyme de *Laughing Out Loud*) et de l'interjection *haha* ; en revanche, les thèmes de *food* ou encore de *complaint* ne sont pas identifiables par les mots du nuage de gauche représentant le contenu brut des tweets géolocalisés.

En dernier lieu, l'identification de thèmes (*topic modelling*) est fréquemment rencontrée sous l'application de la *LDA* (*Latent Dirichlet Allocation*). (Alam *et al*., 2018) présentaient la LDA comme une méthode permettant de mettre en évidence des structures lexicales cachées dans une masse de tweets générés en réponse aux ouragans ; ayant préalablement catégorisé leurs tweets filtrés en fonction de classes thématiques (victimes, dégâts, dons, soutien, conseils, personnes disparues, personnes affectées, communication personnelle.), ils appliquaient la LDA à l'ensemble des tweets regroupés dans leur dernière classe *autre information utile* (qui contient donc tous les tweets non étiquetés dans les huit catégories de leur taxonomie). Concrètement, la LDA est un modèle statistique probabiliste qui permet d'identifier des thèmes contenus dans un ensemble de documents ; elle effectue des séries d'échantillons de mots contenus dans le corpus de documents et identifie ceux qui ont tendance à apparaître ensemble (Ghosh et Guha, 2013), qu'on nommera ici cooccurrences ou associations lexicales : ces cooccurrences lexicales sont affichées dans des ensembles de mots, les *topics*. En revanche, contrairement à un algorithme de clustérisation qui affecte chaque objet dans un cluster précis, la LDA associe fréquemment un mot identique à différents *topics* mais avec des probabilités d'appartenance variables. Pour chaque *topic*, il est ainsi possible d'identifier les mots dont les probabilités d'appartenance sont les plus fortes ; de la même manière, il est possible de connaître, pour chaque document du corpus, le *topic* avec lequel il a la plus forte probabilité d'être lexicalement lié. La LDA fait tout de même appel à une supervisation : c'est l'utilisateur qui doit choisir le nombre de *topics* construits par le modèle ainsi que le nombre de mots pris en compte dans chaque *topic*. Dans un second temps, il appartient également à l'utilisateur d'interpréter le sens des *topics* générés par le modèle : (Alam *et al*., 2018) étaient ainsi parvenus à identifier des *micro-événements* dans leur classe de tweets *autre information utile* : pendant l'ouragan Harvey, un fabricant de matelas mettait ses magasins à disposition des sinistrés ; pendant l'ouragan Irma, une critique du président Trump qui manifestait son intention d'expulser 800 000 individus du programme DACA[37] vers le Mexique, au moment même où le Mexique envoyait du ravitaillement aux Etats-Unis en aide aux sinistrés de l'ouragan.

[37] Mis en place sous la présidence Obama en 2012, le programme DACA permet de protéger de l'expulsion les immigrés irréguliers arrivés aux Etats-Unis dans leur enfance. Source : https://www.lemonde.fr/ameriques/article/2018/11/06/jeunes-immigres-dreamers-aux-etats-unis-trump-saisit-la-cour-supreme-pour-mettre-fin-au-programme-daca_5379720_3222.html (Consulté le 30/07/2019).

3.3.2. Méthodes statistiques et d'analyse spatiale

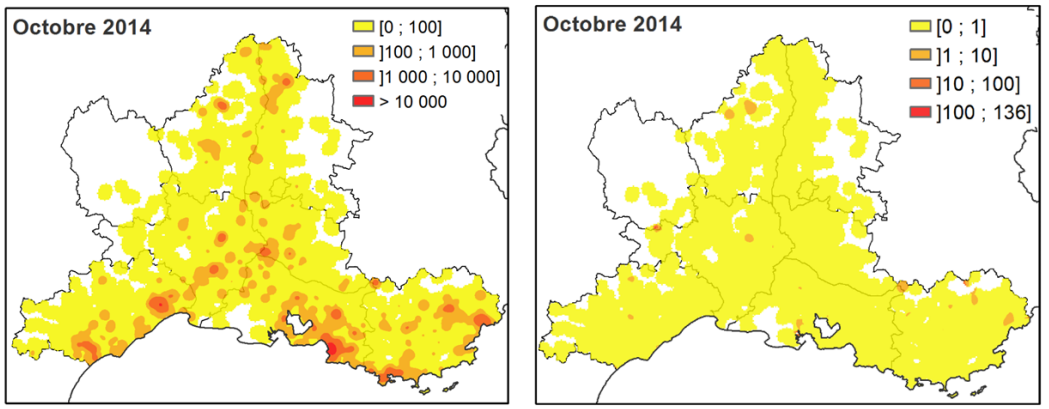
Certains travaux croisant tweets et jeux de données traditionnelles cherchent à évaluer le potentiel du tweet comme variable prédictive d'un comportement ou explicative de disparités sociales. Tests bivariés et analyses multivariées sont ainsi fréquemment employés chez les géographes : (Ghosh et Guha, 2013) identifiaient une corrélation négative entre le taux d'obésité et le nombre de tweets liés à ce sujet sanitaire. Ils expliquaient alors ce constat par le fait que les individus virtuellement actifs sur le sujet se situaient dans les territoires à niveau de vie et d'éducation élevés, et discutaient des enjeux de santé publique que ce problème représente sans être physiquement concernés par l'obésité. De la même manière, (Gomide *et al.*, 2014) mettaient en évidence par régression linéaire la capacité des tweets à prédire, à hauteur de 90%, le nombre de cas de dengue dans les métropoles brésiliennes.

Afin de modéliser les liens éventuels entre variables socio-démographiques et opinions positives ou négatives énoncées sur les types d'aliments dans les tweets, (Widener et Wenwen, 2014) accomplissaient une régression logistique. Ils mettaient alors en exergue qu'un tweet localisé dans un territoire à faibles revenus et à faible accès aux produits alimentaires non transformés fait diminuer la probabilité que le tweet énonce un avis positif sur la nourriture saine ; à l'inverse, plus l'âge médian et le nombre d'hommes sont élevés sur une entité de recensement, plus la probabilité de rencontrer des tweets énonçant un avis positif sur les aliments sains est forte. Dans le domaine des risques naturels, (Sakaki *et al*., 2010) représentaient graphiquement la variabilité des émissions de tweets émis en réponse à une secousse sismique et mettaient en évidence l'existence d'une relation exponentielle décroissante permettant de prédire le nombre de tweets émis en fonction du temps écoulé depuis la survenue du séisme.

Les méthodes statistiques quantitatives usuelles peuvent également être utilisées via la création d'indices. En effet, les quantités de tweets émis en un lieu peuvent être biaisées par de multiples facteurs : (Li *et al*. 2012), montraient déjà que les densités de contenus géolocalisés créés sur les plateformes Twitter et Flickr reflétaient les densités de population (en conséquence, les espaces urbains ou métropolitains densément peuplés concentrent davantage d'activité tweeting que les espaces ruraux). Cet effet de densités de population sur l'activité tweeting peut toutefois être nuancé par l'existence d'utilisateurs *virtuellement hyperactifs* qui peuvent générer à eux seuls de grandes quantités de tweets dans des territoires faiblement peuplés (Cavalière *et al*., 2016). Pour gommer des effets de densités de population ou de tweets dans le cadre d'études quantitatives, on peut alors normaliser les quantités de tweets émis en réponse à un événement social ou à un phénomène naturel en les rapportant au nombre d'habitants (à une unité spatiale fine, [Cavalière *et al*., 2016]) ou encore à l'activité quotidienne normale (Saravanou *et al*., 2015 ; Lucchini *et al*., 2016). La figure 3.14 ci-après représente, à gauche, la cartographie globale de l'activité tweeting géolocalisée émise dans huit départements du sud-est de la France en octobre 2014, sous forme de carte de chaleur : cette première représentation fait nettement transparaître le réseau urbain

(vallée du Rhône, réseau des villes méditerranéennes). La carte de droite rapporte cette activité tweeting globale au nombre d'habitants (données issues du carroyage de l'INSEE) : la nouvelle carte de chaleur construite met alors en évidence :

- les territoires dans lesquels on trouve moins ou autant de tweets géolocalisés que d'habitants recensés (ratio compris entre 0 et 1) : l'activité écrasante des pôles urbains par rapport aux territoires périurbains ou ruraux est ainsi gommée ;
- les territoires dans lesquels on trouve plus de tweets que d'habitants (ratio supérieur à 1), qui correspondent à ces territoires périurbains ou ruraux très rarement représentés dans les études focalisées sur les tweets géolocalisés.



a). Densités de tweets géolocalisés *b). Ratio Tweets géolocalisés / Nombre d'habitants*

Figure 3.14 : Densités brutes de tweets géolocalisés et densités normalisées par le nombre d'habitants (Source : Cavalière et al., 2016)

Les méthodes d'analyse reposant sur la dimension spatiale des tweets géolocalisés, quant à elles, peuvent faire appel à de simples outils des SIG jusqu'aux modèles de classification plus élaborés : (Chen et Yang, 2014) modélisaient l'environnement alimentaire d'un tweet en appliquant des zones tampon de 0,5 et 1 mile (soit respectivement 0,8 et 1,6 km) autour de chaque tweet et en comptabilisant le nombre d'établissements localisés dans cet espace délimité. (Lucchini *et al.*, 2016) avaient recours à des grilles carroyées destinées à détecter des événements virtuels à une échelle spatiale fine et affranchie des unités administratives traditionnelles. (Kounadi *et al.*, 2015) cherchaient à modéliser les effets de distance au lieu de survenue d'un événement dans la réponse virtuelle (figure 3.15) : la représentation graphique du nombre de tweets en fonction de la distance au lieu des homicides indiquait en effet l'existence d'une relation décroissante d'ordre logarithmique (plus la distance au lieu augmente, moins on trouve de tweets relatifs à l'événement[38]).

[38] Ce constat peut certainement être nuancé dans les réponses virtuelles aux événements les plus violents comme les attentats : à l'échelle de la ville, on pourra sans doute observer une correspondance entre lieux frappés et lieux de l'activité virtuelle mais celle-ci va se diffuser bien au-delà des lieux affectés dans le réel.

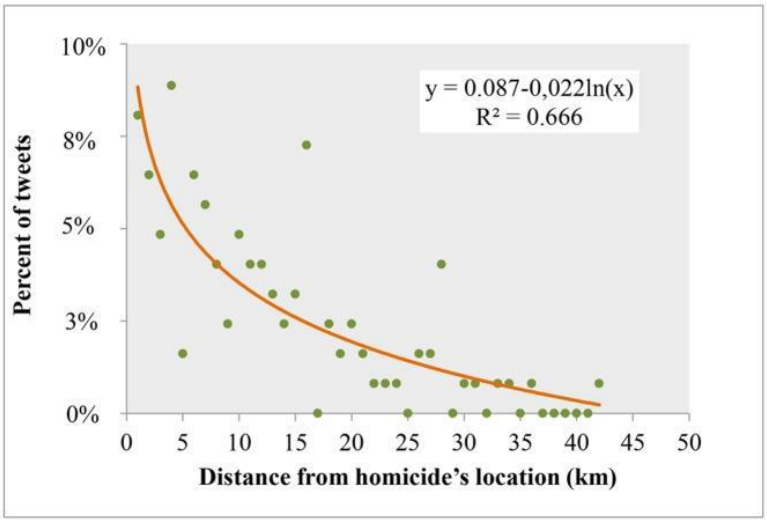


Figure 3.15 : Effets de la distance d'un homicide sur le pourcentage de tweets émis liés au crime en question (Source : Kounadi et al., 2015)

D'autres travaux s'appuient sur les méthodes de partitionnement spatial (Andrienko *et al*., 2013 ; Saravanou *et al*., 2015) afin de transformer le tweet géolocalisé sous forme de points en données interprétables et destinées à servir de support à d'autres analyses. Quelles sont alors les méthodes rencontrées, dans la littérature des tweets géolocalisés, pour l'analyse des semis de points ? En tout premier lieu, la calcul des densités de Kernel s'avère être une méthode de représentation cartographique fréquente dans la mesure où elle permet d'identifier et de quantifier rapidement les lieux de l'activité tweeting (Hertfort *et al*., 2014 ; Cavalière *et al*., 2016). Mais l'une des questions récurrentes qui émergent lorsqu'il est question d'analyse spatiale et de représentation cartographique de tweets géolocalisés reste le passage d'une trace ponctuelle à une donnée polygonale. Dans leurs travaux sur les inondations au Royaume-Uni, (Saravanou *et al*., 2015) cherchaient dans un premier temps à identifier des *régions* en fonction de l'intensité du phénomène ; ils affirmaient ainsi "*Given that we need to identify areas or regions, we need to go beyond the GPS coordinates of a single tweet. Therefore, we need to aggregate the geotagged information of our tweets to form these larger areas*". Leur première étape consistait donc à agréger spatialement les tweets par clustérisation en fonction de la distance euclidienne et en donnant un nombre de clusters à former. A la seconde étape, ils pouvaient alors générer un diagramme de Voronoï dessinant les polygones formant des régions supposées homogènes autour clusters de tweets (figure 3.16).

Figure 3.16 : Diagrammes de Voronoï des tweets clustérisés (Source : Saravanou et al., 2015)

Comme l'indique la figure précédente, se pose la question de la résolution spatiale des agrégats constitués : à quel niveau d'agrégation spatiale, défini ici par le nombre de clusters à former et la distance euclidienne entre deux points, pourra-t-on dégager des tendances ? En l'occurrence, les résultats indiquaient une résolution spatiale fine des tweets : plus le nombre de clusters augmentait, plus les différentes tendances spatiales se distinguaient (pour rappel, le premier objectif des auteurs consistait à identifier les territoires qui avaient été les plus affectés, en analysant soit le nombre total de tweets, soit en calculant un indice comme nous l'avons indiqué plus haut dans le texte).

Pour autant, l'un des outils d'agrégation spatiale les plus fréquemment employés reste l'algorithme de clustérisation DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), qui effectue une partition de données ponctuelles en fonction des critères de densité et de voisinage (Sakkari *et al*., 2019). L'application de cet algorithme requiert la définition préalable de deux paramètres : l'utilisateur doit indiquer une distance seuil au-delà de laquelle deux points ne seront pas agrégés dans le même cluster ainsi que le nombre minimal de points agrégés pour former un cluster. Tout point dont la distance aux points voisins est supérieure à la distance seuil est considéré comme du *bruit* et sera exclu de la partition finale (Andrienko *et al*., 2013). Cet outil présente deux avantages : l'utilisateur n'a pas besoin de spécifier le nombre de clusters à former et en conséquence (Gomide *et al*., 2011), les clusters formés présentent *une forme arbitraire, en adéquation au fait que la distribution spatiale des individus ne s'organise pas selon des logiques régulières*[39] (Rodriguez Dominguez *et al.*, 2017).

Les contenus géolocalisés émis sur le Web 2.0 ayant tendance à révéler une alternance entre lieux concentrant de fortes densités de contenus et lieux de faibles densités de

[39] Des logiques spatiales peuvent néanmoins apparaître lors d'événements particuliers : par exemple, lors d'une manifestation, le parcours suit le tracé linéaire du réseau de voirie avant d'aboutir généralement sur une place.

contenus, DBSCAN reste régulièrement employé pour détecter les hauts-lieux de l'activité virtuelle des milieux urbains. (Rodriguez Dominguez *et al*., 2017) ainsi que (Capdevila *et al*., 2017) utilisaient l'algorithme DBSCAN afin de détecter des événements localisés dans des lieux d'activité sporadique s'agitant lors de phénomènes naturels ou d'événements culturels perturbant la routine quotidienne. Quelles préconisations peut-on alors suivre pour paramétrer distance seuil d'agrégation et nombre minimal de points pour former un cluster ? (Rodriguez Dominguez *et al.*, 2017) calculaient une matrice de distance entre tous les points géolocalisés et sélectionnaient la distance minimale repérée entre deux points ; (Soliman *et al*., 2015) fixaient ce seuil à 250 mètres, étant donné qu'ils souhaitaient identifier les objets précis qui polarisent les tweets dans les lieux de l'activité tweeting, en prenant en compte l'incertitude spatiale de la mesure GPS. Concernant le nombre minimal de points, (Rodriguez Dominguez *et al.*, 2017) le fixaient arbitrairement à trois points ; (Soliman *et al*., 2015) le fixaient à quatre points afin de s'assurer que la présence de tweets proches n'était pas due à un hasard mais à la présence d'un objet particulier.

Dans tous les cas, il restera toujours, quelle que soit la méthode de partitionnement spatial utilisé, une part de risque et de subjectivité induite par le choix d'une distance seuil, d'un nombre minimal de points ou d'un nombre d'entités spatiales à former. Comme le montrent les travaux consultés, il est alors nécessaire de procéder à plusieurs essais de paramétrages et d'explorer la configuration qui permettra de distinguer des structures spatiales.

3.3.3. Les questionnements épistémologiques en faveur d'un changement de paradigme

Au regard des travaux effectués depuis une dizaine d'années et dont les lignes essentielles ont été présentées dans les paragraphes précédents, une conclusion se dessine : les chercheurs placent les traces numériques géolocalisées comme une opportunité exceptionnelle offrant des perspectives inédites pour porter un nouveau regard sur la compréhension des événements sociaux et des phénomènes physiques dont les manifestations varient dans l'espace et dans le temps. Pour autant, si l'on étend la recherche bibliographique, on pourra rapidement s'apercevoir de la tendance à la redondance des questions de recherche, et notamment en ce qui concerne la problématique des risques naturels. (Steiger *et al*., 2015) listaient une douzaine de publications focalisées sur la détection de phénomènes ou la gestion de crise, et les méthodes employées s'avéraient similaires d'un papier à l'autre : extraction de tweets utiles par mots-clés, cartographie de densités de tweets, clustérisation, exécution de la LDA, carroyage, *etc.*. Et malgré le regain d'attention accordée aux tweets suite à la saison cyclonique de 2017, les papiers les plus récents cités dans les paragraphes précédents s'inscrivent dans cette même lignée. La recherche peine-t-elle alors à fouiller davantage le contenu et les structures qui peuvent se dégager de l'analyse des tweets géolocalisés ?

(Dashti *et al.*, 2014) soulevaient l'épineuse question des capacités des méthodes d'analyse spatiale usuelles de la géographie à traiter et à exploiter les traces numériques, celles-ci ayant en effet été développées à une période où la connaissance des territoires était exclusivement fabriquée par l'analyse des jeux de données institutionnels : ainsi, on ne sait pas par avance si ces méthodes sont adaptées aux traces numériques dans l'objectif de déployer leur plein potentiel (Kitchin, 2013), de même que, lorsqu'on collecte un jeu de tweets, on ne peut pas prédire à l'avance des structures spatiales, temporelles ou sémantiques qu'on pourra (ou qu'on ne pourra pas) mettre en évidence[40]. Ainsi, cette perspective d'une nouvelle compréhension des populations et des territoires se heurterait au manque de moyens qui permettraient d'approfondir l'analyse de traces spatio-temporelles, sémantiques et hétérogènes : et en effet, si le chercheur qui travaille avec des *Big Data* développe des algorithmes capables de mettre en évidence des tendances dans des jeux de données qui se mesurent désormais en téra- ou pétaoctets, les questions relevant des sciences humaines sont traitées à une résolution plus fine et s'appuient sur des jeux de données peu volumineux. L'acquisition de connaissances à partir de l'analyse exploratoire de ces données rencontre ainsi cet obstacle du manque de méthodes statistiques et d'analyse spatiale appropriées (Steiger *et al*., 2015). (Kitchin, 2013) affirmait effectivement que les méthodes statistiques utilisées dans l'analyse des données géographiques avaient peu évolué depuis les années 1990 et que les "*new forms of data sciences are in their infancy*". C'est pourquoi l'évolution des méthodes s'est plutôt traduite, très récemment en géographie, par un questionnement épistémologique sur les possibilités de donner du sens à ces traces en prenant en compte leurs dimensions géographique et sociale (Kitchin, 2013). Un nouveau paradigme a alors émergé au sein de cette communauté scientifique : la science guidée par les données, ou *data-driven science* (Miller et Goodchild, 2015). Les prémices des *data-driven sciences* ont ainsi préconisé un rejet systématique de l'inférence déductive afin de privilégier l'attention portée à l'observation de structures inattendues dans les données, à la formulation d'hypothèses *a posteriori* et à leur test, et ce, dans le but de construire progressivement une connaissance ancrée dans les données (Kitchin, 2013).

L'idée de construction d'une connaissance ancrée sur l'observation des faits qui se dégagent des données a d'ores-et-déjà été introduite dans certaines disciplines il y a quelques décennies, et se trouve de nouveau portée en avant par une poignée de chercheurs en géographie (Kitchin, 2013 ; Miller et Goodchild, 2015), animés par la volonté de faire progresser leur discipline et de l'intégrer au *data deluge* (Anderson, 2008). Envisager autrement la production de connaissances scientifiques devient impérieux pour les sciences humaines, qui tendent désormais à intégrer des postures de recherche rejetant la norme "*if there is no hypothesis, it is not science*" (Kell et Oliver, 2003), celle-ci signifiant que toute connaissance naît d'une hypothèse validée par l'expérience. Les paragraphes suivants

[40] Comme le montreront les chapitres 5 et 6 de la seconde partie du manuscrit, les résultats des analyses exécutées à partir des jeux de tweets collectés et des données complémentaires utilisées dans cette recherche se démarqueront des résultats fournis par les auteurs cités dans les sections 3.1 et 3.2 de ce chapitre.

présentent ainsi les théories qui encadrent la production de la connaissance dans une posture située en dehors de la démarche hypothético-déductive.

"Simply gathering data without having any specific question in mind is an approach to science that many people are doubtful about. Modern science is supposed to be hypothesis-driven" (Kell et Oliver, 2003). Les démarches traditionnelles de la recherche reconnaissent en général les approches fondées sur l'hypothèse initiale qui oriente le choix des données et le déroulement de l'expérience. L'*hypothesis-free science* ne pose pas d'hypothèse initiale mais cherche à la construire. Comment ? (Kell et Oliver, 2003) préconisent d'appréhender le phénomène observé comme un système, c'est-à-dire un ensemble d'éléments en interaction, afin d'identifier les paramètres qui lient ces différents éléments et les variables qui les contrôlent. L'étude du système favoriserait alors la recherche des facteurs qui expliquent la survenue d'une observation particulière. La primauté est ainsi donnée à l'*observation* et à la *sérendipité,* terme qui qualifie la découverte fortuite voire accidentelle. L'hypothèse émerge donc progressivement par juxtaposition d'observations et qualifie une *proposition spécifique relative au comportement d'un système, fondée sur un raisonnement logique et permettant de construire une prédiction expérimentalement vérifiable*[41] (Kell et Oliver, 2003).

Théorisée en 1967 par Glaser et Strauss, la *Grounded Theory* s'ancrait déjà dans cette posture consistant à se détacher de tout *a priori* sur les données. Considérée comme une approche inductive qui appréhende les données comme prémices de toute théorie, elle s'oppose à la démarche hypothético-déductive, non seulement dans la formulation des postulats mais encore dans le regard porté aux données "*qui* [dans la démarche déductive] *ne servent que d'exemples pour valider des théories existantes"* (Guillemette, 2006). La *Grounded Theory* se positionne comme démarche de déploiement des données, c'est-à-dire qu'elle focalise son attention aux tendances qui émergent du terrain ou des acteurs qui vivent les phénomènes étudiés. Elle vise alors à développer son analyse selon des questionnements indépendants de tout cadre théorique et conceptuel existant afin de découvrir des idées inédites ; ce sont ces questionnements, qui découlent directement des faits émergents du terrain, qui orientent les analyses (Walters, 2012). Cette approche rejette ainsi toute formulation d'hypothèse initiale et tout recours systématique à l'état de l'art pré-analyse, afin de favoriser l'adoption d'une posture d'ouverture à l'inattendu et de faire abstraction de tout cadre préconçu qui s'imposerait à l'étude ; de la même manière, il est recommandé d'éviter de formuler toute question de recherche ou problématique qui pousserait à orienter inconsciemment l'analyse des données (Walters, 2012). En conséquence, objet de recherche et question de départ sont provisoires et se construisent progressivement, en réponse à

[41] L'exploration des traces numériques géolocalisées rejoint justement cette considération ; comme indiqué précédemment, on ne sait pas par avance quels phénomènes on pourra (ou pas) mettre en évidence par l'analyse des tweets géolocalisés (même si, dans le cas des réponses aux phénomènes naturels, l'on se pose un certain nombre de questions préalables auxquelles on souhaite savoir s'ils ont la capacité ou non d'apporter des pistes de réponses). Il faut donc commencer l'analyse sans idée préconçue et construire la théorie à partir des comportements observés de l'activité virtuelle.

l'imprévisible qui se manifeste en cours d'analyse et à l'ouverture de nouvelles pistes exploratoires.

Hypothesis-free science et *Grounded Theory* font toutes deux appel au raisonnement abductif : celui-ci a été envisagé par quelques chercheurs comme inférence clé dans l'exploration des données et la découverte de phénomènes intéressants cachés dans les traces numériques hétérogènes et données volumineuses (Miller et Goodchild, 2015). L'abduction fait référence à la capacité à mettre en relation des observations et à générer des hypothèses pour tester les relations éventuelles existant entre elles (Miller et Goodchild, 2015). Il s'agit donc d'une démarche explicative fondée sur la perception d'un fait illustrant un phénomène et la recherche des facteurs qui le provoquent et l'influencent (Walters, 2012). Théorisée par le philosophe Charles Pierce à la fin du XIX^{ème} siècle, l'abduction consiste à examiner *'[…]a mass of facts and allowing these facts to suggest a theory'* (Hoffmann, 1999). Elle s'entreprend de la manière suivante : on observe un phénomène B ; on infère l'hypothèse A comme facteur explicatif plausible de B (et l'essence de l'abduction se situe à cette étape de réflexion et de recherche de facteurs explicatifs). L'hypothèse A suggère l'expérience : si A est vraie, alors le phénomène observé B peut s'imposer comme une conséquence évidente de A. En d'autres termes, si on observe B dans un contexte différent, on peut suspecter l'existence de A. La répétitivité de l'expérience transformera l'hypothèse A en théorie, par induction. L'abduction n'est donc pas une méthodologie de recherche en soi mais une étape du raisonnement inductif.

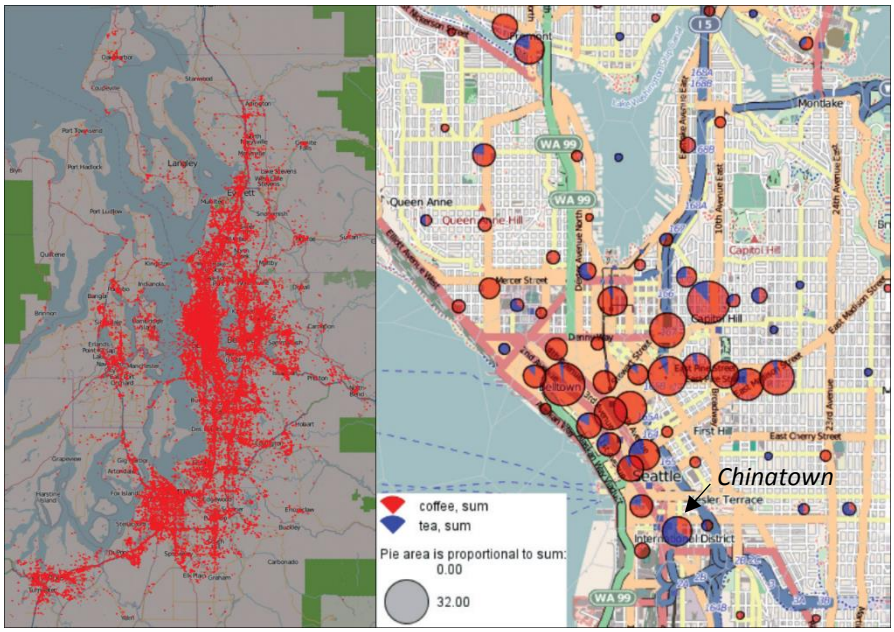
Les postures et approches évoquées ci-avant ont toutes une caractéristique en commun : elles tendent à s'affranchir de la formulation d'une hypothèse initiale qui est perçue comme un étau limitant le champ des possibles aux observations escomptées. En fait, certains travaux géographiques s'inscrivent, sans le mentionner, dans ces types de raisonnement : par exemple, (Ghosh et Guha, 2013) observaient, *contrairement aux attentes*, que le thème de l'obésité dans les tweets était en fait marginal dans les territoires à faibles revenus et à faible accès aux aliments frais ; par la cartographie des tweets recoupés à des variables socio-démographiques, ils en identifiaient alors le facteur explicatif (pour rappel, ce sont les utilisateurs des classes plus aisées qui discutent des enjeux liés à l'obséité, via le réseau, sans être concernés par la pathologie). Il est également nécessaire de signaler que l'inscription dans une démarche de recherche adoptant les postures évoquées, et notamment la validation par induction d'une théorie construite à la suite d'une première série d'expériences, requièrent la multiplication des études de cas (soit une recommandation déjà énoncée par [Goodchild, 2013]), et ce, afin de vérifier la validité des théories construites à partir d'un premier cas d'étude dans des cas différents[42].

[42] Dans la problématique des risques naturels, on peut tester la théorie construite à partir de l'étude d'un événement virtuel consécutif à la survenue d'un phénomène réel de deux manières différentes : la théorie est-elle vérifiée lors de la survenue d'un phénomène de même nature sur le même territoire mais à une date ultérieure ? La théorie s'applique-t-elle sur un autre territoire aux caractéristiques analogues ?

3.3.4. Les outils cartographiques au service de l'application des nouvelles approches épistémologiques

Si la question de l'adéquation entre traces numériques géolocalisées et méthodes traditionnelles utilisées en analyse spatiale de données géographiques reste pour l'heure en suspens, peut-on alors envisager d'inscrire la cartographie dans une posture de construction d'hypothèses par l'observation ? Permet-elle, au-delà des outils d'analyse spatiale, de *faire parler* les tweets ? Comme indiqué précédemment, le premier enjeu de la cartographie des tweets géolocalisés consiste à transformer l'amas de points aux coordonnées XY en information quantitative, qui offre ainsi l'avantage de mettre en évidence ces lieux qui concentrent l'activité tweeting, généralement sous forme de cartes de chaleur (Hertfort *et al*., 2014 ; Cebeillac *et al*., 2017), de cartes de clusters ou encore de cartes en symboles proportionnels (Andrienko *et al*., 2013).

Pour autant, certaines propositions cartographiques dépassent cette représentation quantitative des tweets et se présentent comme soutien à l'exploration et à la valorisation de leurs différentes composantes, et notamment sémantique (Andrienko *et al*., 2013). C'est ce que montre la figure 3.17 (en page suivante) : le semis de points (a) qui représente l'affichage de tweets bruts sur la métropole de Seattle donne une idée très générale de la distribution spatiale des tweets collectés, localisés en amas concentrés sur les différents espaces urbains centraux, en îlots dans les espaces périurbains et suivant également les axes routiers principaux. Cet amas de tweets est ensuite représenté sous forme de cercles proportionnels (b), en appliquant la méthodologie suivante : (1) génération d'un diagramme de Voronoï autour des tweets géolocalisés agrégés en clusters spatiaux ; (2) inventaire du nombre de tweets géolocalisés contenant les mots-clés *coffee* ou *tea* dans chaque polygone généré à l'étape précédente ; (3) représentation du nombre de tweets en cercles proportionnels, eux-mêmes implémentés sous forme de diagrammes circulaires. Ainsi, plus le cercle est grand, plus le nombre de tweets contenus dans un polygone est élevé ; les couleurs des surfaces représentent la proportion de tweets contenant chaque mot-clé sélectionné.



a). Tweets bruts *b). Tweets transformés en données*

Figure 3.17 : Affichage brut de tweets géolocalisés et transformation des traces pour l'identification des lieux de consommation (Source : Andrienko et al., 2013)

Ainsi, dans la figure ci-dessus, la transformation des traces brutes par leur cartographie combinant localisation (sous forme d'agrégation spatiale), sémantique et quantités de tweets permet d'identifier les lieux précis de consommation des boissons emblématiques de la métropole. En conséquence, l'apport de cette cartographie s'ancre dans une logique double : d'une part, elle soumet une proposition de transformation de traces numériques géolocalisées brutes en ayant recours aux composantes spatiale et sémantique des tweets géolocalisés. D'autre part, elle s'inscrit dans une démarche exploratoire qui propose une expérience en réaction à la constatation d'un fait : la publication d'(Andrienko *et al.*, 2013) est focalisée sur l'ensemble des tweets géolocalisés émis sur la métropole de Seattle entre juin et octobre 2011, mais elle n'est pas articulée sur une facette particulière du réseau social. La première tâche lancée par les auteurs consistait alors à étiqueter les tweets en fonction de thèmes généraux (musique, famille, amis, travail, transports, sorties, etc.) ; parmi ces thèmes, 6% des tweets faisaient directement référence à une consommation de boissons (café ou thé). La carte arrive donc en réponse à la question suivante : puisque le thème de la consommation de boissons est présent dans le jeu de tweets, comment identifier les lieux et la variabilité spatiale de ces thèmes ? En l'occurrence, la carte affiche la tendance claire de forte concentration de ces tweets dans le centre de la métropole (soit le lieu qui contient les plus fortes densités de cafés où consommer la boisson phare de la ville), à l'exception du quartier asiatique, dans lequel le thé prédomine (il s'agit du cercle que nous avons fléché sur la carte [b] de la figure 3.17).

Que proposent les autres cartographies pour valoriser les différentes composantes du tweet géolocalisé ? La carte (b) de la figure précédente soumet d'ores-et-déjà une proposition

de transformation de traces numériques géolocalisées brutes en ayant recours aux composantes spatiale et sémantique des tweets géolocalisés. Dans la même publication, les auteurs s'avancent sur la géovisualisation qu'ils présentent comme une technique de cartographie interactive permettant d'explorer les trois composantes du tweet (spatiale, temporelle et sémantique), d'identifier des points d'intérêts détectés par la cartographie statique traditionnelle préalable et de parcourir des tweets individuels. Dans la géomatique, la géovisualisation repose sur le principe de l'analyse visuelle : (Yi *et al.*, 2008) présentent l'analyse visuelle comme un processus exploratoire de construction de la connaissance, fonctionnant sur l'interaction homme-machine et destiné à soutenir le raisonnement humain autour des données. Ce processus fonctionne donc selon une approche associant ordinateur (outils de traitement et de représentation des données) et raisonnement humain pour une compréhension progressive et approfondie des données (Luo et MacEachren, 2013). (Yi *et al.*, 2008) indiquent que ce processus d'appréhension des données se déroule en quatre étapes qui s'enchaînent et de répètent de manière itérative :

- étape 1 : disposer d'un aperçu global des données, qui doit suggérer un début de piste d'analyse afin de lancer l'exploration plus approfondie des données ;
- étape 2 : ajuster l'analyse aux données d'intérêt, c'est-à-dire au sous-ensemble des données sur lequel le chercheur se focalise (l'intérêt de ces données a pu être repéré lors de l'étape précédente) ;
- étape 3 : détecter des structures intéressantes dans les données d'intérêt : distributions particulières, anomalies, clusters, tendances ;
- étape 4 : faciliter le processus cognitif en fournissant une représentation visuelle des données qui vise à réduire l'écart entre les structures cachées dans les données et la représentation mentale que s'en construit l'utilisateur.

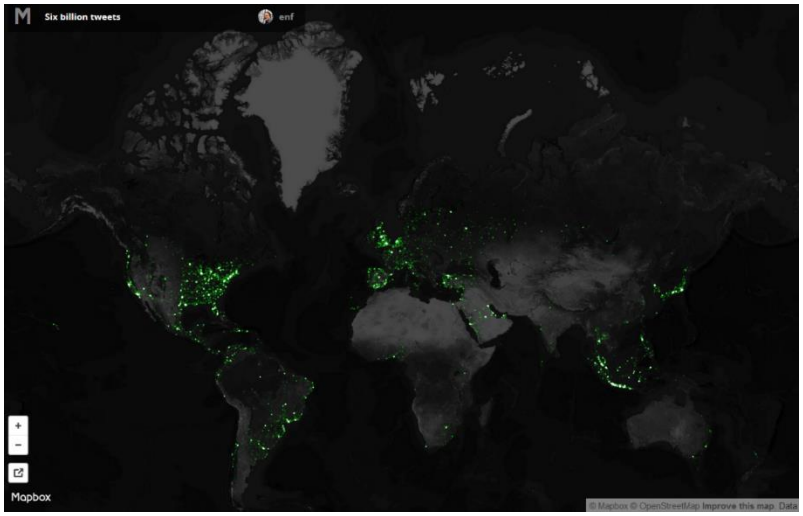
L'ensemble des étapes mentionnées ci-dessus associent donc les capacités d'observation, d'interprétation et de construction (Luo et MacEachren, 2013). (Roberts, 2005) résumait ainsi l'hypothèse de fonctionnement cognitif de l'analyse visuelle : "*When the user sees the information in different views and in different ways, they get a deeper understanding of the information".* Dans les faits, l'analyse visuelle de données ou de traces géolocalisées repose sur une combinaison d'outils de visualisation de l'information, qui peuvent se présenter sous différentes formes et dont le fonctionnement s'appuie sur l'interactivité homme-machine (Keim *et al*., 2005). L'interface de géovisualisation correspond alors à un environnement constitué de multiples fenêtres synchronisées les unes aux autres (Roberts, 2005). En d'autres termes, toute action effectuée par l'utilisateur sur l'une des fenêtres (zoom, sélection) se répercute sur l'ensemble des autres fenêtres synchronisées. L'environnement de géovisualisation hérite bien évidemment de la carte, à laquelle viennent se greffer de multiples représentations graphiques des données[43] (MacEachren *et al*., 2004) ; mais surtout, elle implique un changement du rôle de la représentation cartographique. Contrairement à la

[43] Les pages suivantes offrent un aperçu de ces différentes représentations.

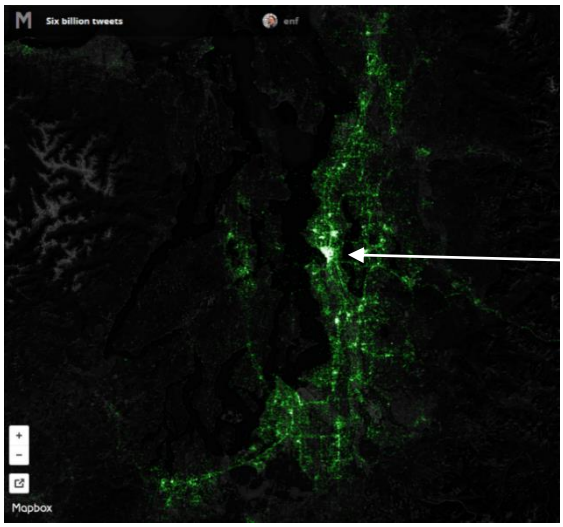
carte statique traditionnelle qui constitue un objet à lire, résultat d'une analyse et d'un mode de représentation, la carte intégrée dans l'environnement de géovisualisation s'inscrit dans la démarche exploratoire progressive décrite par (Yi *et al*., 2008) : elle soutient le raisonnement humain, la formulation d'hypothèses et finalement, la construction progressive de connaissances (Keim *et al*., 2005 ; MacEachren et Kraak, 2001).

Dans le cas particulier des tweets géolocalisés, les interfaces de géovisualisation existantes sont autant diversifiées que les cartes du Géoweb que nous avons présentées dans le chapitre 2 du manuscrit : les cartographies interactives de tweets géolocalisés peuvent ressembler aux cartes à punaises du Géoweb (Mericskay, 2016) : elles permettent de visualiser la localisation des tweets à diverses échelles, indiquant ainsi les structures spatiales de la contribution à l'activité tweeting sans plus d'approfondissement (Croitoru *et al*., 2017). La figure 3.18 correspond à la carte d'Eric Fischer "*Six Billion Tweets*[44]" qui propose d'afficher les tweets géolocalisés émis dans le monde entier et collectés via l'*API Streaming* de Twitter et ce, depuis 2011 : la carte (a) présente la vue du planisphère ; la carte (b) représente la vue globale de l'aire métropolitaine de Seattle ; la carte (c) zoome sur le centre de la ville (*Downtown Seattle*).

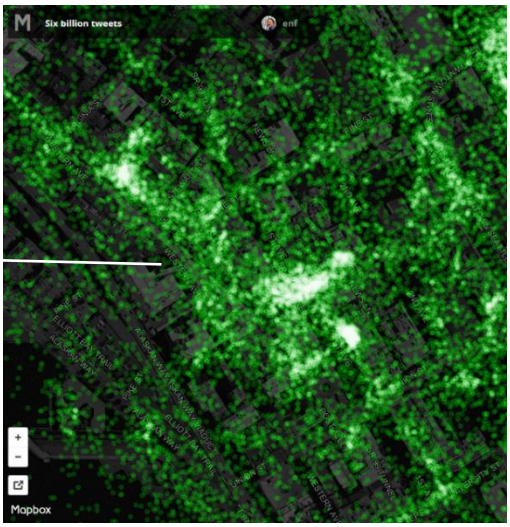
[44] Source : https://blog.mapbox.com/making-the-most-detailed-tweet-map-ever-b54da237c5ac (Consulté pour la dernière fois le 03/10/2019)



(a). Vue du planisphère



(b). Aire métropolitaine de Seattle



(c). Downtown Seattle

Figure 3.18 : Cartographie des tweets "Six Billion Tweets" d'Eric Fischer (Mapbox)

Au-delà de dégager les formes et structures de l'activité virtuelle à diverses échelles (foyers de peuplement à l'échelle du globe, à l'exception des Etats bloquant l'accès à la plateforme, maillage des villes et des réseaux de communication et enfin, à l'échelle de la métropole, identification des lieux et objets du territoire qui polarisent l'activité tweeting[45]), la carte n'offre pas plus de possibilité d'exploration des paramètres de cette activité virtuelle[46].

[45] En revanche, si l'utilisateur de la carte n'a pas connaissance du territoire parcouru virtuellement, le choix du fond de carte rend l'identification des objets polarisant l'activité tweeting (amas de tweets vert clair sur la carte [c]) quasi impossible.

[46] Notons que pour Eric Fischer, le véritable défi posé par la réalisation de cette cartographie interactive était avant tout technique (filtrage des tweets dupliqués, affichage de l'ensemble des points en fonction des niveaux de zoom).

Aux côtés de ces cartographies interactives du Géoweb, les interfaces de géovisualisation scientifiques fondées sur les traces numériques géolocalisées s'articulent autour de trois enjeux découlant des propriétés mêmes de ces traces : on peut visualiser et explorer leur contenu en fonction des structures réticulaires sous-jacentes aux traces (le principe de fonctionnement des médias sociaux étant l'interaction entre utilisateurs), de leur dynamique spatio-temporelle et de leur contenu sémantique (Croitoru *et al.*, 2017). La recherche de structures réticulaires s'ancre sur la construction de graphiques réseaux permettant de visualiser les interactions entre les utilisateurs (Croitoru *et al.*, 2017) : cette construction s'appuie sur des algorithmes de dessin basé sur les forces qui consistent à identifier des nœuds (*nodes*) et à modéliser les liens tissés entre ces nœuds par des arcs (*edges*). C'est ce que montre la figure 3.19 (Croitoru *et al.*, 2017) : elle représente les échanges de tweets géolocalisés entre utilisateurs, consécutifs à l'attentat survenu lors du marathon de Boston en 2013. Les points rouges correspondent aux nœuds du graphe réseau et représentent des clusters d'utilisateurs[47] ; les lignes bleues correspondent aux arcs du graphe et indiquent les échanges de tweets entre les nœuds d'utilisateur.

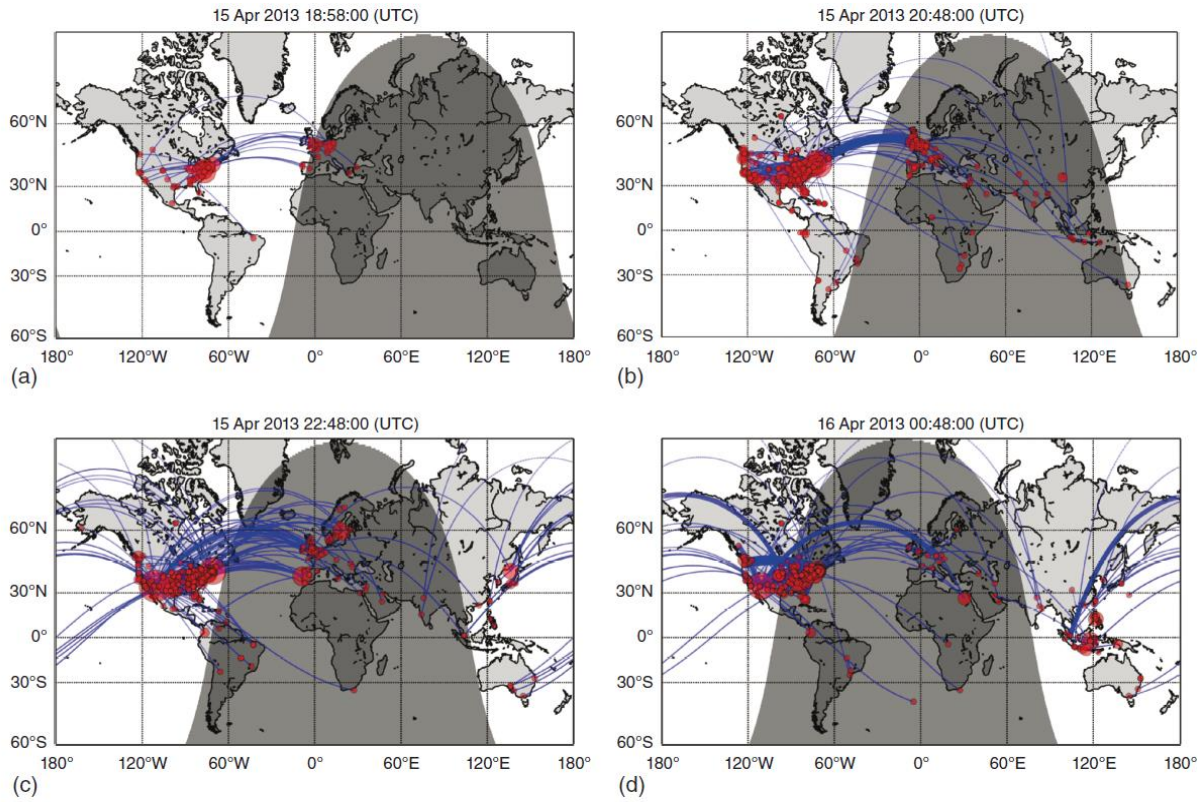
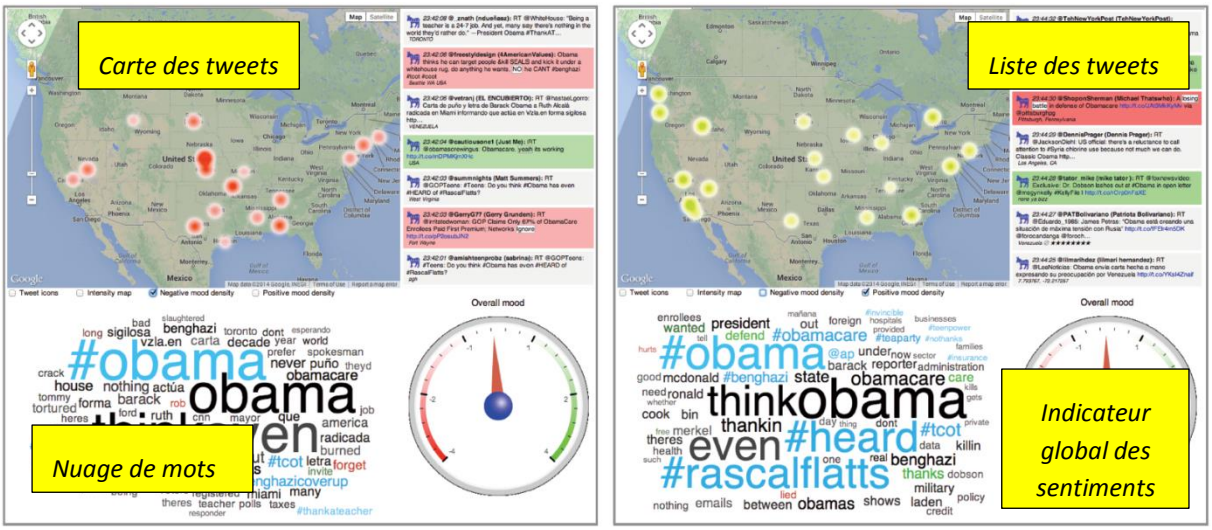


Figure 3.19 : Evolution des échanges de tweets au cours du temps entre les groupe d'utilisateurs, en réponse à l'attentat de Boston en 2013 (Croitoru et al., 2017)

[47] Les auteurs ne précisent cependant pas les modalités de construction des clusters d'utilisateurs pour former les nœuds du graphe réseau.

A l'origine, cette cartographie (figure 3.19) est conçue comme une animation : elle met en évidence le passage, au cours du temps, d'un événement virtuel dont les flux majoritaires restent locaux, à un événement virtuel mondial dans les deux heures suivant les attaques terroristes.

Les interfaces de géovisualisation permettant l'exploration d'un contenu lexical associé à un espace géographique s'avèrent les plus représentées dans la littérature (Nguyen et Schumann, 2010 ; De Chiara *et al*., 2012 ; Thom *et al*., 2012 ; Andrienko *et al*., 2013 ; Croituro *et al*., 2017). (Croituro *et al*., 2017) proposent une interface associant cartographie et contenu sémantique des tweets émis pendant la période des élections présidentielles de 2012 des Etats-Unis et mentionnant le nom du Président Obama. Les traitements sémantiques se distinguent en deux temps : une analyse des sentiments classe les tweets en fonction de leur tonalité positive ou négative et un nuage de mots affiche les termes les plus fréquents dans les deux corpus de tweets constitués (le corpus positif et le corpus négatif). La figure 3.20 présente la forme de l'interface construite :



(a). Tweets négatifs *(b). Tweets positifs*

Figure 3.20 : Géovisualisation des tweets contenant le nom du Président Obama en fonction des résultats de la l'analyse des sentiments (Croitoru et al., 2017)

Chaque visualisation contient la carte de chaleur des tweets (les tweets à sentiment négatif sont représentés en rouge [a] et les tweets à sentiment positif sont représentés en vert [b]), le nuage de mots (les 50 mots les plus fréquents) construit à partir des tweets respectivement étiquetés comme négatifs ou positifs, un échantillon de l'ensemble des tweets inclus dans le jeu global (qui respecte le code des couleurs indiqué ci-dessus[48]), ainsi qu'un indicateur global du score assigné aux tweets par l'analyse de sentiments[49]. Si les auteurs

[48] NB : si le tweet n'a aucune couleur de surlignage, il est classé comme neutre.
[49] Les auteurs ne précisent cependant pas comment la calcul du score *Overall mood* est paramétré.

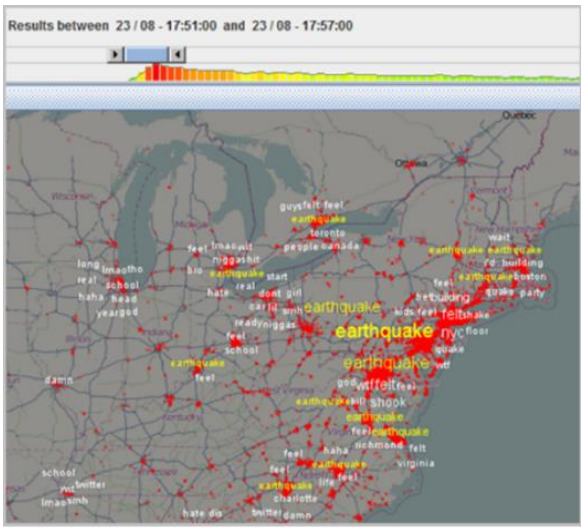
indiquent que la cartographie des tweets en fonction de leur sentiment positif ou négatif oppose les métropoles situées dans des Etats votant plutôt pour le parti démocrate (côte Ouest) aux métropoles situées dans des Etats plus favorables au parti républicain (Etats du Sud), on peut en revanche émettre quelques réserves sur l'efficacité des nuages de mots qui, au final, manquent d'éléments de contextualisation :

- on ne sait pas si les auteurs ont défini un code couleur pour représenter les différents mots en fonction de leur sens péjoratif ou mélioratif : si l'on regarde les deux nuages de mots, on peut trouver, figurant en rouge, les mots plutôt négatifs *lied*, *hurts*, *rob* mais des mots comme *bad*, *tortured* et *slaughered* ne figurent pas dans cette même couleur en dépit de leur sens tout aussi négatif ;
- le mot *Obamacare*, en référence à la mise en place d'une assurance maladie universelle, figure dans les deux nuages de mots mais on ne sait pas comment les tweets positifs et négatifs l'évoquent (autrement dit, les mots positifs ou négatifs auxquels se trouve associée l'*Obamacare*).

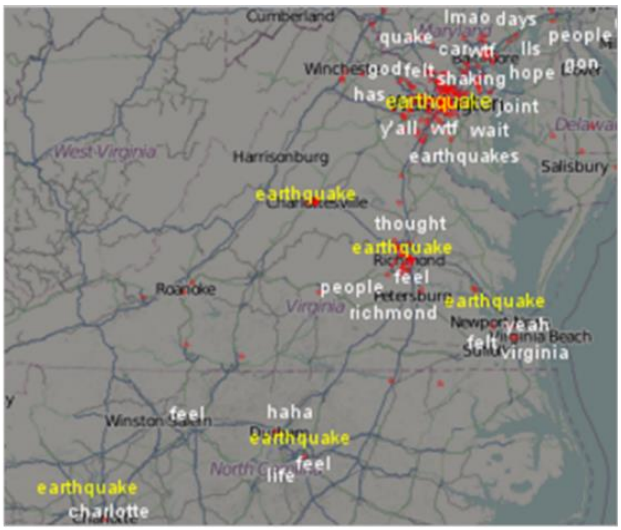
D'autres formes de géovisualisation intègrent l'analyse spatio-temporelle à la cartographie et à la représentation sémantique du contenu des tweets. Etant donné les propriétés volatiles des tweets géolocalisés, l'intégration de la dimension temporelle à la cartographie et à l'analyse sémantique est essentielle pour observer la variabilité spatio-temporelle de la distribution des utilisateurs dans l'espace réel et de l'occurrence de thématiques précises (Croituro *et al*., 2017). Certaines interfaces d'exploration intègrent ainsi l'ensemble des composantes des tweets géolocalisés : (Andrienko *et al*., 2013) ont développé une première interface associant cartographie, outils de filtrage sémantique, curseur de sélection temporelle des tweets ainsi qu'un dernier outil, sous forme de *loupe*, permettant de naviguer et de sélectionner des tweets sur la carte : le contenu des tweets intersectant la loupe figure sous forme d'un nuage de mots. L'analyse combinée des trois composantes est ainsi présentée comme une fenêtre de recontextualisation des messages émis dans un lieu et pendant une période précise (Andrienko *et al*., 2013) ; ces auteurs mettaient ainsi en évidence l'existence, au-delà de l'information routinière et des thèmes récurrents dont l'apparition varie en fonction des jours/heures (transports, travail, loirsirs, etc.), d'événements locaux inhabituels se déroulant dans un lieu particulier à un moment donné (par exemple, l'existence d'un festival musical et artistique dans un quartier de la ville, pendant une courte période incluse dans la collecte, ou encore la présence d'un *spammer* localisé en un lieu unique mais tweetant de manière régulière afin d'assurer la promotion de ses créations musicales).

L'environnement *Tagmap* (figure 3.21) proposé par (Thom *et al*., 2012) fonctionne en suivant le même principe, intégrant espace, temps et sémantique, mais en étant focalisé sur les anomalies de tweeting consécutives à la survenue d'une perturbation sociale ou d'un phénomène naturel. L'interface est constituée d'une fenêtre catographique représentant les

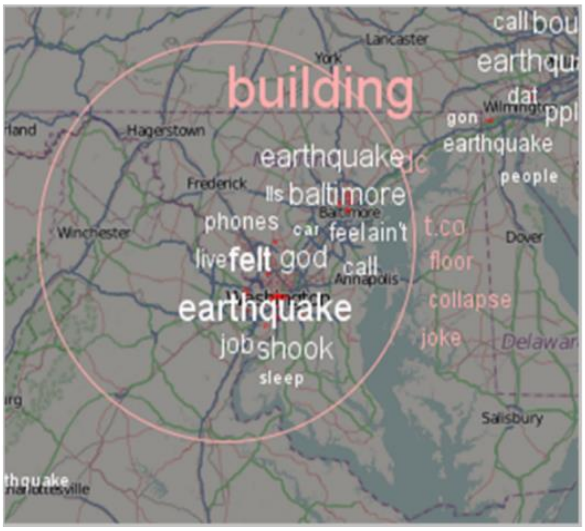
tweets clustérisés[50] en rouge, d'un curseur temporel dont la résolution est définie par l'utilisateur ainsi que d'un nuage de mots représentant les cinq à dix mots les plus rencontrés dans les clusters calculés. Curseur temporel et carte se trouvent associés à un graphique affichant les flux des émissions de tweets, correspondant à la résolution temporelle sélectionnée par l'utilisateur. L'exemple donné dans la figure 3.21 s'applique au séisme survenu sur la côte Est des Etats-Unis le 23 août 2011 (magnitude de 5,8).



(a). Tweets émis en réponse au séisme dans les six premières minutes



(b). Tweets émis en réponse au séisme dans la première minute



(c). Loupe d'exploration sémantique à une échelle plus fine (métropole)

Figure 3.21 : Environnement Tagmap pour l'exploration de la réponse virtuelle consécutive à la survenue d'un séisme sur la côte Est des Etats-Unis (Thom et al., 2012)

[50] Les tweets géolocalisés s'agrègent ets désagrègent en fonction des niveaux de zoom/dézoom.

Vis-à-vis de ce phénomène naturel rare et imprévu, les auteurs (Thom *et al*., 2012) résument le potentiel de l'interface comme suit :

- la visualisation de l'ensemble des composantes des tweets géolocalisés (a), dans les premières minutes suivant la survenue du séisme, fournit l'aperçu de l'ensemble de l'espace au sein duquel la secousse est ressentie ; mais surtout, le graphique de la distribution temporelle des tweets affiche le profil typique de la réponse virtuelle à un phénomène soudain et imprévu, c'est-à-dire une augmentation soudaine et brutale des émissions, suivie d'une lente décroissance ;

- si l'on place le curseur temporel dans la première minute témoignant de cette augmentation brutale des tweets contenant le mot-clé *earthquake*, on peut estimer la localisation de l'épicentre du séisme ou au moins les premières zones habitées dans lesquelles la secousse a été ressentie (b) ;

- à un niveau de zoom élevé, l'interface intègre le même outil de loupe qui permet d'afficher le contenu lexical des tweets l'intersectant (c) : dans le cas de ce phénomène, il apparaissait, au niveau local, que les tweets fournissaient de bons indicateurs de ressenti (en l'occurrence, les utilisateurs témoignaient davantage de leur réaction émotionnelle au moment de la secousse que de dommages ou dégâts).

Enfin, l'environnement *SensePlace3* (Pezanowski *et al*., 2017), qui inclut des fonctionnalités plus complètes, est dédié à l'analyse des tweets émis pendant les périodes de crise d'origine naturelle. L'interface complète se présente sous la forme suivante (figure 3.22) et intègre un premier volet incluant des fonctionnalités de filtrage et une liste des tweets répondant aux critères de sélection énoncés par l'utilisateur (volet de gauche), une fenêtre cartographique localisant les tweets sélectionnés ainsi que des outils d'analyse sémantique.

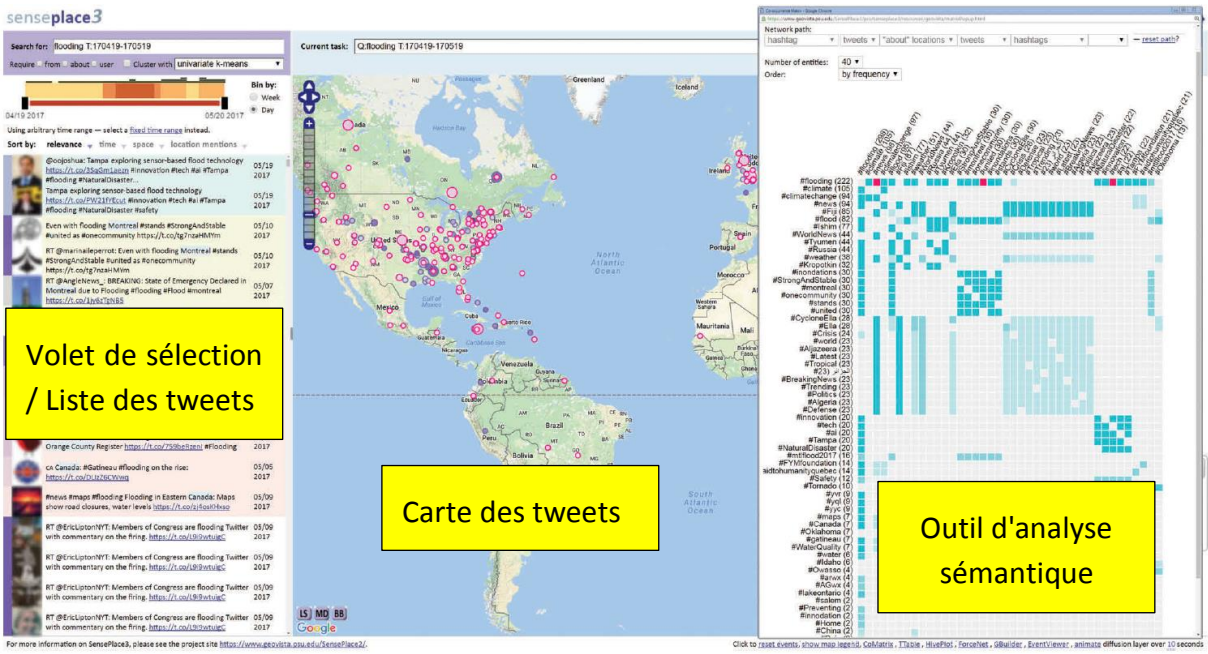


Figure 3.22 : Recherche et exploration lexicale sur SensePlace3 des tweets contenant le mot flooding, émis entre le 19/04/2017 et le 19/05/2017 (Pezanovski et al., 2017)

Dans un premier temps, l'utilisateur doit soumettre une requête, par le volet de gauche, dans laquelle il indique ses critères de sélection des tweets collectés par l'*API Streaming* : cette sélection est sémantique et temporelle. Dans la figure 3.23 ci-après, l'utilisateur requiert des tweets contenant le mot-clé *flooding* et émis entre le 19 avril 2017 et le 20 mai 2017, sans indiquer de pays particulier d'origine (case *from* non cochée). Les cases colorées (dégradé de couleurs chaudes) figurant au-dessus du curseur temporel offrent un aperçu des effectifs de tweets contenant le mot-clé recherché en fonction des jours (plus la couleur est foncée, plus l'effectif de tweets est conséquent). La liste des tweets retournés apparaît au-dessous de cette fenêtre de requête.

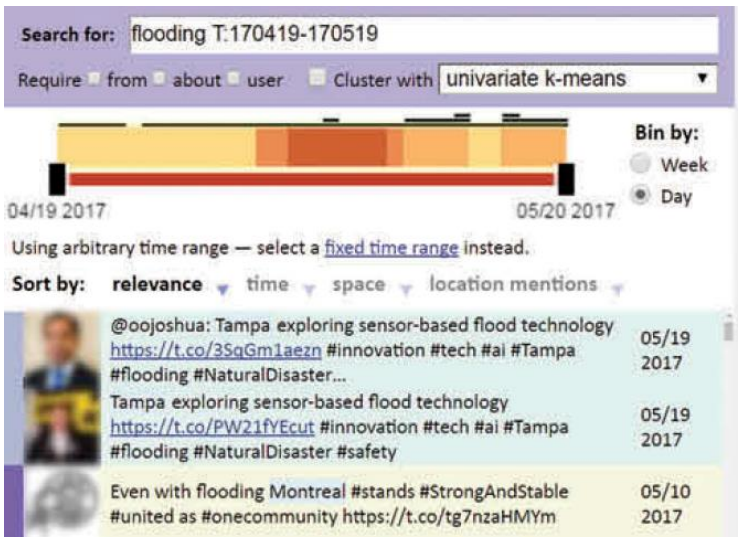
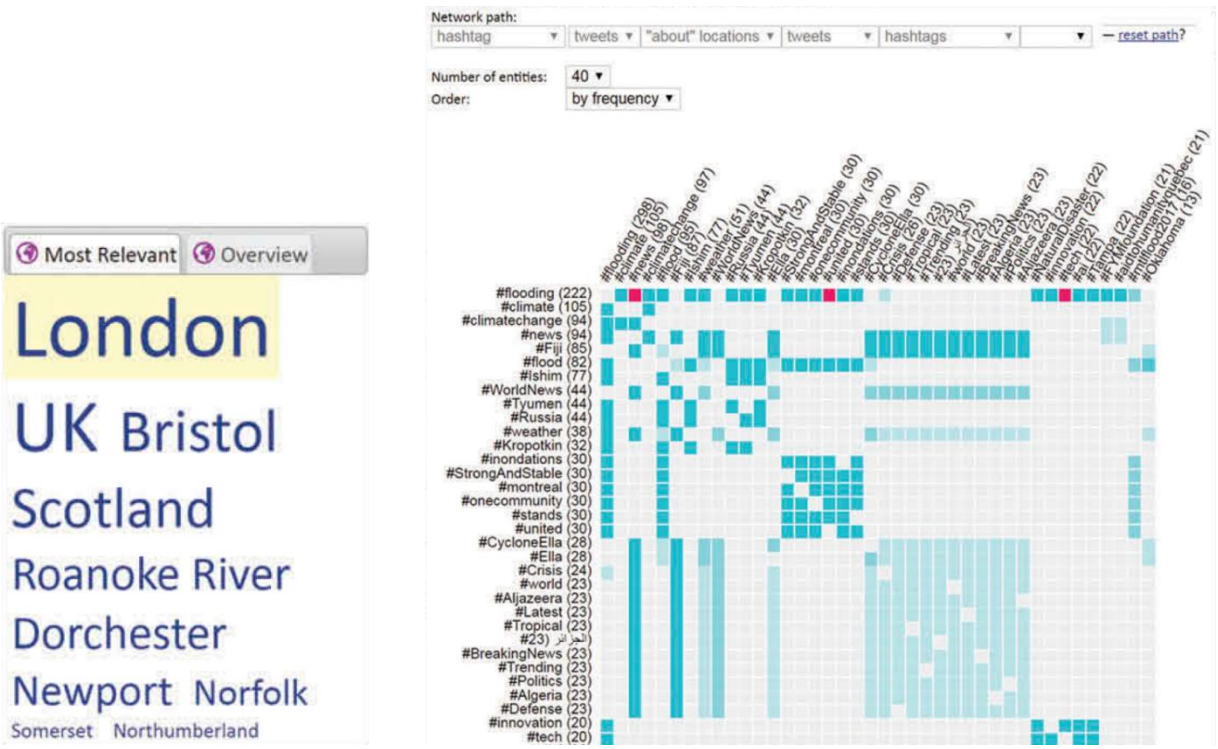


Figure 3.23 : Requêtes de sélection des tweets par SensePlace3 (Pezanovski et al., 2017)

Les tweets géolocalisés retournés par la requête de l'utilisateur figurent sur la carte sous forme de clusters. Dans un second temps, la sémantique de ces tweets peut être explorée par des outils d'analyse lexicale. La figure 3.24 présente deux types de fonctionnalités lexicales : la partie gauche de la figure (a) soumet un nuage de mots constitué à partir d'une requête utilisateur sélectionnant les tweets géolocalisés émis au Royaume-Uni entre le 22 avril et le 17 mai 2017 et contenant le mot-clé *flood*. Le nuage de mots généré ne représente pas l'ensemble des mots-clés les plus fréquents contenus dans les tweets sélectionnés mais se trouve en fait exclusivement focalisé sur les lieux mentionnés, en lettres, dans les tweets (l'interface contient en effet un outil de *geoparsing*[51] permettant d'identifier les noms de lieux ou d'objets localisables inclus dans les tweets, et à diverses échelles spatiales, de la région à la rivière). La partie droite (b) de la figure affiche une matrice de cooccurrences entre les hashtags présents dans les tweets filtrés par la requête exprimée dans les figures 3.22 et 3.23 (pour rappel, il s'agit des tweets contenant le mot-clé *flooding* et émis entre le 19 avril 2017

[51] Le *geoparsing* est une méthode de construction d'index de localisation, fondée sur le recours à des algorithmes de reconnaissance automatique d'entités nommées dans des chaînes de caractères.

et le 20 mai 2017) : cette matrice permet de visualiser les hahstags ayant tendance à être associés dans la composante sémantique des tweets : ici, plus la couleur est foncée, plus la tendance d'association entre deux hashtags est forte (et le rouge est utilisé pour distinguer les associations les plus fréquentes, comme *#flooding* et *#news*). Dans le cas précis de la figure 3.24, la construction de cette matrice de cooccurrences est paramétrée comme suit : seuls les tweets contenant une information lexicale mentionnant un toponyme sont pris en compte (paramètre *"about" locations*) et seuls les 40 hashtags les plus fréquents apparaissent dans la matrice de cooccurrences.



(a). Nuage de mots à partir des entités nommées

(b). Matrices des cooccurrences lexicales entre les hashtags contenus dans les tweets

Figure 3.24 : Deux exemples de fonctionnalités d'analyse lexicale intégrées à l'environnement SensePlace3

L'avantage indéniable de cette interface est qu'elle s'avère la première à prendre en compte l'ensemble des dimensions spatiales susceptibles d'être intégrées dans un tweet : la géolocalisation directe par GPS, les informations de lieux ajoutées dans le profil de l'utilisateur ainsi que les mentions lexicales de lieux dans le tweet. Par la combinaison carte et matrice de cooccurrences hashtag/lieux, on peut alors rapidement identifier si des utilisateurs évoquent des phénomènes locaux auquels ils assistent ou des phénomènes distants pour lesquels ils se positionnent comme des relais de l'information. En outre, par la matrice, on peut également afficher des tendances de tweeting relatives à un lieu précis, et par conséquent, identifier les territoires dans lesquels des phénomènes sont en cours (par exemple : *#flooding #Oklahoma*, *#Russia #flood*).

Que peuvent apporter ces interfaces de géovisualisation à l'approfondissement de l'analyse et des connaissances acquises en ce qui concerne les comportements virtuels spatiaux ? Comme nous l'avions d'ores-et-déjà indiqué dans la section 3.3.3, les fonctionnalités de ces interfaces restent fondées sur des méthodes d'analyse de données courantes et non spécifiques aux traces numériques issues du Web social (clustérisation, carroyage, nuages de mots, etc.). L'environnement *SensePlace3*, dans sa version la plus récente, se démarque néanmoins par ses deux fonctionnalités novatrices :

- la représentation des liens entre les mots (et non plus chaque mot représenté indépendamment de ses cooccurrences lexicales) ;
- la prise en compte de la déclinaison de la composante spatiale des tweets géolocalisés avec la possibilité d'évaluer les effets de téléprésence (par exemple, si l'on identifie des tweets géolocalisés dans la capitale des Etats-Unis, Washington, qui contiennent le mot *flooding*, mais que celui-ci se trouve fréquemment associé au hashtag *#Oklahoma*, alors on se trouve très vraisemblablement face à un effet de téléprésence et non face à un phénomène d'inondation localisée).

L'avantage principal de ces interfaces reste qu'elles ont la possibilité d'intégrer, dans une unique vue, l'ensemble des outils mobilisés pour l'analyse globale des tweets : cartographie, flux temporels et analyse sémantique. Les étapes d'exploration sont donc associées dans un même environnement et non plus séparées sur des logiciels différents. Conception et fonctionnement des environnements de géovisualisation articulés aux traces numériques géolocalisées s'inscrivent donc dans le cadre théorique et méthodologique de l'analyse visuelle : ils consistent en effet à représenter graphiquement et de manière synchronisée les différents attributs du twees dans une interface interactive et ce, afin de favoriser une immersion progressive dans les données (MacEachren et Kraak, 2001). Par ailleurs, ce cadre théorique et méthodologique s'enracine lui-même dans les postures épistémologiques présentées dans la section 3.3.3 de cette partie, et notamment dans l'association entre abduction (construire une hypothèse à partir de la juxtaposition d'observations établies sur les données) et induction (répéter les expérimentations afin de valider les conditions dans lesquelles l'hypothèse posée se vérifie). Si l'on reprend les quatre étapes de l'approche de l'analyse visuelle proposée par (Yi *et al.,* 2008) et si l'on considère simultanément une approche épistémologique ancrée sur les faits qui émergent progressivement de l'analyse des données, la conception d'un environnement de géovisualisation, construit dans l'objectif d'explorer les traces numériques géolocalisées générées en réponse à un phénomène d'origine naturelle, devrait répondre à une logique d'exploration des données respectant l'ordre des étapes suivantes :

- l'aperçu général est donné par des fonctionnalités de primo-exploration des traces (clusters, cartes de chaleur, graphiques temporels représentant les flux de tweets, nuages de mots, soit l'ensemble des représentations les plus courantes). L'objectif consiste à répondre à

la question suivante : où peut-on identifier la survenue d'un phénomène donné dans des temporalités précises ?

- l'analyse s'ajuste ensuite sur les traces incluses dans le premier lieu d'intérêt retenu (changement d'échelle) identifié par l'aperçu général de la distibution des traces ;

- la recherche de structures spatiales et de tendances s'appuie conjointement sur des outils d'analyse (spatiale, statistique, sémantique) plus approfondis et sur les facultés cognitives de l'utilisateur (en d'autres termes, ses capacités à observer les tendances qui se dégagent de l'analyse des traces et à rechercher des facteurs explicatifs de ces tendances). Dans notre cas, cette recherche de facteurs explicatifs peut s'appuyer sur le test de variables environnementales, sociales et spatiales qui influenceraient la distribution des tweets géolocalisés émis en réponse au phénomène étudié.

- En dernier lieu, il est nécessaire de construire un modèle d'analyse exploratoire permettant de répéter l'ensemble des étapes appliquées au premier lieu d'intérêt dans un nouveau lieu. L'objectif consiste à vérifier, dans une logique inductive, si les structures mises en évidence dans le premier lieu exploré s'appliquent à d'autres lieux d'intérêt, et ce, afin de valider la justesse des hypothèses posées.

Conclusion du chapitre 3

Quel bilan des connaissances géographiques acquises sur les tweets géolocalisés pouvons-nous alors retenir ? Peut-on déjà envisager l'existence de lois encadrant les logiques comportementales de l'activité virtuelle géolocalisée ?

On peut en effet souligner, dans l'état de la question, un certain nombre de faits constatés, dans le comportement des tweets, de manière répétée, observables dans différents contextes sociaux, spatiaux et phénoménologiques. Dans un premier temps, la distribution spatiale des tweets géolocalisés révèle un contraste entre des lieux concentrant de fortes densités de tweets géolocalisés alors que d'autres lieux ne contiennent que des traces éparses ; ils forment ainsi un maillage irrégulier sur le territoire. Cette variabilité permet d'identifier les pulsations du territoire, c'est-à-dire les hauts-lieux de l'activité virtuelle (quelle que soit l'échelle géographique considérée, du pays aux objets précis dans une ville qui polarisent l'activité virtuelle), ainsi que leur variabilité temporelle qui peut s'expliquer par divers facteurs : la présence d'une activité touristique, la présence d'objets localisés ou d'événements prévus générateurs d'une activité virtuelle, ou encore de la survenue d'un événement imprévu perturbateur de la routine.

Dans un deuxième temps, le discours et les flux temporels d'émission de tweets témoignent de l'existence irréfutable d'une réactivité systématique aux perturbations, qu'elles soient d'origine sociale ou environnementale. Si l'on considère les émissions de tweets en rapport avec l'une de ces perturbations, celle-ci se détecte rapidement par un effet déluge de tweets qui se manifeste par la survenue d'un pic de tweets dans les premières minutes suivant l'apparition de la perturbation. Par ailleurs, la composante sémantique des tweets peut également apparaître comme un indicateur de réactivité à l'environnement : dans les risques naturels, ce comportement se traduit par la capacité des tweets géolocalisés à retranscrire les différentes phases d'un phénomène naturel et de ses répercussions sur les populations affectées.

Enfin, conséquence des deux premiers constats répétés, le contenu sémantique et les dynamiques d'émission des tweets n'apparaissent pas aléatoires : les concentrations de tweets en des lieux particuliers ne sont pas la conséquence d'un hasard mais plutôt de la présence d'un objet ou d'un événement éphémère générateur d'activité virtuelle, de même que dans le cas des phénomènes naturels dommageables, on peut constater l'existence d'une correspondance locale entre lieux affectés et localisation et contenu des tweets émis en réponse au phénomène. Un pan de la recherche tente ainsi de modéliser les comportements de l'activité virtuelle par des tests statistiques témoignant de l'existence de certaines corrélations entre les tweets et des variables sociales ou environnementales.

En dépit de ces constats positifs vis-à-vis du potentiel géographique associé aux tweets géolocalisés, il nous faut inventorier certaines mises en garde qui découlent de la nature même du matériau tweet géolocalisé ; elles sont inventoriées dans le tableau 3.2 ci-dessous, qui soumet les pratiques méthodologiques présentées dans le chapitre ainsi que les doutes et manques qui leur sont associés :

Tableau 3.2 : Bilan des pratiques associées à l'exploration des tweets géolocalisés et des questions non résolues liées à ces pratiques

	Pratiques	Questions en suspens
Tests statistiques	Démonstration de l'existence de corrélations entre activité virtuelle géolocalisée et variables sociales ou facteurs environnementaux.	Que vaut la représentativité de ces résultats vis-à-vis de l'ensemble des individus présents sur le territoire mais inactifs sur le réseau? Peut-on généraliser les résultats à d'autres populations d'utilisateurs géolocalisés dans d'autres lieux ?
Analyse sémantique	Vérification de l'existence d'une sensibilité à des thèmes précis, qui contrate avec l'activité routinière. Restranscription lexicale du déroulement d'une perturbation.	Approfondir l'analyse au-delà de la simple représentation des mots-clés majoritaires en nuages de mots ou de la discrétisation des tweets par les algorithmes d'analyse de sentiments.
Analyse spatiale	Utilisation des méthodes traditionnelles dans les SIG : zones tampon, analyse des distances, clusters, autocorrélation spatiale, cartes de chaleur, etc.	Ces méthodes sont-elles adaptées face à l'hétérogénéité et au caractère volatil des traces numériques géolocalisées ?
Géovisualisation	Regroupement de fenêtres d'analyse et de visualisation des différentes composantes des tweets dans un unique environnement, facilitant l'efficacité de l'exploration.	Quels outils d'analyse fournissent des résultats probants pour être intégrés dans un environnement garantissant une exploration efficace des tweets ?

Face à ces difficultés énumérées dans le chapitre (et résumées dans le tableau 3.2), c'est au final le renouvellement des postures épistémologiques ainsi que la démarche exploratoire théorique adoptée par l'analyse visuelle qui nous paraissent comme les paramètres les plus pertinents pour approfondir les connaissances des événements virtuels créés en réponse aux phénomènes naturels générateurs de dangers. En effet, comme nous l'avons indiqué précédemment, devant la variabilité de l'engagement des populations à la création de contenus virtuels géolocalisés, on ne peut pas prédire du potentiel d'un jeu de tweets, ni des méthodes et pistes de recherche qui fourniront des résultats exploitables ou conduiront à des impasses ; de même que nous ne savons pas par avance, dans les phénomènes à l'étude de cette recherche, si les résultats constatés par les publications citées dans ce chapitre seront observables dans notre cas, ou révèleront des comportements de l'activité virtuelle diamétralement opposés.

Partie I : De l'unicité à la diversité des formes des données géographiques et de leurs usages - Conclusion

L'intégration des outils du Web social à la vie quotidienne, conjuguée à l'adoption massive des smartphones, marque l'amorce d'une nouvelle dynamique dans la production de contenus web dont les propriétés de localisation qui leur sont souvent associées, apparentent ces contenus à des données géographiques. Ainsi, les conditions de production et d'utilisation du matériau traditionnel du géographe cartographe lui échappe : technologies 2.0 et nouvelles pratiques contributives du Web social entraînent un foisonnement des contenus géolocalisés de formes variées et de leurs cartographies. En conséquence, la nouvelle cartographie du Web social est généralement réalisée par des non géographes et entraîne une rupture avec les pratiques professionnelles standardisées des héritages séculaires de la discipline. En revanche, les cartes du Web construites à partir des données du Web ne s'incrivent pas forcément dans une démarche analytique ou exploratoire : elles se contentent généralement d'afficher les contenus bruts créés sur le Web ou tentent de le valoriser, parfois maladroitement (puisqu'en dehors des règles de représentation scientifique des données géographiques). Peut-on alors valoriser les contenus géolocalisés du Web en information géographique ?

Cette recherche s'articulant autour du potentiel géographique et cartographique d'une forme de données du Web, les tweets géolocalisés, nous proposons d'appréhender ce matériau, parmi la mosaïque des nouvelles données du Web social, selon les critères suivants :

- le tweet géolocalisé est une *trace numérique* et non une donnée géographique : il est en effet produit de manière spontanée pour faire parler de soi sur un réseau virtuel dont la production de connaissances géographiques n'est pas la première finalité, ou pour partager de l'information acquise sur le terrain (ou encore lors de la navigation web) auprès des utilisateurs de la plateforme. L'exploitation scientifique des traces, contrairement aux données géographiques officielles, ne peut pas être envisagée sans soulever une certain nombre de questions quant à leurs propriétés ;
- le tweet géolocalisé est une trace numérique *contributive*, dans la mesure où la dimension géographique précise (localisation par GPS ou par *bounding box*) ne peut pas être acquise au détriment de l'utilisateur. La question qui reste à soulever vis-à-vis de ce choix d'ajout d'une composante spatiale à un tweet reste la suivante : existe-t-il une intention géographique derrière le tweet géolocalisé ? Autrement dit, pourquoi un utilisateur choisit-il de géolocaliser ses tweets et a-t-il conscience de la réutilisation de ses traces par des acteurs tiers ?
- Conséquence du point précédent, nous considérons que la plupart des tweets géolocalisés ne sont pas créés dans une posture volontaire (à la différence de la VGI) : l'activité tweeting ne s'inscrit pas dans une logique de construction et de diffusion d'information utile,

structurée et réutilisable (même si le réseau peut parfois être utilisé par des communautés mettant en place des codes précis encadrant la production d'une information directement exploitable, comme c'est le cas avec le hashtag *#MSGU*) ;

- le tweet géolocalisé est une trace biaisée : l'activité tweeting géolocalisée, et plus particulièrement en situation de mobilité, ne peut être produite que si l'utilisateur dispose d'un smartphone et d'un accès à l'Internet mobile ; or, tous les individus et tous les territoires ne disposent pas des mêmes conditions d'accès à ces deux paramètres. La production de tweets géolocalisés est liée à l'activité d'individus dont il est encore difficile de caractériser le profil socio-démographique et les motivations parmi l'ensemble des utilisateurs de Twitter. En conséquence, même si l'on parvient à produire une connaissance géographique des interactions entre individus et territoires par le biais de l'activité virtuelle, nous n'avons aucune garantie quant à sa généralisation à l'ensemble des individus parcourant le même territoire mais invisibles sur le réseau.

En dépit de ses défauts, nous considérons le tweet géolocalisé comme un capteur potentiel de l'interaction entre le lieu (ou un objet du lieu) et l'individu qui survient dans le contexte précis du territoire en crise, pour lequel le chercheur ne peut pas entrevoir d'autres sources de données capturées sur le terrain et en temps réel. Ce potentiel de capture des interactions humain/lieu a d'ores-et-déjà été mis en valeur dans diverses problématiques et valorise le tweet géolocalisé comme source alternative pour la détection et la caractérisation de comportements spatiaux et de micro-événements non captés par les jeux de données officielles. En effet, les pistes de recherche explorées ont déjà modélisé l'existence de corrélations entre des facteurs sociaux ou environnementaux et des comportements individuels ou collectifs enregistrés sur le réseau ; il reste néanmoins que la recherche peine à approfondir les connaissances acquises et que, d'après les remarques de certains auteurs (Kitchin, 2013 ; Miller et Goodchild, 2015 ; Steiger *et al*., 2015), les travaux entrepris n'ont pour l'instant dégagé que la partie superficielle de la géographie du réseau social géolocalisé. Le problème alors avancé se résume au défaut d'outils d'analyse et de méthodes, pour les géographes, adaptés au traitement de jeux de données géolocalisées peu volumineuses et arborant les propriétés des *Big Data.*

En conséquence, la solution que nous envisageons d'adopter ici correspond à l'approche ancrée dans l'observation des faits émergeant progressivement des étapes de traitement des traces croisées à d'autres jeux de données. Le cadre épistémologique associe alors l'abduction (la génération progressive des hypothèses afin d'identifier des facteurs explicatifs des observations effectuées à partir d'un traitement) et l'induction (la vérification de la validité d'une hypothèse posée dans des cas différents de celui qui a permis de la soumettre). D'autre part, les étapes de l'analyse des tweets géolocalisés se conformeront au cadre théorique de mise en œuvre de l'analyse visuelle : aperçu global des composantes des tweets géolocalisés, sélection d'un lieu d'intérêt, phase de traitement et de représentation, interrogation de la validité des résultats obtenus et critique des méthodes employées pour l'exploitation et la représentation des traces numériques géolocalisées.

Partie II

Contributions à l'extraction et au traitement des tweets de crise géolocalisés

Partie II : Contributions à l'extraction et au traitement des tweets de crise géolocalisés

Après avoir présenté, dans la partie précédente, les propriétés des traces numériques que sont les tweets géolocalisés, les verrous sous-jacents à leur analyse et à leur représentation cartographique, ainsi que le bilan des connaissances acquises et des approches méthodologiques utilisées ou préconisées, la seconde partie du manuscrit est centrée sur l'exposition des contributions proposées pour l'approfondissement de la question relative aux apports des tweets géolocalisés dans le contexte particulier des risques et catastrophes naturels. Elle s'articule ainsi autour de la présentation des contributions sur deux niveaux :

- pour la contribution méthodologique, la définition de procédures d'extraction de tweets en réponse à une problématique particulière et de leur traitement ;
- pour la contribution thématique et en réponse au cadre méthodologique posé, il s'agit de mesurer le potentiel de construction de connaissances sur la géographie d'un espace virtuel que peuvent fournir les tweets géolocalisés sur notre terrain d'étude dans le cas précis des phénomènes hydrométéorologiques à l'étude.

Cette seconde partie se subdivise en trois nouveaux chapitres : le chapitre 4 expose et illustre les démarches méthodologiques et le cadre épistémologique adoptés par rapport aux travaux effectués et aux réflexions engagées sur les difficultés éprouvées face au traitement des traces numériques pour les questions géographiques, que nous avons soulignées dans le chapitre 3. Les chapitres 5 et 6 présentent les résultats des méthodologies mises en œuvre pour l'extraction et le traitement des tweets géolocalisés : le chapitre 5 se focalise sur le retour des démarches d'extraction de tweets et les premières explorations effectuées à l'échelle globale des perturbations et du terrain d'étude. Enfin, le chapitre 6 change d'échelle et s'attache à l'exploration des événements virtuels consécutifs aux phénomènes naturels dans le territoire de prédilection du numérique, la métropole.

4. Proposition d'une démarche méthodologique pour l'extraction et l'analyse de tweets géolocalisés

Ce chapitre présente les réflexions thématiques et les méthodologiques engagées et les démarches appliquées après le premier cadrage spécifique à l'utilisation académique des tweets géolocalisés, présenté dans le chapitre précédent. D'une manière générale, l'ensemble des travaux consultés s'entendent sur la capacité des tweets à détecter des tendances et des dynamiques spatio-temporelles. Si l'on considère les processus d'extraction et de traitement des tweets, ceux-ci sont focalisés sur deux aspects : l'extraction de tweets utiles à la résolution de la problématique d'une part, et l'analyse des tweets en elle-même, soit automatisée par le recours aux algorithmes d'apprentissage, soit supervisée et accomplie au moyens d'outils d'analyse spatiale qui ne sont pas inédits, ou par des environnements de géovisualisation qui offrent l'avantage de regrouper la représentation graphique de l'ensemble des composantes des tweets au sein d'une même interface.

Dans ce quatrième chapitre, il s'agit tout d'abord de positionner notre recherche par rapport à l'existant, en termes de questionnement géographique ; au commencement de la recherche, une question générale se pose : qu'est-ce-que la géographie d'un réseau virtuel géolocalisé lorsque le territoire réel qui sert de support de capture de traces numériques géolocalisées se trouve en crise ? Sur le plan technique, il s'agit ensuite d'identifier et de tester des méthodes reproductibles d'extraction de tweets liés à cette problématique de crise d'origine naturelle, et d'analyse (statistique, temporelle, spatiale, sémantique) ; l'objectif consiste à valoriser les trois composantes des tweets géolocalisés afin d'apporter des réponses non seulement à la question générale posée ci-avant, mais encore aux interrogations et doutes émis sur le matériau tweet géolocalisé et exposés dans les chapitres 2 et 3 de la première partie[1].

Le chapitre est alors organisé comme suit : le premier axe introduit d'une part, les limites des travaux existants, autrement dit les points qui nous ont interpellés à la lecture de l'état de l'art et d'autre part, notre positionnement thématique et méthodologique concernant les outils d'extraction et d'analyse des tweets. Le deuxième axe présente l'approche et les outils mobilisés dans le cadre de l'extraction de tweets utiles à la problématique. Enfin, le troisième et dernier axe dresse l'inventaire des données officielles complémentaires qui seront associées aux tweets et expose les outils et méthodes d'analyse des tweets et données.

[1] Pour rappel, nous faisons référence à la forte hétérogénéité de la sémantique des tweets, aux biais de leurs condition de capture, de représentativité statistique et spatiale, mais également à la question de l'adaptabilité des outils et méthodes d'analyse spatiale traditionnels aux traces numériques géolocalisées.

4.1. Comment rechercher dans les tweets géolocalisés ?

Cet axe soumet, dans un premier temps, un exposé des points qui suscitent notre attention après la présentation des pistes de recherche géographique explorées dans le chapitre précédent. Dans un deuxième temps, et en réponse à ces questionnements, il retrace les interrogations thématiques et méthodologiques qui se posent dans cette recherche sur l'activité tweeting dans le cadre de la survenue de phénomènes naturels dommageables. Enfin, il décrit la démarche épistémologique mise en œuvre dans la conduite de la recherche, articulée autour des concepts introduits dans la section 3.3.3 du chapitre 3.

4.1.1. Quelles limites peut-on identifier dans la recherche accomplie sur les tweets géolocalisés ?

4.1.1.1. Constitution des jeux de tweets filtrés pour l'étude d'une question précise

Dans un premier temps, si l'on se réfère aux différentes méthodologies exposées dans le chapitre précédent, un certain nombre de points problématiques méritent d'être soulevés. (Steiger *et al*., 2015) indiquaient d'emblée que la première étape concernant l'extraction de tweets utiles à la problématique d'étude se montrait souvent trop rapide, dans la mesure où elle n'était basée que sur une poignée de mots-clés et dans une logique supervisée (le ou les mots-clés de recherche sont déterminés par les chercheurs). C'est en effet ce qu'on peut constater dans la problématique des risques naturels : chez (Sakaki *et al*., 2010), on trouvait les mots-clés *earthquake* et *shake*, sans plus de précisions quant à l'éventuelle prise en compte de mots supplémentaires ; pour (Alam *et al*., 2018), l'extraction de tweets liés aux ouragans consistait à associer diverses combinaisons entre le mot-clé *hurricane* et les noms des cyclones à l'étude ; (Hertfort *et al*., 2014) appliquaient une liste constituée de deux mots-clés trouvés dans la définition de *Hochwasser* (inondation) donnée par le dictionnaire *Duden* (en l'occurrence, les synonymes du mot *Hochwasser* : *Überschwemmung* et *Flut*), auxquels ils ajoutaient trois autres mots-clés : *flood*, *Sandsack* (sac de sable) et *Deich* (digue). Au contraire, (Saravanou *et al*., 2015), en exécutant une analyse grammaticale complète isolant chaque mot dans chaque tweet, dressaient un lexique final et validé manuellement de 456 mots-clés sur les inondations survenues en Angleterre.

Que peut-on alors reprocher aux démarches précédentes ? Tout d'abord, une extraction lexicale fondée sur quelques mots-clés représente le risque d'écarter tout tweet relatif au phénomène, mais qui le décrit dans un vocabulaire en dehors des standards scientifiques. Le VISOV a bien intégré cette nuance : la recherche de tweets mentionnant des phénomènes émergents d'inondation n'est pas uniquement focalisée sur le terme *inondation* ; elle intègre la faute d'orthographe la plus commune *innondation* mais également des termes plus

familiers comme *déborder*, employé pour décrire une rivière en crue[2]. Dans un second temps, les tweets émis en réponse à un phénomène physique ne contiennent pas nécessairement de *hashtag* mentionnant le nom du phénomène : si les cyclones sont baptisés à partir du stade de la tempête tropicale, les séismes ou les phénomènes récurrents de pluies/inondations ne portent pas de noms permettant de les identifier (si ce n'est le nom des lieux affectés). En outre, comme l'ont montré (Sakaki *et al*., 2010) et (Bossu *et al*., 2018), le tweet signalant dans l'immédiat la survenue d'un phénomène est souvent court. En conséquence, si l'on recherche des tweets qui contiennent le seul hashtag *#HurricaneHarvey* ou *#HurricaneIrma*, on risque d'exclure du jeu filtré, tous les tweets rapportant des observations rapides des conditions environnementales locales. Enfin, l'ajout de mots-clés choisis par les chercheurs risque également d'introduire un biais. Les mots-clés donnés par (Hertfort *et al*., 2014) peuvent concerner à la fois une phase d'anticipation et une phase de crise : les sacs de sables sont installés en prévention des inondations, les digues peuvent également être renforcées, ou céder pendant la crue du fleuve. Mais si l'on souhaite intégrer d'autres mots, différents d'*inondation* et de ses synonymes, pourquoi ne pas alors aussi prendre en compte d'autres situations envisageables dans les différentes phases du phénomène ? Par exemple, on pouvait intégrer les idées suivantes : faire ses provisions, fermeture des écoles, coupures d'électricité, personnes bloquées, *etc.*. Dans le cas contraire, le risque immédiat est de réaliser des analyses sur un jeu de tweets extraits selon des critères subjectifs.

D'un autre côté, même si la tâche se montre longue et fastidieuse (Saravanou *et al*., 2015), la constitution d'une liste de 456 mots-clés par examen du contenu sémantique des tweets se positionne comme une approche adéquate et plus approfondie de l'extraction lexicale supervisée. On pourra néanmoins soulever la question suivante (en dehors de la lenteur de la tâche) : étant donné qu'une poignée de mots-clés sont associés à un grand nombre de tweets[3], faut-il nécessairement établir des corpus d'extraction composés de plusieurs centaines de mots-clés ? En réponse à l'ensemble de ces remarques, les méthodologies d'extraction de tweets utiles à la problématique de recherche constituent ainsi le premier axe des travaux effectués dans ce manuscrit.

4.1.1.2. Comment appréhender la distance entre les tweets ?

Dans un deuxième temps, et en ce qui concerne les méthodologies d'analyse des tweets extraits, un certain nombre de points soulèvent également nos interrogations. Le premier point qui pose problème selon nous est celui de la distance et de la proximité. Lorsqu'ils étudiaient l'influence de l'environnement alimentaire de proximité sur le comportement des utilisateurs, (Chen et Yang, 2014) définissaient ce critère comme une distance maximale de 1

[2] Source : https://www.francebleu.fr/infos/societe/le-visov-des-pompiers-volontaires-du-web-pour-aider-les-secours-en-vaucluse-1499937149 (Consulté pour la dernière fois le 05/08/2019)

[3] Le mot *rain* est présent dans 11 235 tweets ; le mot *weather* dans 3 331 tweets ; le mot *snow* dans 1 006 tweets ; le mot *showers* dans 273 tweets et le mot *flooded* dans 274 tweets. Néanmoins, les auteurs de la publication ne précisent pas le nombre total de *flood-related tweets* extraits par le corpus de 456 mots-clés.

mile au tweet (soit 1,6 km). Dans la problématique des risques naturels, (Hertfort *et al*., 2014) annonçaient une correspondance spatiale entre la distance des tweets géolocalisés aux bassins versants inondés et le rapport de ces tweets au phénomène. En outre, après avoir classifié les tweets en fonction de leur contenu sémantique, il apparaissait que les catégories mentionnant des hauteurs d'eau et des actions d'entraide étaient les plus *proches* des bassins versants affectés. Néanmoins, la distance moyenne des tweets géolocalisés au centre de gravité des bassins versants affectés s'avérait de 39 km. Que considère-t-on alors comme *proche* avec les tweets géolocalisés ?

Sur le plan technique, le critère de distance apparaît également dans les algorithmes de création des clusters de tweets ; par exemple, comme indiqué dans le chapitre précédent lors de l'introduction de DBSCAN, l'utilisateur doit indiquer une distance seuil au-delà de laquelle deux points ne sont pas agrégés dans le même cluster. Dans cette perspective, nous avons relevé les interrogations des auteurs sur les critères de détermination de cette distance seuil[4]. Celle-ci revient en effet à poser la question des effets d'échelle géographique : si l'on étudie un phénomène spatial à une échelle mal ajustée, le risque est de manquer l'essentiel du sens de l'information (Steiger *et al*., 2015). Dans le cas des tweets géolocalisés, le cœur du problème se manifeste par le fait que cette distance seuil doit être fixée à partir d'une distribution spatiale irrégulière, dense en certains lieux, lâche voire inexistante en d'autres ; en conséquence, le réseau virtuel géolocalisé ne peut pas s'apparenter à un échantillon spatial mais s'assimilerait plutôt à un maillage irrégulier emplis de biais qui se répercutent sur l'attractivité virtuelle des lieux[5] (Quesnot, 2016). Et c'est sans doute pourquoi (Miller et Goodchild, 2015) posaient l'hypothèse selon laquelle nos méthodes actuelles ne seraient pas suffisamment adaptées aux propriétés spatiales des traces numériques. C'est par ailleurs pour cette même raison que nous interrogeons la pertinence de délimiter des régions, par les polygones de Voronoï (Saravanou *et al*., 2015), sur une distribution ponctuelle dont la représentativité sociale et spatiale est soumise à question, et sachant que le contenu des tweets s'avère très hétérogène par l'absence de contrainte de formalisme. Au final, et dans notre cas des risques naturels, la question de la pertinence des distances revient à savoir si des tweets géolocalisés voisins[6], émis en réponse à un phénomène naturel, contiennent un discours sémantiquement proche ; en d'autres termes, la première loi de Tobler[7] se vérifie-t-elle dans le cas des tweets géolocalisés émis en réponse à un phénomène naturel, et si oui, à quel pas de distance ?

[4] Pour rappel, cette distance pouvaient être fixée arbitrairement ou en faisant appel à des analyses de distance entre les points.

[5] En d'autres termes, les quantités de tweets émis en un lieu précis.

[6] Encore faut-il définir la distance à laquelle on considère deux tweets comme voisins en fonction des territoires considérés et de leur activité virtuelle.

[7] "*Everything is related to everything else, but near things are more related than distant things*" (Luo et McEachren, 2013).

4.1.1.3. Une géographie humaine des traces numériques géolocalisées ?

En ce qui concerne les thématiques de recherche, un point central fait, selon nous, défaut : la recherche en sciences humaines et sociales s'est lancée dans l'exploration des possibilités thématiques qu'offraient les traces numériques géolocalisées pour saisir les rapports entre l'individu, l'espace et le temps mais sans avoir au préalable étudié la géographie humaine et sociale du réseau virtuel géolocalisé. Or, comme nous l'avons indiqué dans le chapitre 2, les utilisateurs émetteurs de tweets géolocalisés forment une population à part entière parmi l'ensemble des utilisateurs de Twitter. En conséquence, on étudie et on tente d'apporter des réponses à des enjeux collectifs en se focalisant sur les traces émises par une catégorie d'individus, sans parvenir à déterminer avec une grande précision leur profils socio-démographiques. (Ghoshs et Guha, 2013) avaient d'ores-et-déjà mis en évidence ce paradoxe : les utilisateurs tweetant sur le thème de l'obésité ne sont pas ceux qui sont les plus confrontés à l'enjeu de santé publique. Dans la problématique des risques naturels, la question qui se pose est alors la suivante : les utilisateurs géolocalisés actifs pendant les différentes phases répondant à un phénomène naturel sont-ils représentatifs des individus et des espaces affectés ?

Dans la continuité de ces réflexions, les paragraphes 4.1.2 et 4.1.3 présentent les positionnements thématique et méthodologique de la recherche doctorale, c'est-à-dire l'ensemble des questions préalables posées après connaissance de l'état de l'art, ainsi que la posture épistémologique adoptée vis-à-vis du traitement des traces numériques géolocalisées.

4.1.2. Cadre thématique d'amorce de la recherche

Cette section introduit la première étape de l'analyse des tweets géolocalisés et des données complémentaires. Cette première étape correspond à ce que nous avons décrit comme l'*aperçu global des données* dans la démarche méthodologique de l'analyse visuelle (Yi *et al*., 2008), introduite dans le chapitre précédent, et qui se concrétise par un premier lot de traitements et de représentations des traces et données. La poursuite de la définition des expériences (c'est-à-dire l'*ajustement* sur un lieu ou sur un nouveau questionnement émergeant des résultats) dépendra des résultats observés (en conformité avec les cadres épistémologiques présentés dans le chapitre 3 et la démarche exploratoire qui est ici adoptée en conséquence[8]).

[8] et elle-même décrite dans la section 4.1.3.1 de ce chapitre.

4.1.2.1. La géographie générale de l'activité virtuelle

Le matériau tweet géolocalisé étant de nature nouvelle et comme nous n'avons que peu d'informations sur ses limites dans le cadre d'une utilisation géographique, l'amorce de la recherche soulevait un certain nombre d'interrogations qu'on peut ici subdiviser selon deux questions générales et les types d'analyse et verrous qu'elles impliquent. Le premier questionnement fait écho à la limite mise en évidence dans le paragaphe 4.1.1, à savoir le manque d'études de la géographie humaine et sociale d'un réseau virtuel géolocalisé (tableau 4.1). Dans l'introduction de ce manuscrit, nous avons soumis un certain nombre d'exemples relatifs aux dynamiques d'intégration/exclusion des individus et des territoires par le développement croissant des plateformes et des services numériques. Il s'agit alors d'observer les manifestations qui témoigneraient de l'existence de cette variabilité socio-spatiale d'intégration au numérique sur notre terrain.

Tableau 4.1 : Questions relatives à la mesure d'une fracture numérique sur notre terrain d'étude

Question 1 - Le tweet géolocalisé comme matériau géographique : une fracture numérique socio-spatiale tangible ?
Quels sont les lieux qui concentrent l'activité tweeting géolocalisée ? Peut-on déterminer des profils de lieux attractifs/répulsifs de l'activité géolocalisée ?
Quelles sont les caractéristiques des populations productrices de contenus géolocalisés sur Twitter ?

Comme l'indique le tableau 4.1, l'objectif consiste donc à identifier les dynamiques d'intégration des territoires et des populations à la publication de contenus géolocalisés sur le Web, par les tweets. L'approche opérationnelle est, dans un premier temps, exclusivement quantitative : elle consiste à cartographier la distribution des tweets sur le territoire afin d'observer l'existence éventuelle de lieux ou objets qui polarisent l'activité tweeting et, si l'on peut percevoir une logique spatiale de distribution de ces tweets, à tenter de la modéliser. Les tests statistiques bi- ou multivariés peuvent être employés pour tenter de caractériser des profils de populations productrices de tweets géolocalisés mais il est alors impératif de prendre en compte le fait que le tweet géolocalisé est essentiellement acquis en situation d'activité et de mobilité, et non au domicile, contrairement aux données de recensement qui fournissent les variables indiquant le profil socio-démographique des individus et des lieux.

4.1.2.2. La géographie spécifique à l'activité virtuelle de crise

Le second questionnement général est orienté sur la déclinaison de l'événement virtuel et se focalise alors sur le tweet comme marqueur virtuel du phénomène physique et des événements consécutifs dans le réel : l'événement virtuel géolocalisé capture-t-il l'ensemble

des facettes d'un phénomène physique et de ses effets socio-spatiaux ? En d'autres termes, peut-on envisager de décrire un comportement systémique avec ses variables de contrôle expliquant les paramètres (quantité d'émissions, distribution spatio-temporelle, variabilité de la sémantique) de l'activité tweeting de crise ? Le tableau 4.2 établit la liste des premières questions mises en évidence dans la définition des éléments de cadrage de la recherche, dans la problématique des risques et catastrophes naturels.

Tableau 4.2 : Questions relatives à la thématique de l'événement virtuel généré en réponse à la survenue d'un phénomène naturel dommageable

Question 2 - Le tweet géolocalisé comme marqueur géographique d'un phénomène et d'événements réels : quelles informations et quelles limites ?
1 : Quels sont les lieux de l'activité tweeting de crise ? Ces lieux s'agitent-ils en fonction de la dynamique spatio-temporelle et de l'intensité des phénomènes et événements du réel ?
2 : Observe-t-on une décroissance de l'activité tweeting de crise en fonction de la distance aux lieux de survenue des phénomènes ?
3 : Des tweets géolocalisés émis dans une certaine proximité spatio-temporelle sont-ils lexicalement cohérents ?
4 : Quelles dynamiques spatio-temporelles peut-on mettre en évidence dans les tweets seuls ? Peut-on observer la diffusion d'un message d'alerte ? Les utilisateurs sont-ils mobiles dans l'espace en crise et si oui, que peut-on apprendre de leur parcours dans l'espace en crise ?
5 : Quel est le degré d'implication individuelle dans l'activité tweeting de crise ?
6 : La quantité de traces disponibles est-elle compatible avec les changements d'échelle et le passage à une résolution fine ?
7 : Dans l'événement virtuel, peut-on distinguer les différentes phases de la crise ? Existe-t-il une variabilité spatio-temporelle des territoires vis-à-vis du passage dans les différentes étapes de la crise ?
8 : Observe-t-on des variations spatiales du lexique dans une même crise ? Si oui, quels facteurs explicatifs peuvent être introduits ?
9 : Peut-on distinguer des lieux dans lesquels les utilisateurs sont plus actifs ? Si oui, peut-on en introduire les facteurs explicatifs ? Quels facteurs environnementaux ou sociaux font alors varier la mobilisation virtuelle des utilisateurs ?

Ces questions se focalisent ainsi, d'une part, sur l'identification des dynamiques spatio-temporelles et sémantiques de l'événement virtuel, qui seront comparées avec des données officielles de réalité-terrain, et d'autre part, sur l'identification des facteurs qui régissent ces dynamiques et les quantités de tweets émis en réponse à la survenue d'une perturbation. En revanche, ces questions soulèvent ou font écho à certaines difficultés à considérer dans les méthodologies d'analyse :

- la proximité spatio-temporelle et la cohérence lexicale entre les tweets (question 3) : dans la section 4.1.1.2 de ce chapitre, nous avons introduit le problème de la définition du seuil de distance au-delà duquel on va considérer que deux tweets ne sont pas voisins. A ce paramètre de distance seuil, il faut également prendre en considération le fait que la distribution des tweets forme un réseau irrégulier contrastant de vides et de pleins ; si l'on considère une zone de fortes densités et une zone de faibles densités de tweets, le critère du *proche* doit-il être identique ? La même question se transpose à la dimension temporelle : on peut envisager l'existence de périodes pendant lesquelles les rythmes d'émissions

s'accélèrent[9] ou ralentissent. Doit-on alors appliquer un pas de temps commun à l'ensemble des périodes lorsqu'on agrège des tweets ou définir des pas de temps spécifiques à chaque période détectée ? Enfin, si l'on considère l'hétérogénéité de la composante sémantique des tweets, ajoutée à la possible subjectivité du regard de l'utilisateur, on ne sait pas à l'avance si un même phénomène est appréhendé de manière identique.

- La plupart des questions énoncées dans le tableau sous-entendent une dépendance à l'analyse quantitative (questions 1, 2, 5, 6 et 9), fondée sur une distribution spatiale dont on connaît déjà l'irrégularité : le problème sous-jacent reste que nous ne savons pas si, en période de crise, la présence d'un tweet est plus significative que la présence d'un amas de tweets.

- La question 6 aborde les thèmes du changement d'échelle et de résolution : ceux-ci sont également liés aux problématiques quantitatives de la distribution des tweets géolocalisés. Peut-on appliquer une résolution spatiale et temporelle fine à l'exploration d'un événement virtuel dans un haut-lieu d'activité tweeting, de la même manière que dans un lieu où les tweets sont épars ? Cette question rejoint en outre l'enjeu décrit ci-avant : la quantité apporte-t-elle du sens à l'événement virtuel ?

Si l'on considère maintenant les outils à mobiliser pour explorer l'ensemble de ces questions, ceux-ci associent la cartographie spatio-temporelle (par les SIG) aux analyses statistiques et sémantiques. Concernant ce volet, nous pouvons également anticiper quelques écueils :

- la cartographie doit intégrer l'ensemble des composantes des tweets géolocalisés, auxquelles s'ajoutent les séries de données officielles permettant de recontextualiser le déroulement, l'intensité et les conséquences des phénomènes physiques sur les territoires. Bien qu'elle se veuille en tout premier lieu support exploratoire des analyses, la cartographie risque rapidement la surcharge visuelle et en conséquence, le manque de lisibilité graphique.

- La représentation efficace de la composante sémantique pose également problème : comme évoqué plus haut, nous ne savons pas à l'avance prévoir la cohérence ou la grande diversité du contenu sémantique des tweets. Cette difficulté rejoint la question d'ores-et-déjà soulevée de la significativité du nombre : l'affichage d'une poignée de mots-clés extraits des tweets suffit-il afin d'observer des tendances ou le sens lexical des tweets se dégage-t-il à travers la visualisation d'un ensemble de mots plus conséquent ? En outre, le nuage de mots, qui reste la représentation la plus courante de l'information sémantique, la valoriseraient davantage s'ils affichaient les mots fréquemment associés, afin de mettre en évidence les différentes tonalités adoptées en réponse à un phénomène ou événement du réel[10].

[9] Ce phénomène était notamment palpable avec les séismes, vis-à-vis desquels le réseau enregistrait une augmentation rapide des tweets dans les premières minutes suivant la survenue de la secousse.

[10] Soit le manque qu'on avait identifié chez (Croitoru *et al.*, 2017) qui soumettaient deux nuages de mots générés à partir de tweets mentionnant le Président Obama, mais ceux-ci ne mettaient pas en évidence en quels termes les tweets positifs ou négatifs classés par l'analyse de sentiments évoquaient l'*Obamacare*.

4.1.2.3. Synthèse des questionnements généraux

D'une manière générale, ce questionnement s'articule sur la méthode traditionnelle de recherche "*Qui, Quoi, Où, Quand, Comment, Combien, Pourquoi ?*". L'ensemble des questions présentées dans les encadrés précédents sont donc guidées par deux axes :

- le premier axe tient à mesurer l'efficacité du tweet géolocalisé comme marqueur de la primo-existence virtuelle des territoires et des populations ;
- le second axe consiste à vérifier et à approfondir les connaissances acquises sur les comportements spatio-temporels et sémantiques de l'événement virtuel engendré en réponse à un phénomène naturel.

Dans la problématique de l'étude géographique des risques naturels, ils nous paraît en effet essentiel de nous interroger sur le tweet comme indicateur géographique : il existe, à notre connaissance, peu de travaux qui mesurent le tweet géolocalisé comme un indicateur humain et social (et en l'occurrence, les résultats des travaux varient en fonction de la composante étudiée de la trace). (Ripberger *et al*., 2014) soulevaient la question des émissions de tweets comme un indicateur de l'attention portée par les utilisateurs aux messages d'alerte diffusés sur le réseau par les comptes officiels des organismes responsables en la matière. Ils établissaient alors deux modèles de régression entre le nombre de tweets officiels lanceurs d'alerte *tornade* par jour et le nombre de tweets non officiels contenant le mot-clé *tornado*, puis entre le nombre d'habitants potentiellement affectés et le nombre de tweets émis contenant le mot-clé *tornado*. Les auteurs validaient les modèles par l'existence d'une forte corrélation positive entre les séries de variables indiquées, soulignant ainsi l'engagement d'une dynamique virtuelle globale en réponse à l'émission des tweets marqueurs d'alerte officiels, mais encore l'existence d'une relation entre le nombre de personnes potentiellement affectées et le volume de la réponse virtuelle.

Mais certains travaux récents affichent des résultats qui se détachent de cette posture témoignant du tweet géolocalisé comme un apport positif à la discipline géographique. (Zou *et al*., 2018) cherchaient à mesurer l'éventuelle existence de disparités socio-spatiales dans la diffusion de tweets géolocalisés pendant les différentes phases de l'ouragan Harvey, dans les comtés frappés du Texas et de la Louisiane. Les résultats confirmaient l'hypothèse posée par les chercheurs : la fracture du numérique persistait pendant les différentes phases de réponse à l'ouragan, c'est-à-dire que les communautés à forte empreinte numérique pendant le phénomène correspondaient à des invididus plutôt aisés et moins vulnérables[11].

Ainsi, lorsqu'on relit les questions posées dans les tableaux 4.1 et 4.2, et au vu des informations collectées dans l'état de l'art, il serait quasi instinctif de poser certaines hypothèses selon lesquelles :

[11] Résultats ayant ainsi tendance à confirmer les tweets d'Anthony Robinson, cités dans le chapitre 3.

- les hauts-lieux de concentration de l'activité tweeting correspondent aux espaces urbains (Andrienko *et al*., 2013 ; Lucchini *et al*., 2016 ; Cebeillac *et al.*, 2017), ceux-ci bénéficiant généralement d'une meilleure connectivité aux infrastructures numériques ; de même, les populations plus aisées disposent d'une meilleure accessibilité financière à l'Internet mobile et par conséquent, sont plus aptes à participer régulièrement à la création de contenus géolocalisés sur les plateformes numériques (Li *et al*., 2012). Tous les individus et tous les territoires ne sont pas équitablement visibles à travers les contenus numériques géolocalisés (Zou *et al*., 2018).

- Les utilisateurs ayant tendance à tweeter à propos de ce qu'ils font ou voient à l'instant *t*, l'événement virtuel se positionne comme concomitant à la dynamique des phénomènes et événements du réel (Sakaki *et al*., 2010) : on peut donc suivre l'évolution d'un phénomène physique et des événements sociaux résultants comme le passage d'une phase de préparation à une phase de gestion de crise puis de résilience (Blanford *et al*., 2014).

- La réponse virtuelle varie, en termes de volume, en fonction de l'intensité des phénomènes : plus un phénomène est violent, plus l'agitation virtuelle est forte (De Longueville et al., 2008 ; Hertfort *et al*., 2014).

Mais au final, compte-tenu de notre méconnaissance relative à la pertinence du matériau tweet géolocalisé comme marqueur géographique, à l'éventuelle répétitivité des résultats observés dans les publications consultées, aux risques de biais socio-spatiaux et à l'imprévisibilité humaine, nous ne savons pas d'avance si l'ensemble des questions posées dans les tableaux pourront être abordées et si les résultats des analyses s'accorderont aux intuitions logiques formulées ci-dessus.

4.1.3. Cadre méthodologique adopté pour guider la recherche

4.1.3.1. Théorie du positionnement épistémologique pour l'analyse des tweets utiles extraits

En premier lieu, afin de préciser les remarques concluant le paragraphe 4.1.2 et les méthodologies de recherche introduites au paragraphe 3.3.3 (*hypothesis-free science*, *grounded theory* et *raisonnement abductif*), nous présentons notre choix de démarche de recherche : celui-ci s'articule autour du raisonnement abductif et de la méthode inductive (Hoffmann, 1999 ; Kell et Oliver, 2003). Contrairement aux principes de la *grounded theory*, nous n'entrons pas dans l'analyse des tweets géolocalisés sans connaissance ou questions de départ (cf. tableaux 4.1 et 4.2). Au contraire, les questions exprimées dans le paragraphe 4.1.2 ont le rôle de prétexte : elles orientent un premier lot d'analyses dont on ne sait pas si les résultats seront conformes aux quelques hypothèses intuitives énoncées ci-avant, quelles que soient les échelles géorgaphiques considérées, de même que nous ne savons pas si une seule expérience fournira une réponse précise à toute question posée ou si cette réponse pourra se décliner sous plusieurs aspects. C'est pourquoi nous considérons que toute analyse focalisée

sur les traces numériques géolocalisées seules ou les croisant à des jeux de données officielles peut aboutir à des résultats imprévus qui soulèvent de nouveaux questionnements plus approfondis. Nous préférons alors définir une méthodologie d'expérience fondée sur la *primo-question*, l'observation des résultats de la *primo-analyse*[12] effectuée en réponse à cette question, et la recherche d'une hypothèse explicative à l'observation faite ou l'exploration d'une structure inattendue. Mais également, dans une logique inductive, nous cherchons à savoir si les structures mises en évidence sont répétitives ou si elles sont le fruit d'un hasard. Le schéma ci-après (figure 4.1) offre l'aperçu de la démarche épistémologique globale, utilisée dans l'analyse des tweets inscrits à la problématique de recherche :

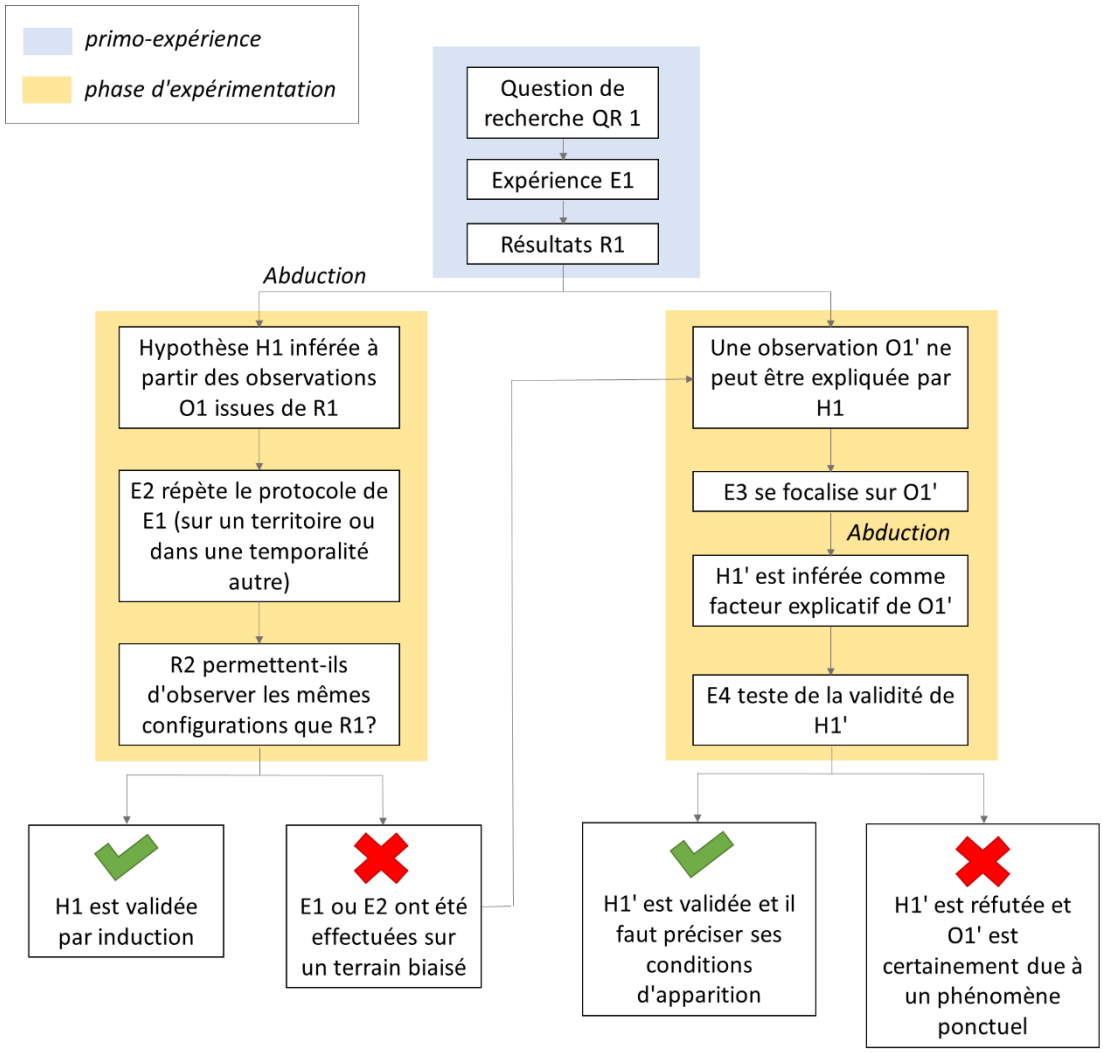


Figure 4.1 : Schéma explicatif de la démarche exploratoire des traces et données, et de construction de la connaissance (C.Cavalière)

[12] Ce qui correspond à l'étape de l'*aperçu global des données* de la méthode de l'analyse visuelle donnée par (Yi *et al.*, 2008).

Cette démarche fonctionne alors comme suit : une première expérience E1 est définie à partir d'une question de recherche posée QR1. Les premiers résultats R1 permettent d'inférer par abduction une hypothèse plausible qui explique l'observation de R1. La deuxième expérience R2 répète scrupuleusement le protocole de E1 sur un terrain analogue ou sur le même terrain mais dans une temporalité différente (est-ce-que la structure mise en évidence sur le territoire A à *t(0)* se répète sur le territoire B qui présente les mêmes caractéristiques à *t(0)*, ou sur ce même territoire A, mais à *t+n* ?). Si les résultats R2 permettent d'observer les mêmes configurations identifiées dans R1, alors on peut valider la véracité de l'hypothèse explicative H1. Dans le cas contraire, il est possible que les expériences E1 ou E2 aient été effectuées, sans le savoir, dans des conditions biaisées ; en comparant les configurations observées dans R1 et dans R2, on pourra sans doute détecter O1', une observation correspondant à une anomalie de terrain qui ne peut pas être expliquée par l'hypothèse inférée H1. Une nouvelle expérience, E3 explore alors le contenu de la configuration O1', dans le but d'inférer une nouvelle hypothèse explicative H1' d'O1'. Cette hypothèse H1' doit être testée dans une nouvelle expérience E4, fondée sur un protocole identique à E3. Si H1' est vérifiée, il faudra alors identifier les facteurs qui encadrent ses conditions de validité. Dans le cas contraire, l'anomalie observée O1' peut être assimilée à un phénomène inhabituel qui ne s'est enregistré qu'une seule fois dans les traces numériques géolocalisées.

Pour illustrer cette démarche, on peut prendre un exemple simple, relatif à la distribution spatiale des tweets géolocalisés. Le questionnement de recherche de départ QR1 cherche à mettre en évidence les facteurs de concentration de l'activité tweeting à l'échelle d'un Etat. L'expérience E1 consiste à identifier ces lieux en réalisant une carte de chaleur ; l'abduction intervient dès lors qu'on recherche un facteur explicatif à R1. En l'occurrence, en superposant R1 à une couche SIG des villes, on se rend rapidement compte que les foyers de forte concentration de tweets géolocalisés correspondent aux milieux urbains et métropolitains. On infère alors l'hypothèse H1 : les densités de tweets géolocalisés dépendent des milieux et des densités de population : ainsi, les plus forts foyers d'activité se trouvent dans les grandes métropoles alors que les déserts de tweets géolocalisés se trouvent dans les territoires ruraux. En explorant plus en détails la carte de chaleur, on se rend compte de la présence d'un foyer de tweets en dehors de toute agglomération : cela entraîne une nouvelle étape d'abduction car il faut identifier un facteur explicatif à l'existence de ce foyer. Dans ce foyer, on peut alors examiner la sémantique des tweets ou encore compter le nombre d'utilisateurs actifs. Ici, c'est ce dernier paramètre qui est révélateur : un seul utilisateur hyperactif est responsable de la présence du foyer de tweets. On infère alors une nouvelle hypothèse H1' : un unique utilisateur très actif (qu'il corresponde à un individu, à un *spammer* ou à un robot) peut faire apparaître un îlot de tweets.

On souhaite maintenant savoir si H1 est propre au comportement des utilisateurs du premier pays étudié ou si ce comportement peut être généralisé à d'autres populations d'utilisateurs dans un pays différent. On génère de nouveau une carte de chaleur (E2) à partir des tweets géolocalisés du second pays d'étude qu'on superpose à la couche des villes. La

seconde série de résultats R2 permet d'identifier les mêmes configurations que celles de R1. Les foyers de l'activité tweeting géolocalisée (O2) correspondent aux espaces urbains et aux métropoles ; la pertinence de H1 est donc très vraisemblable. Toutefois, on distingue une nouvelle configuration anormale (O2') signalant un foyer en dehors d'un milieu urbain. On teste H1' mais les résultats sont ici négatifs : on distingue des dizaines d'utilisateurs géolocalisés actifs dans le foyer. Dans une nouvelle étape d'abduction, on explore le contenu du foyer pour inférer une nouvelle hypothèse comme facteur explicatif de O2'. Comme l'analyse quantitative n'a rien donné, on explore ici le qualitatif : l'analyse sémantique du contenu des tweets permet alors d'identifier un événement culturel ponctuel ayant provoqué un afflux temporaire d'utilisateurs géolocalisés. On peut alors inférer H2 : des événements inhabituels et ponctuels peuvent générer des îlots d'agitation temporaire. Il conviendra alors de renouveler les expériences afin de savoir si H1' et H2 se répètent sur d'autres territoires ou dans d'autres conditions.

4.1.3.2. Définition de méthodes d'extraction de tweets utiles à la problématique

Dans un deuxième temps, nous nous intéressons à la question de l'extraction de tweets géolocalisés utiles, émis en réponse à la problématique des phénomènes naturels dommageables d'origine hydrométéorologique. (Steiger *et al.*, 2015) indiquaient, comme point faible récurrent des méthodologies d'extraction de tweets utiles dans les travaux existants, que celles-ci étaient très souvent limitées à une extraction focalisée sur une poignée de mots-clés sélectionnés par les chercheurs (cf. chapitre 3). Ici, nous souhaitons approfondir ces méthodologies, en dépassant la simple approche supervisée et en intégrant, d'une part, la dimension spatiale des tweets géolocalisés et d'autre part, le type de phénomène physique ainsi que l'échelle d'étude considérée.

4.1.3.3. Positionnement vis-à-vis des outils d'analyse existants

Comme le laissaient supposer les questions contenues dans les tableaux 4.1 et 4.2, les méthodologies d'analyse des tweets géolocalisés reposent, même si elles sont intégrées dans des interfaces de géovisualisation, sur des outils *hérités* (cartes de chaleur, agrégation de données à diverses échelles, animation spatio-temporelle, tests statistiques, nuages de mots[13], *etc.*). Même si leur efficacité et leur adaptabilité aux traces numériques s'avèrent remises en question par certains chercheurs (Kitchin, 2014 ; Miller et Goodchild, 2015) comme indiqué dans le chapitre 3, ce ne sont pas les outils et leurs capacités que nous interrogeons mais notre manière d'appréhender les processus d'exploration et de construction de connaissances avec le nouveau matériau qu'est le tweet géolocalisé. En conséquence, il s'agit de savoir s'il est envisageable, par les outils dont les géographes/géomaticiens disposent

[13] Le concept de la représentation de l'information sémantique en nuage de mots apparaît dans les années 1990. Sur le Web 2.0, Flickr fut la première plateforme de partage (de photos) à implémenter les nuages de mots. Source : https://en.wikipedia.org/wiki/Tag_cloud (Consulté pour la dernière fois le 14/10/2019).

actuellement, de construire une recherche qui déduit progressivement ses hypothèses en fonction des observations effectuées sur les résultats d'analyse préalables.

D'un autre côté, on pourrait également envisager le développement premier d'une interface de géovisualisation et l'évaluation consécutive de son potentiel (Pezanowski *et al*., 2017). En effet, tels qu'ils ont été introduits dans le chapitre 3, les principes d'utilisation de la géovisualisation s'accordent avec notre positionnement épistémologique, dans la mesure où ils favorisent l'observation et la formulation d'hypothèses à partir de données dont les connaissances préalables sont faibles (Dykes *et al.*, 2015). Mais ici, notre positionnement vis-à-vis des méthodes d'analyse ne s'engage par sur cette voie. Selon nous, il est essentiel de mesurer, avant d'envisager tout développement d'environnement de géovisualisation, le potentiel du tweet en tant qu'indicateur géographique dans un contexte particulier. Nous souhaitons donc identifier les outils et méthodes par lesquels on pourra construire une connaissance (en suivant la logique analyse/observation/hypothèse) et selon quelles résolutions spatio-temporelles. Nous rappelons que sur le plan cartographique, nous considérons le carte en tant qu'instrument de soutien à l'exploration, tout comme en géovisualisation.

4.2. Définition d'une méthodologie d'extraction de tweets utiles

Comme nous l'avons précédemment indiqué, les méthodes approfondies pour l'extraction de tweets utiles à une problématique particulière sont rarement développées. Cette partie présente la méthodologie proposée pour améliorer les conditions de déroulement de cette étape primordiale, qui croise une approche supervisée (le chercheur définit les critères de sélection des tweets selon sa propre appréhension du phénomène à l'étude) à une approche non supervisée (le chercheur définit les critères de sélection des tweets en fonction de leur contenu). Dans un second temps, l'axe valorise cette méthodologie en développant les cas d'utilisation possibles et en introduisant les fonctionnalités intégrées dans une interface exploratoire destinée à en automatiser les étapes.

4.2.1. Théorie encadrant la définition de la méthode

4.2.1.1. Eléments de cadrage de la méthode

Dans le chapitre précédent, nous avons défini comme *tweet utile* tout tweet dont le contenu sémantique est en rapport avec la question étudiée par les chercheurs (lorsque celle-ci focalise son intérêt sur une facette précise de l'ensemble des messages émis au quotidien). Dans notre problématique, tout tweet utile est considéré comme un *tweet de crise*. Ces tweets

de crise se trouvent également noyés dans le bruit créé par l'information quotidienne. La première étape des travaux entrepris consiste donc à définir une méthodologie reproductible d'extraction de tweets de crise. De la même manière que les recherches présentées dans le chapitre précédent, nous considérons le tweet de crise comme un tweet géolocalisé contenant au moins l'un des mots-clés que nous aurons inventoriés, et émis dans le contexte approprié.

La définition de la méthodologie d'extraction de tweets de crise s'est appuyée sur plusieurs types de phénomènes résultant d'aléas de natures différentes. En consultant régulièrement les actualités liées à la survenue de phénomènes hydrométérologiques sur les années 2016 et 2017 aux Etats-Unis et en France, nous avons identifié et focalisé notre intérêt sur :

- les phénomènes extrêmes récurrents et saisonniers d'intensité variable (pluies intenses, crues lentes ou rapides) ;
- les phénomènes extrêmes plus rares (tempêtes de blizzard, ouragans).

Pour chaque phénomène suivi, nous avons assuré une veille rapide sur le Web afin d'observer comment se comportait la diffusion des informations relatives à ces phénomènes en termes de résonance géographique (autrement dit, à quelle échelle géographique trouve-t-on des articles de presse ou des tweets mentionnant un type de phénomène ?). Il s'agit alors de suivre les tendances de l'activité Twitter qu'on peut actualiser en fonction de villes ou de pays (cf. figure 2.3 du chapitre 2), et des articles de presse en ligne proposés par *Google Actualités* (en faisant varier la langue et le pays concerné). Il est rapidement apparu que les phénomènes extrêmes rares, susceptibles d'affecter des millions d'individus, bénéficient d'une très forte médiatisation qui résonne souvent à l'échelle du globe. La résonance virtuelle de ces phénomènes peu fréquents dépasse largement l'enveloppe spatiale des territoires affectés. A l'inverse, les phénomènes habituels ne provoquent pas une telle mobilisation virtuelle, bien que ceux-ci puissent provoquer de nombreux dégâts et faire des victimes. Ces phénomènes surviennent à une échelle locale et affectent un espace donné de manière cyclique (ils sont habituels, et non exceptionnels) : période des tornades, des crues éclairs, *etc*. L'enveloppe spatiale de l'événement virtuel résultant ne dépasse pas, ou de peu, le territoire affecté dans le réel. Par exemple, dans *Google Actualités*, les inondations régulières survenues au Texas en avril 2016 (recherche en fonction des mots-clés *inondations Texas*) ne trouvent pas d'écho dans les quotidiens français en ligne, ce qui n'est pas le cas de l'ouragan Harvey en 2017. Dans la presse américaine, la recherche *Texas floods* fournit des résultats sur les pages en ligne des grandes chaînes de télévision nationale (*CNN*, *Foxnews*) et des quotidiens ou hebdomadaires locaux (*The Texas Tribune*, *Click2Houston*, *DallasNews*, *etc.*). L'ouragan Harvey

trouvait son écho dans toutes les échelles de la presse américaine et son spectre s'avère toujours présent, deux ans plus tard, dans les journaux locaux[14].

A l'aide de ces premières observations et des réflexions amorcées dans le chapitre 2, nous théorisons les trois facteurs suivants comme guides pour l'orientation de la méthodologie de recherche et d'extraction de tweets de crise :

- *Facteur 1* : le *type* de phénomène étudié. Comme indiqué ci-avant, nous avons noté deux types de phénomènes naturels présents sur les réseaux, qui ne se traduisent pas par un impact réel et virtuel analogue. Le phénomène extrême rare provoque une mobilisation massive (l'utilisateur qui l'évoque peut avoir vécu le phénomène ou peut ne pas l'avoir vécu). Le phénomène local et récurrent n'engendre manifestement pas cet effet de masse : l'utilisateur qui ne l'a pas vécu reste passif ou n'est quasiment pas informé.

- *Facteur 2* : la *quantité* de tweets de crise émis en réponse à un phénomène physique. Comme indiqué dans le chapitre 2, seules certaines catégories de la population sont actives dans la production de contenus géolocalisés via les réseaux sociaux. Les facteurs socio-démographiques se présenteraient donc comme des variables de contrôle de la quantité de traces créées à l'échelle locale. On pourrait également en citer deux autres, à cette même échelle : le comportement individuel de l'utilisateur. Certains individus sont sensibles aux phénomènes naturels et tweetent à la moindre pluie ; d'autres utilisateurs tweetent lorsqu'ils jugent que les effets du phénomène dépassent leur expérience vécue. Enfin, selon sa situation à l'instant, l'individu peut choisir de se mettre immédiatement à l'abri ou de filmer et de tweeter en dépit de sa sécurité. Le lieu de survenue du phénomène pourrait également être considéré comme un facteur de régulation de la quantité de tweets émis : des aléas peuvent en effet survenir dans des espaces peu peuplés et/ou dans lesquels la population est peu utilisatrice de Twitter ; la disponibilité de tweets de crise sera alors très restreinte[15]. A une plus petite échelle géographique, on peut anticiper le *type* de phénomène mentionné précédemment comme variable de contrôle de la réponse virtuelle. Les phénomènes extrêmes rares sont diffusés à l'échelle du globe et alimentés par les médias traditionnels et le Web, amplifiant alors la réponse virtuelle : la sensibilité des utilisateurs semble aiguisée par les images locales diffusées à l'échelle internationale. De même, un utilisateur situé à distance du phénomène peut se trouver indirectement affecté et se manifester sur le réseau (en cas

[14] Exemples d'articles publiés en ligne dans la presse locale : https://www.texastribune.org/2019/04/18/under-the-dome-episode-10-hurricane-harvey-texas/; https://www.houstonchronicle.com/business/columnists/tomlinson/article/Builders-battle-most-basic-rule-changes-following-14276239.php

[15] Par exemple, si l'on recherche les tendances tweeting en Inde et au Pakistan entre le 10 et le 12 août 2019, on ne trouve aucun hashtag ou sujet de discussion majeur mentionnant les moussons meurtrières ayant frappé les deux pays. Les tendances gravitent autour de l'Aïd-El-Kebir : en 2019, la fête a coïncidé avec le phénomène hydrométéorologique saisonnier et semble avoir éclipsé l'événement virtuel mousson et ce, malgré son intensité exceptionnelle.

de perturbation du trafic aérien ou encore s'il s'inquiète pour des proches localisés dans le territoire affecté).

- *Facteur 3* : la *pertinence* des tweets. Ce paramètre reste ici mal encadré, la production de tweets géolocalisés n'étant régie par aucune des normes applicables aux données traditionnelles. Dans le cadre des traces numériques géolocalisées, ce paramètre peut sous-entendre des propriétés variées : les chercheurs considèrent fréquemment que tout tweet pertinent contient un *hashtag* ([Lin *et al*., 2013], la pertinence est alors rattachée à l'identification d'un sujet précis dans le tweet). Les travaux publiés par (Dashti *et al*., 2014) ont plutôt tendance à considérer la proximité aux lieux du phénomène. Ici, c'est donc la première loi de Tobler[16] qui est considérée comme indicateur de pertinence du tweet pour l'analyse : le tweet situé dans un territoire affecté par le phénomène contient des informations en rapport avec les effets de ce phénomène. On pourrait également prendre en compte la longueur d'un tweet (dans le cadre de la détection de phénomènes) ou la présence d'un vocabulaire particulier (verbes conjugués, adjectifs d'intensité, adverbes temporels), la présence d'une photographie ou d'une vidéo, *etc*. Au final, ce critère de la pertinence devrait sans doute être défini indépendamment de l'existant, en fonction des exigences codifiées par les chercheurs dans leurs travaux respectifs[17].

4.2.1.2. Définition de la méthodologie et cas d'application

Au regard des considérations précédentes, nous soumettons une proposition méthodologique d'extraction de tweets de crise, qui se décline selon trois approches prenant en compte le type de phénomène et l'échelle géographique étudiée. Ces différentes approches, qui commandent un type d'extraction ici appliquée à des cas d'étude tests, sont présentées dans le tableau 4.3 ci-après :

[16] "*Everything is related to everything else, but near things are more related than distant things*" (Luo et McEachren, 2013).

[17] Le problème qui se pose est le suivant : on collecte un jeu de tweets bruts pour étudier un phénomène survenu sur un territoire délimité. A partir de ce jeu de tweets bruts, on crée deux jeux de tweets de crise : le premier est uniquement filtré en fonction de mots-clés définis par les chercheurs (comme on trouve couramment). Pour le second, on a préalablement identifié les lieux affectés du territoire et on sélectionne les tweets situés dans le voisinage de ces lieux et contenant des mots-clés précis. Si l'on applique ensuite la même méthodologie d'analyse aux deux jeux de tweets de crise, observera-t-on des résultats similaires ?

Tableau 4.3 : Principes des méthodologies proposées pour l'extraction de tweets de crise géolocalisés

Type et échelle du phénomène	Caractéristiques du jeu de tweets de crise	Approche préconisée	Cas d'étude test
Phénomène extrême / analyse à l'échelle globale	Echelle globale d'un phénomène à réponse virtuelle massive : filtrage par un critère de qualité	Lexicale par hashtags	Tempête de blizzard Jonas, janvier 2016, Etats-Unis
Phénomène récurrent de moindre ampleur spatiale et sociale / analyse d'un phénomène extrême à l'échelle locale	Passage à l'échelle locale : priorité donnée à la quantité de tweets	Lexicale par mots-clés	Episodes pluvio-orageux et inondations au Texas, mars-juin 2016 Ouragan Harvey dans les comtés frappés du Texas et de la Louisiane, août 2017
Tous types de phénomènes / analyse à l'échelle locale	Echelle locale : priorité donnée à la proximité des tweets à un lieu témoin	Spatiale et lexicale autour d'objets d'intérêt identifiés sur le territoire affecté	Crue de la Seine, France, juin 2016

En considérant les trois facteurs (type de phénomènes, quantités et pertinence des tweets) définis, quelle approche appliquer et dans quel cas d'étude ? A partir de l'exemple de la tempête de blizzard Jonas survenue dans le centre et nord-est des Etats-Unis en janvier 2016, on souhaite tout d'abord constituer un jeu de tweets de crise pour étudier l'événement virtuel à l'échelle du pays (donc à petite échelle géographique). La tempête de blizzard correspond à l'un de ces phénomènes extrêmes rares qui bénéficient d'une couverture médiatique internationale et dont le retentissement virtuel dépasse l'enveloppe spatiale de l'espace réel physiquement affecté. Cet événement virtuel se caractérise par une mobilisation massive des utilisateurs de Twitter qui s'activent sur le réseau, même s'ils ne sont pas physiquement présents et ne l'ont pas physiquement vécu. Dans le cadre de ces réactions massives au réel, on peut filtrer les tweets en fonction des hashtags et privilégier ainsi l'information identifiée par un ou plusieurs thèmes précis. Le principe de cette approche s'appuie alors sur les travaux de (Lin *et al*., 2013) qui ont souligné l'existence d'un comportement propre à ces événements virtuels majeurs : un certain nombre de hashtags sont créés et massivement adoptés par les utilisateurs (surnommés les hashtags *Winners*, toujours par [Lin *et al*., 2013]). En parallèle, d'autres hashtags sont créés et associés avec ces *Winners*. En général, les tweets émis dans le cadre d'une réponse massive sont marqués de plus d'un hashtag. A l'échelle globale du phénomène, on peut donc orienter l'extraction lexicale sur la mise en évidence progressive de ces différentes catégories de hashtags.

Dans un second temps, le même phénomène est étudié à l'échelle d'une aire métropolitaine physiquement affectée par la tempête de blizzard. Si l'on se fonde sur le jeu

de tweets de crise précédent, construit uniquement à partir de la recherche de hashtags, on risque de tenir à l'écart une part non négligeable des tweets émis à l'échelle locale en réponse au même phénomène ; à grande échelle, il convient de ne pas négliger la quantité d'information de crise qu'on va extraire. Si, à l'échelle du pays, tout utilisateur non affecté peut évoquer son opinion ou son ressenti sur le phénomène en ajoutant un hashtag, il est probable que l'individu affecté sur le terrain crée une information davantage spontanée et rapide, sans ajout systématique de plusieurs hashtags (à la manière des tweets courts détectant les séismes, mentionnés par [Sakaki *et al*., 2010] et [Bossu *et al*., 2018]). Dans un tel cas, on considère une extraction lexicale globale, focalisée à la fois sur la recherche de hashtags et de mots-clés simples. On pourra également appliquer cette méthode dans le cas des phénomènes météorologiques récurrents qui ne provoquent pas cet effet de réponse massive sur le réseau et qui font l'objet d'analyses à leur échelle de survenue.

Enfin, on souhaite explorer l'événement virtuel à grande échelle, en fonction des lieux physiques qui enregistrent une réponse virtuelle. Dans un lieu de la métropole, on identifie des objets affectés, qu'on appelera ici des *objets d'intérêts* (OI), et on cherche à spatialiser l'emprise de l'événement virtuel autour de ces OI. Le schéma de la figure 4.2 illustre le principe de l'approche spatiale.

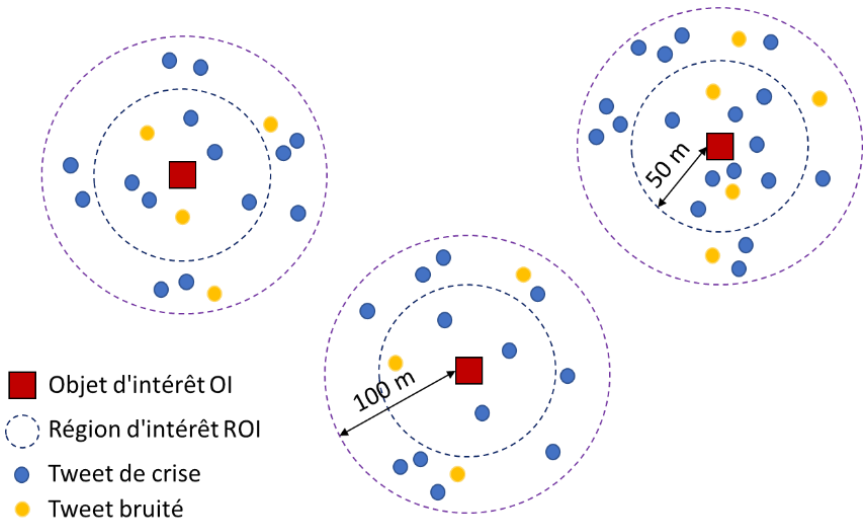


Figure 4.2 : Méthode d'extraction spatiale de tweets géolocalisés autour d'objets d'intérêt locaux

Quel que soit le type de phénomène naturel en cause, la première étape de cette approche consiste à identifier les objets du territoire qu'on considère comme les OI (un cours d'eau, un bâtiment, un tweet marqueur d'alerte, un quartier inondé, *etc*.). Il convient ensuite de tracer une *région d'intérêt* (ROI, par exemple, une zone tampon) qui entoure chaque OI, à une distance déterminée. Les tweets bruts contenus dans cette ROI sont sélectionnés puis soumis à l'extraction lexicale globale (mots-clés et hashtags) précédemment décrite. On peut

alors créer une nouvelle ROI autour des OI initiaux pour élargir le champ de recherche des tweets de crise (et en relançant la recherche de mots-clés à partir des tweets extraits dans cette nouvelle ROI). Cette dernière méthode repose alors sur le postulat de la première loi de Tobler, précédemment citée.

4.2.2. Opérationnalisation de la méthode théorique

4.2.2.1. Collecte initiale des tweets

A l'automne 2014, le LIG a mis en service une infrastructure de collecte et de stockage des tweets contenant un attribut de localisation[18]. Cette infrastructure dispose d'une connexion permanente à l'*API Streaming* de Twitter et d'une base de données PostgreSQL (nommée *twitterdb*) qui stocke les tweets retournés en continu par la requête du serveur[19]. La base de données *twitterdb* contient un total de 32 tables. Dans un premier temps, seules deux tables nous intéressent : la table *tweet* stocke l'ensemble des attributs de tout tweet retourné au serveur par l'*API* ; la table *countries* contient les géométries de l'ensemble des pays du globe. La figure 4.3 présente l'ensemble des champs inclus dans les deux tables mentionnées, ainsi que leurs types respectifs :

tweet	
tweet_id	*bigint*
user_id	*bigint*
text	*text*
created_at	*timestamp with time zone*
captured_at	*timestamp with time zone*
user_tweet_num	*integer*
saved_at	*timestamp with time zone*
gps	*geometry*
collector_id	*integer*
lang_id	*integer*
place_id	*bigint*

countries	
id	*text*
countryname	*text*
geom	*geometry*

Figure 4.3 : Liste des champs des tables tweet et countries dans la base de données twitterdb

Dans le cadre de cette recherche, les champs suivants de la table *tweet* sont utilisés : *tweet_id* (numéro unique d'identifiant de tweet), *user_id* (numéro d'identifiant de chaque utilisateur[20]), *created_at* (horodatage de la publication du tweet au fuseau horaire GMT+01), *text* (texte du tweet) et *gps* (coordonnées GPS du tweet). La table *countries* est exclusivement

[18] A l'origine, cette infrastructure de collecte a été mise en place pour répondre aux besoins du projet CNRS *Crowdhealth* (Amer-Yahia *et al.*, 2015), focalisé sur l'étude des tweets géolocalisés pour les questions relatives à la santé et à la nutrition. A noter également que suite à l'application de la RGPD en mai 2018, l'accès aux données de la base *twitterdb* se fait désormais à travers une vue anonymisée.

[19] Le serveur interroge le flux *filter* de l'*API Streaming*, qui permet de collecter des tweets en appliquant des critères sélectifs (cf. paragraphe 2.2.1.2 du chapitre 2) : ici, il s'agit d'une information de localisation.

[20] Les noms d'utilisateurs ne sont pas collectés (bien que Twitter les mette à disposition des développeurs).

utilisée pour exécuter les requêtes spatiales nous permettant de sélectionner les tweets géolocalisés émis dans les pays d'intérêt.

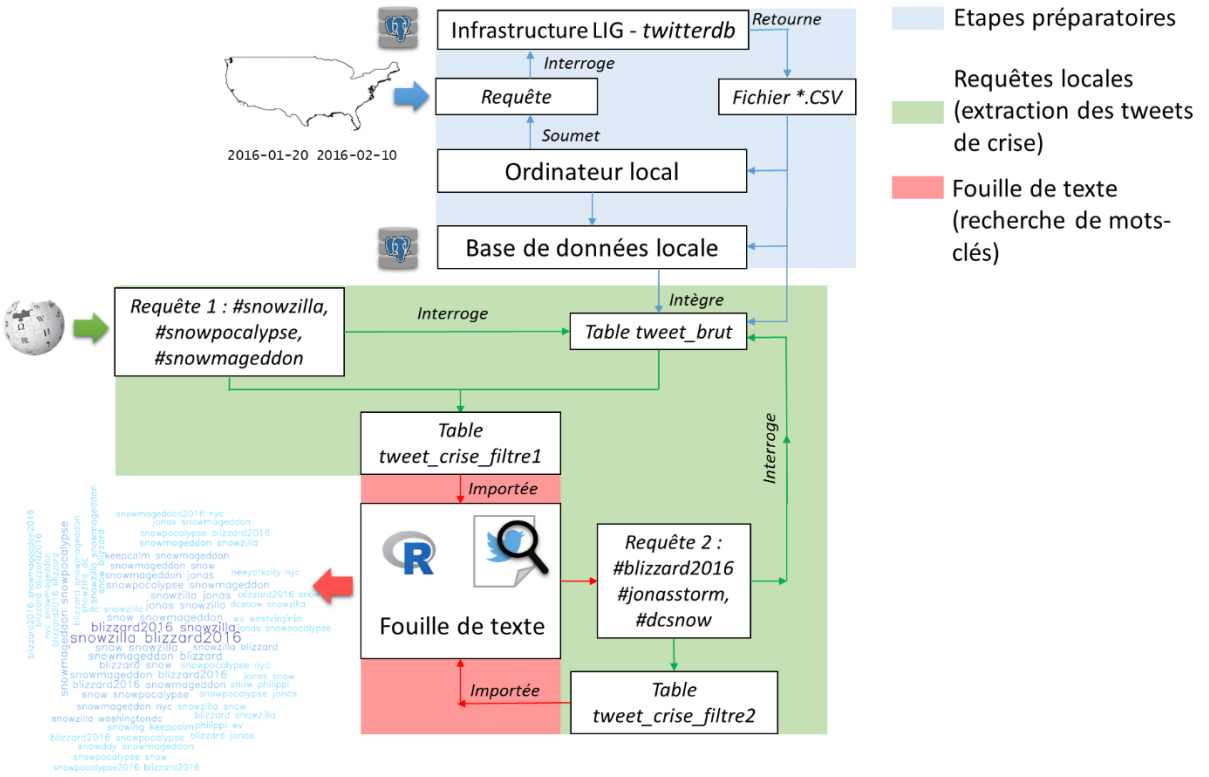
Avant toute démarche de construction de jeux de tweets de crise, il nous faut constituer un jeu de tweets bruts en interrogeant, dans *twitterdb*, les deux tables mentionnées, en fonction de deux types de variables : la période pendant laquelle le phénomène en question s'est manifesté (champ *created_at* de la table *tweet*) et le territoire affecté (requête spatiale croisant le champ *gps* de la table *tweet* et le champ *geom* de la table *countries)*. Les résultats de cette première requête sont stockés dans une nouvelle base de données PostgreSQL, créée dans un ordinateur en local et indépendante de l'infrastructure de collecte du LIG.

4.2.2.2. Application technique de l'approche lexicale

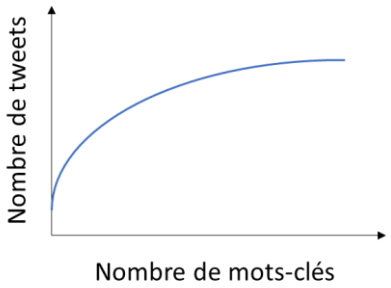
Dans le paragraphe 4.2.1.2, nous avons indiqué, quelle que soit l'échelle d'analyse et l'approche employée, l'existence d'une *étape de recherche* de mots-clés ou de hashtags pour toute extraction lexicale. En fait, pour pallier les manques mis en évidence par (Steiger *et al*., 2015) et donc cette tendance à extraire des jeux de tweets de crise exclusivement à partir de mots-clés déterminés à l'avance par les chercheurs, nous développons une méthodologie d'extraction lexicale mêlant approches supervisée et non supervisée. Cette méthodologie est fondée sur le raisonnement suivant : (1) à partir d'une poignée de mots-clés ou de hashtags définis, on peut extraire un premier jeu de tweets de crise depuis le jeu brut. (2) Ce premier jeu de tweets de crise contient les mots-clés ou hahstags recherchés mais également de nouveaux mots qui se trouvent associés au vocabulaire défini préalablement. (3) Ce jeu de crise est alors analysé par des outils de fouille de texte qui mettront en évidence ces mots ou hashtags destinés à enrichir la liste d'extraction appliquée au jeu de tweets bruts. (4) Le processus est itératif : dès que de nouveaux mots-clés sont ajoutés à la liste, on réinterroge le jeu de tweets bruts pour extraire l'ensemble des tweets de crise contenant les mots-clés et on effectue une nouvelle analyse de texte.

La mise en œuvre méthodologique de l'extraction lexicale.

D'un point de vue opérationnel, la figure 4.4 résume les étapes qui s'enchaînent pour la mise en œuvre de la recherche et de l'extraction lexicale, fondées sur l'exploration du contenu des tweets bruts.



(a). Etapes de recherche et de filtrage lexical



(b). Comportement théorique de saturation

Figure 4.4 : Etapes d'extraction et de recherche de nouveaux hashtags ou de nouveaux mots-clés (C.Cavalière)

Comme indiqué dans le paragraphe 4.2.2.1, les étapes préparatoires (en bleu sur la figure 4.4.a) requièrent une première connexion, depuis un ordinateur local, à la base de données *twitterdb* hébergée sur le serveur de collecte et de stockage des tweets. La requête d'extraction du jeu de tweets bruts, lancée par l'ordinateur local, repose sur deux critères : l'espace géographique concerné et la sélection d'une plage temporelle. La base de données *twitterdb* retourne un fichier au format CSV ensuite injecté dans une base de données locale sous le nom de table *tweet_brut*. La première requête d'extraction du premier jeu de tweets de crise est construite comme suit : dans un premier temps, une lecture rapide de la presse en ligne américaine et française nous révèle l'existence du surnom *Snowzilla*, couramment

employé pour désigner une tempête de blizzard aux Etats-Unis. La page *Wikipédia* indique l'existence de deux autres surnoms fréquemment donnés à ces tempêtes : *Snowmageddon* et *Snowpocalypse*[21]. Nous utilisons alors ces trois surnoms en tant que hashtags pour extraire un premier jeu de tweets filtrés à partir de la table *tweet_brut* (en vert sur la figure 4.4.a). Le premier jeu de tweets de crise extrait (table *tweet_crise_filtre1*) est ensuite importé dans l'interface *RStudio* et soumis à des étapes de fouille de texte (en rouge dans la figure 4.4.a) afin d'identifier les hashtags employés avec les trois hahstags soumis dans la première requête de filtrage lexical. La visualisation des résultats par le nuage de mots permet alors d'identifier de nouveaux hashtags comme *#blizzard2016, #jonas, #jonasstorm, #dcsnow, #snow, #snowing, #snowapocalypse*, *etc*. Ces nouveaux hashtags-clés sont alors ajoutés à la liste de filtrage et interrogent de nouveau la table *tweet_brut* afin d'extraire un deuxième jeu de tweets de crise (en vert sur la figure 4.4.a). L'étape de fouille de texte est alors répétée pour enrichir une nouvelle fois la liste d'extraction.

Quand arrêter le processus ? Si l'on se réfère de nouveau aux travaux de (Lin *et al*., 2013) sur la dynamique des hashtags, on peut supposer que les trois premiers hashtags (*#snowzilla*, *#snowmageddon* et *#snowpocalypse*) font certainement partie de la catégorie *Winners*, autrement dit les hashtags massivement adoptés (ici, par habitude d'utilisation de ces surnoms devant le phénomène). La première requête devrait donc retourner un maximum de tweets de crise. Nous supposons ensuite que la mise en évidence des hashtags concurrents mais non majoritaires augmentera le nombre de tweets de crise retournés, mais dans de moindres proportions, jusqu'à un palier où le nombre de nouveaux hashtags n'ajoute que des poignées éparses de tweets de crise. En fait, si l'on trace la courbe de nombre de tweets retournés en fonction du nombre et de l'ordre des hashtags donnés, on peut s'attendre à observer une courbe d'apparence logarithmique, à la manière de la courbe présentée dans la figure 4.4.b.

Les outils mobilisés pour la mise en œuvre de l'extraction lexicale.

Comme précisé ci-avant, la base de données locale, dans laquelle on importe le jeu de tweets bruts extraits depuis le serveur du LIG, est connectée à l'interface *RStudio* : dans cette interface, on importe uniquement la colonne *text* de la table *tweet_brut*. Les étapes de fouille de texte sont ensuite réalisées d'après la procédure suivante (figure 4.5) :

[21] Source : https://en.wikipedia.org/wiki/Snowmageddon (Consulté pour la dernière fois le 15/08/2019)

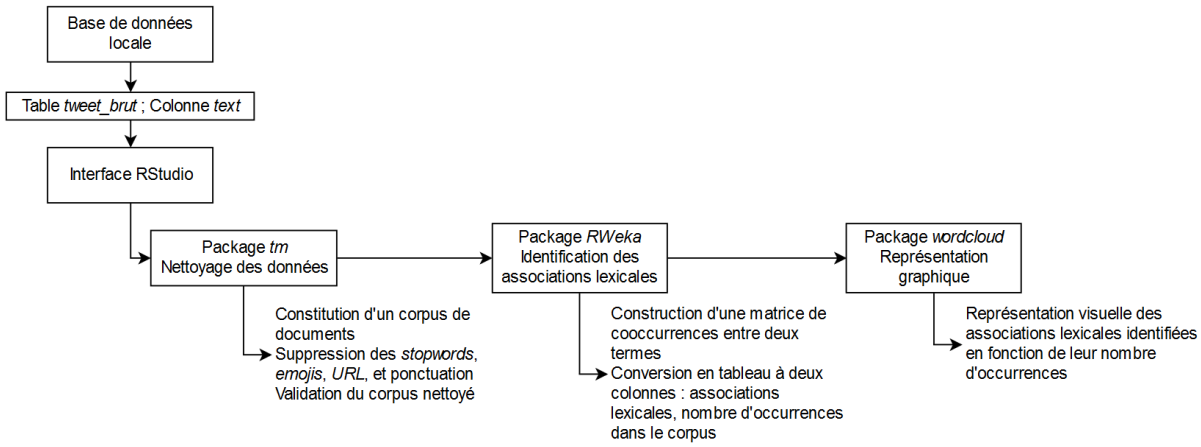


Figure 4.5 : Modules R mobilisés pour les étapes de fouille de texte

Le package *tm* de R est utilisé pour la fouille de texte : il convertit l'ensemble des données contenues dans notre import de tweets en un *corpus de documents* (chaque *document* étant un tweet). L'étape de nettoyage des documents consiste, dans notre cas, à supprimer les *stopwords* (soit les mots les plus courants d'une langue comme les pronoms ou les conjonctions), les *émojis*, les URL ainsi que la ponctuation[22]. En ce qui concerne la mise en évidence des associations lexicales permettant d'identifier les mots ou hashtags employés avec ceux qui ont été recherchés, nous utilisons le package *Rweka* : sa fonction *NGramTokenizer* séquencera les documents (ou tweets) en *n-grammes* en fonction des probabilités d'apparition entre *n* mots (dans notre cas, il s'agira de constituer des matrices en *bi-grammes*, autrement dit des associations de deux mots et leurs fréquences d'apparition respectives dans le corpus ; ces matrices sont ensuite converties en simples tableaux à deux colonnes qui contiennent les associations lexicales ainsi que leur nombre d'occurrences). Comme le soulignaient les deux figures précédentes, 4.4 et 4.5, les résultats sont visualisés sous forme d'un nuage de mots, construit par les fonctions du package *wordcloud*. Le code sémiologique du nuage de mots reste simple : plus la taille de la police est grande et la couleur foncée, plus l'association lexicale formée est fréquente dans le corpus de documents nettoyés.

Enfin, il nous faut également préciser la manière dont nous interrogeons le jeu de tweets bruts, dans la base de données locales, afin d'en extraire les tweets de crise à l'aide de la liste de mots-clés ou hahstags constituée. En effet, si nous avons recours à une requête SQL traditionnelle effectuée sur une chaîne de caractères, il nous faut combiner chaque mot-clé avec le caractère "%". L'inconvénient majeur de ce type de requête est qu'elle retournera les tweets dans lesquels le mot-clé recherché fait partie d'un autre mot. Par exemple, si l'on

[22] Le cas de la ponctuation pouvait être problématique étant donné qu'il nous faut conserver le signe "#" qui introduit tout hashtag. Le problème étant manifestement fréquent, des utilisateurs du forum *Stackoverflow* ont donné des exemples personnalisant la formule *removePunctuation* du package *tm* afin qu'elle ne supprime pas les signes "#". C'est cette formule que nous avons utilisée ici : https://stackoverflow.com/questions/27951377/tm-custom-removepunctuation-except-hashtag

considère la requête suivante "*SELECT * FROM tweet_brut WHERE text LIKE '%rain%'*", nous allons certes retourner tous les tweets contenant le radical *rain* et ses dérivés (comme *rainy* ou encore *raining*) mais également tous les mots qui contiennent cette même racine (*brain*, *train*, *etc.*). Afin d'éviter cet écueil qui introduit du bruit résiduel dans chaque jeu de tweets de crise, nous avons recours aux outils de *full-text search* (recherche de plein texte) de PostgreSQL, et en particulier aux fonctions de requête *to_tsvector()* et *to_tsquery()*, qui présentent l'avantage de se baser sur la racine du mot indiqué. La figure 4.6 ci-dessous présente leur syntaxe particulière :

```
SELECT id, tweet_id, text, created_at, gps, to_tsvector(text) @@ to_tsquery
        (
        'flood|storm|tornado'
        )
FROM tweet_brut_texas ;
```

(a). Requête avec l'opérateur | (OR)

```
SELECT id, tweet_id, text, created_at, gps, to_tsvector(text) @@ to_tsquery
        (
        'flood&victim'
        )
FROM tweet_brut_texas ;
```

(b). Requête avec l'opérateur & (AND)

Figure 4.6 : Exemples de requêtes effectuées avec les outils de recherche plein texte de PostgreSQL

La fonction *to_tsquery()*, encadrée en vert dans la figure 4.6.a, introduit une liste de mots-clés dont il faut vérifier la présence de la racine d'au moins un dans chaque tweet de la table *tweet_brut* (ici, nous avons saisi un total de trois mots-clés). La fonction *to_tsvector()*, encadrée en rouge dans la figure 4.6.a, introduit le nom du champ à tester (ici, il s'agit du champ *text* du tweet) : en d'autres termes, on demande à vérifier la présence d'au moins l'un des radicaux des trois mots-clés, tels qu'ils ont été saisis, dans la composante textuelle des tweets de la table brute. Par ailleurs, dans la syntaxe de la recherche de plein texte, l'opérateur SQL traditionnel *OR* est remplacé par le signe "|". Le résultat retourné par une requête de ce type contient une nouvelle colonne en codage binaire : la valeur T (*TRUE*) indique qu'au moins l'un des mots saisis dans *to_tsquery()* est présent dans la valeur du champ *text* ; dans le cas contraire, F (*FALSE*) indique qu'aucun des mots saisis n'est mentionné dans le champ *text*. Cette fonction permet également de rechercher des associations lexicales (cf. figure 4.6.b) : l'opérateur *AND* est alors remplacé par l'esperluette "&".

4.2.2.3. Application technique de l'approche spatiale

L'approche spatiale englobe bien évidemment les étapes et outils de l'approche lexicale, décrits dans le paragraphe précédent, auxquels s'ajoutent l'identification préalable d'objets d'intérêt sur le territoire et le tracé de régions d'intérêt autour de ces objets (cf. figure 4.2). Cette dernière approche inclut alors une primo-étape de recherche visant à identifier les éléments cibles du territoire, puis à constituer les régions d'intérêt pour la sélection des tweets. La figure 4.7 décrit les étapes à mettre en œuvre dans cette approche :

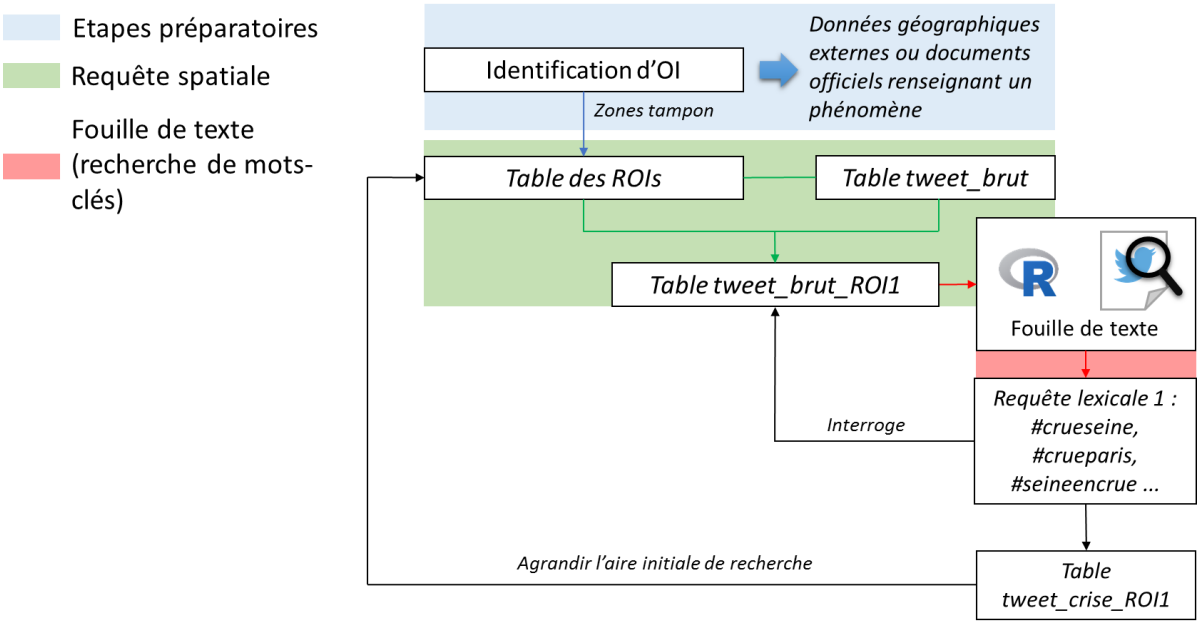


Figure 4.7 : Etapes de l'extraction de tweets de crise par l'approche spatiale et lexicale (C.Cavalière)

L'extraction spatiale de tweets de crise géolocalisés implique, dans un premier temps, la définition d'objets d'intérêt dans le territoire cible : nous considérons comme objet d'intérêt tout objet du territoire, localisable et cartographiable (sous un SIG par exemple), qu'il soit ponctuel, linéaire ou polygonal. Il peut donc s'agir aussi bien d'un bâtiment que d'une rivière ou encore d'une zone inondée. Ces informations peuvent être collectées par des données géographiques externes (tracés des cours d'eau, données de cumuls pluviométriques ou images satellites de territoires inondés, *etc.*), mais encore par de l'information textuelle qu'on peut valoriser en donnée géographique : par exemple, la presse en ligne peut indiquer les noms de quartiers inondés ou d'objets précis ayant subi des dommages, soit des informations qui peuvent être numérisées sous SIG.

Dans un deuxième temps, on génère, autour de ces objets d'intérêt, des régions d'intérêt délimitant un espace seuil permettant de sélectionner les tweets bruts contenus dans cet espace. La définition de ces régions d'intérêt peut se faire par la construction de zones tampon autour des OI ; là encore, le critère de distance dépendra de la résolution spatiale de l'étude, autrement dit, de l'échelle spatiale considérée et de ce que l'utilisateur

appréhende comme voisin ou distant d'un objet dans le contexte particulier d'une étude. Les tweets bruts inclus dans les ROI (et stockés dans une table ici intitulée *tweet_brut_ROI1*) sont alors soumis aux étapes d'analyse lexicale décrites dans le paragraphe 4.2.2.2 afin d'identifier l'existence d'un vocabulaire lié au phénomène en question[23]. De la même manière, en utilisant les outils de recherche de plein texte, on pourra alors interroger la table *tweet_brut_ROI1* afin d'extraire les tweets liés au phénomène d'étude dans une nouvelle table *tweet_crise_ROI1*. On peut alors renouveler le processus de sélection spatiale des tweets/analyse lexicale/extraction de tweets de crise en élargissant l'espace de recherche initial.

Les approches et outils introduits ici sont appliqués à des phénomènes plus ou moins violents et dommageables d'origine naturelle mais ils peuvent être étendus à tout phénomène ou tout événement *inhabituel* impliquant une échelle locale ou des interactions entre l'échelle locale de survenue physique et une échelle virtuelle qui s'agite au-delà du physique : les événements violents, sociaux, culturels ou encore sportifs pourraient ainsi également être testés avec cette proposition méthodologique.

4.2.3. Spécifications d'un environnement d'analyse exploratoire lexicale

Dans l'objectif d'assurer la reproductibilité de la démarche méthodologique de construction d'un jeu de tweets de crise, nous avons défini les spécifications d'un environnement interactif destiné à automatiser l'ensemble des étapes inclues dans les approches proposées. Ces spécifications se présentent sous la forme d'arbres de décision[24] et décrivent le déroulement des étapes à intégrer, avec leurs fonctionnalités et outils, dans l'interface. Cette section présente ainsi une démarche empirique qui est appliquée à un cas d'étude concret, lui-même décliné en fonction des approches proposées. Ce cas d'étude est focalisé sur les phénomènes de pluies-inondations survenus au Texas au printemps 2016.

4.2.3.1. Formalisation d'une démarche orientant le choix d'une méthode de recherche et d'extraction

Dans un premier temps, nous avons formalisé une démarche méthodologique d'extraction, qui s'adapte en fonction des objectifs de construction d'un jeu de tweets utiles ; cette question de l'adaptabilité de l'approche proposée aux besoins reprend la théorie construite et explicitée au paragraphe 4.2.1.1 sur l'articulation entre les types de phénomènes

[23] La différence qu'on pourra noter ici par rapport à la démarche expliquée dans le paragraphe 4.2.2.2 est la suivante : dans le cas de l'approche spatiale, la recherche de mots-clés est totalement non supervisée. Précédemment, on identifiait une poignée de mots ou de hashtags-clés afin d'enrichir progressivement la liste de vocabulaire d'extraction par la mise en évidence des associations lexicales ; ici, on analyse le contenu global des tweets sans filtrage sémantique préalable (celui-ci est remplacé par le filtrage spatial en fonction du critère de proximité à l'objet cible).

[24] Les schémas ont été réalisés avec l'application en ligne *draw.io* : https://www.draw.io/

(récurrents et locaux ou rares et globaux) ainsi que l'échelle géographique de l'étude (événement virtuel analysé à l'échelle locale ou globale). En conséquence, l'interface offrira deux possibilités en entrée : une méthodologie de recherche et d'extraction lexicale de tweets géolocalisés, et une méthodologie reposant sur le premier critère de la sélection spatiale. Dans les deux cas, le formalisme indique les prérequis indispensables à l'utilisation des approches proposées, de même que des exemples d'utilisations possibles (figure 4.8).

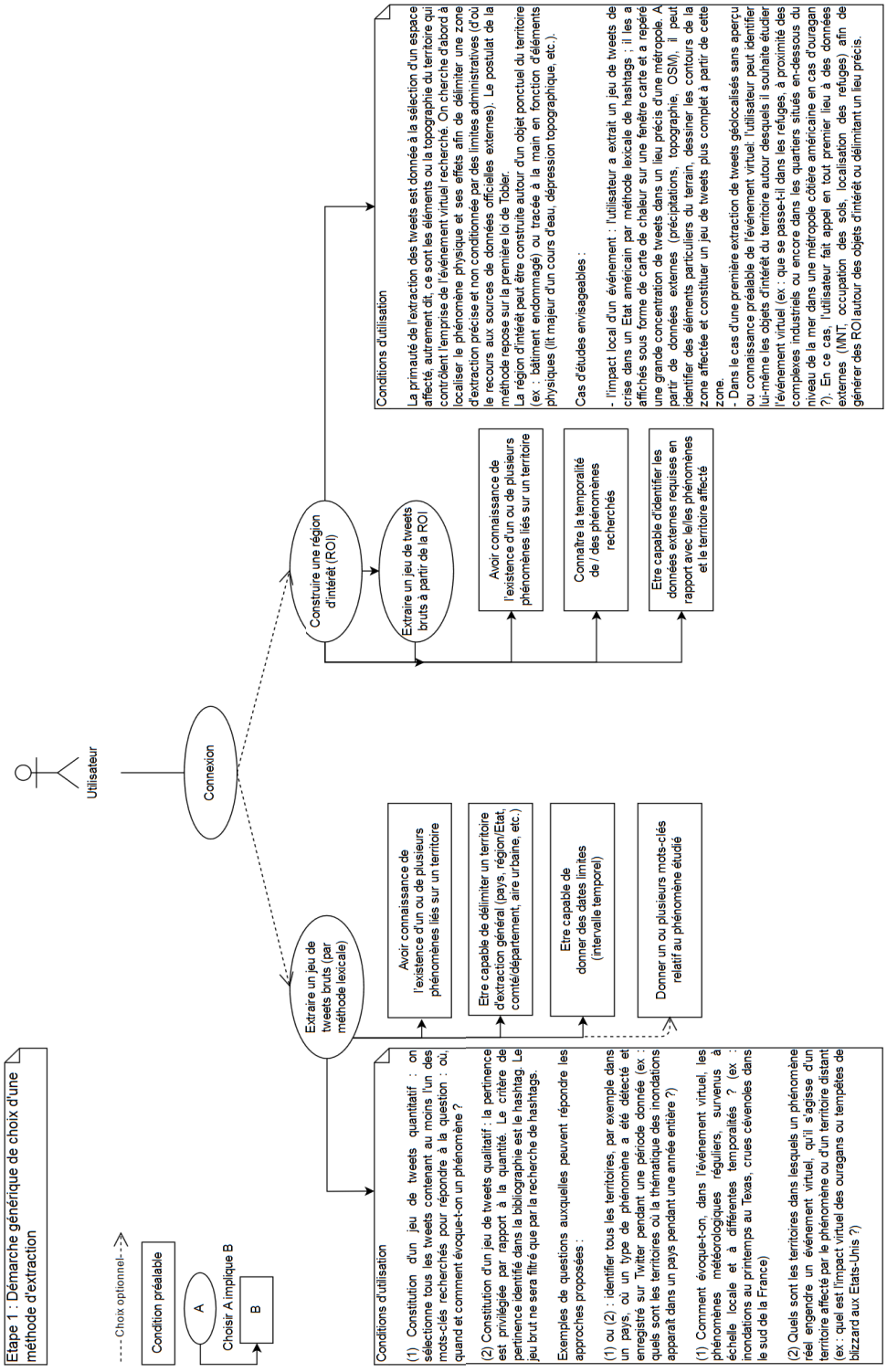


Figure 4.8 : Formalisation de la démarche introductive à la sélection d'une approche pour l'extraction de tweets utiles (C.Cavalière)

Le formalisme exposé se calque ainsi sur les trois facteurs (type, échelle du phénomène et pertinence des tweets) exposés dans le paragraphe 4.2.1.1, ayant orienté la définition des approches d'extraction. La méthode lexicale permet ainsi de collecter des tweets géolocalisés dans une entité administrative et dans un intervalle temporel fourni par l'utilisateur. Le formalisme reprend alors les cas d'applications envisageables en considérant les types de phénomènes et les échelles d'étude privilégiées (recherche par hashtags seuls pour un phénomène extrême à l'échelle d'un pays, recherche de l'ensemble des mots-clés liés à un phénomène à une échelle plus locale). Si la méthode lexicale seule n'est pas appropriée aux besoins de l'utilisateur, notamment en termes d'échelle géographique, c'est la méthode spatiale qui est privilégiée. Celle-ci fait intervenir la construction d'une région d'intérêt (ROI), qui est ici envisagée selon deux possibilités :

- (1) : L'appel à des données externes (par exemple via des services *WFS* fournissant des modèles numériques de terrain pour la topographie, des cartes d'occupation des sols, des données de précipitations, *etc.*) sert à identifier les objets du territoire ou les territoires locaux qui présentent un intérêt dans l'étude de la problématique (un bassin versant, le lit majeur d'un cours d'eau, un territoire urbain en dessous du niveau de la mer, *etc.*). Puis, il permet de tracer une ou plusieurs régions d'intérêts : une zone tampon autour d'un objet du territoire, un polygone tracé à la main pour délimiter une dépression topographique, *etc.*
- (2) : Dans un premier temps, l'exploration des tweets a été effectuée en fonction de la méthode lexicale seule : par exemple, on a recherché, aux Etats-Unis, tous les tweets qui contiennent le mot-clé *flood* émis sur l'année 2017. L'affichage de ces tweets sous forme de carte de chaleur a ensuite permis de repérer une grande concentration de points dans un lieu précis d'une grande métropole américaine. L'espace géographique d'intérêt est alors recentré dans ce point précis de la ville : une ROI est tracée autour du lieu identifié afin de lancer une nouvelle recherche lexicale de tweets[25], plus approfondie.

Dans tous les cas d'utilisation, l'environnement ne contraint pas l'utilisateur à fournir d'emblée une liste de mots-clés ou de hashtags pour extraire directement un premier jeu de tweets de crise : la recherche lexicale peut être accomplie selon une approche non supervisée (étant donné que l'interface intègre des outils de fouille de texte).

En outre, nous adressons une mise en garde relative aux contraintes consécutives aux disparités socio-spatiales d'usage des plateformes web génératrices de traces numériques géolocalisées (figure 4.9) : il s'agit bien évidemment des effets de densités de population ou encore de l'accessibilité au numérique, qui peuvent très vraisemblablement entraver des analyses à très grande échelle. L'objectif consiste à informer l'utilisateur de la possibilité, à

[25] Nous ne souhaitons pas que le résultat final de l'utilisateur soit contraint par le choix initial d'une méthode ; c'est pourquoi il est également essentiel, dans le développement de l'interface, que l'utilisateur ayant choisi une méthode d'extraction lexicale, puisse avoir accès, à un moment donné, à des outils de sélection spatiale et de tracé de ROI.

une échelle géographique fine, qu'un phénomène n'ait pas été enregistré sur le réseau, en fonction du territoire considéré.

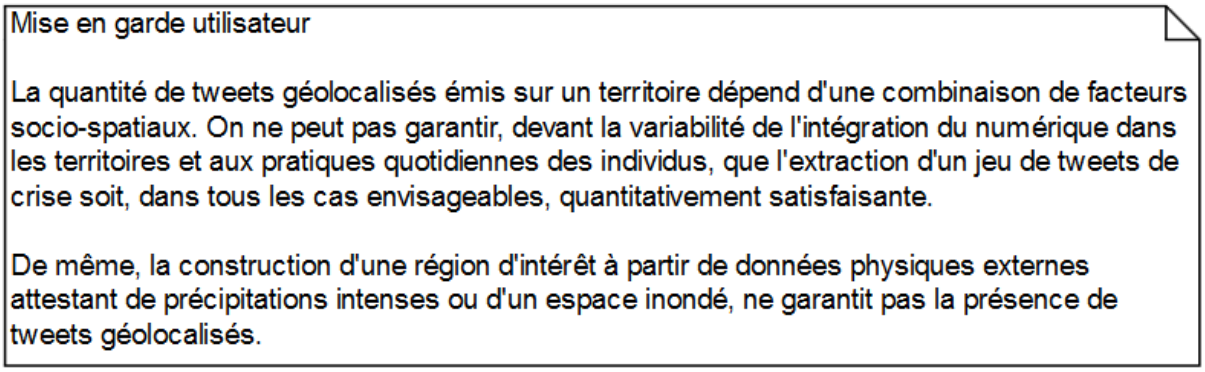


Figure 4.9 : Mises en garde relatives à la variabilité spatiale des émissions de tweets géolocalisés

Comme le sous-entend le formalisme proposé, cette interface est fondée en premier lieu sur l'exploration sémantique des tweets géolocalisés, qui s'articule autour de l'objectif suivant : construire le jeu de tweets de crise le plus pertinent (tri lexical par hashtags ou critère de distance à un objet) ou le plus complet possible (tri lexical par mots-clés) en fonction du phénomène et de l'échelle géographique considérée. L'utilisateur de l'interface est ainsi amené à adopter une posture active vis-à-vis des données : il adapte l'approche d'extraction à ses objectifs et au terrain, choisit les sources de données complémentaires, et peut importer ses propres couches d'information géographique dans l'interface le cas échéant.

4.2.3.2. Du formalisme à l'opérationnalisation de la démarche

Paramétrage initial de l'extraction lexicale.

En premier lieu, nous présentons les fonctionnalités intégrées pour la méthode d'extraction lexicale seule. La première fenêtre de paramétrage de l'extraction lexicale propose les fonctionnalités suivantes (figure 4.10) : la saisie des dates de filtrage des tweets (date de début et date de fin qui doivent être connues par l'utilisateur) ainsi que la zone géographique concernée. Pour ce dernier paramètre, on réserve trois possibilités distinctes :

- le tracé d'une *bounding box* délimitant le territoire d'étude ;
- la saisie manuelle du nom de(s) l'entité(s) géographique(s) d'intérêt ;
- l'import d'une couche d'information géographique contenant les données relatives à cette/ces même(s) entité(s) d'intérêt (par exemple un fichier SIG *shapefile*).

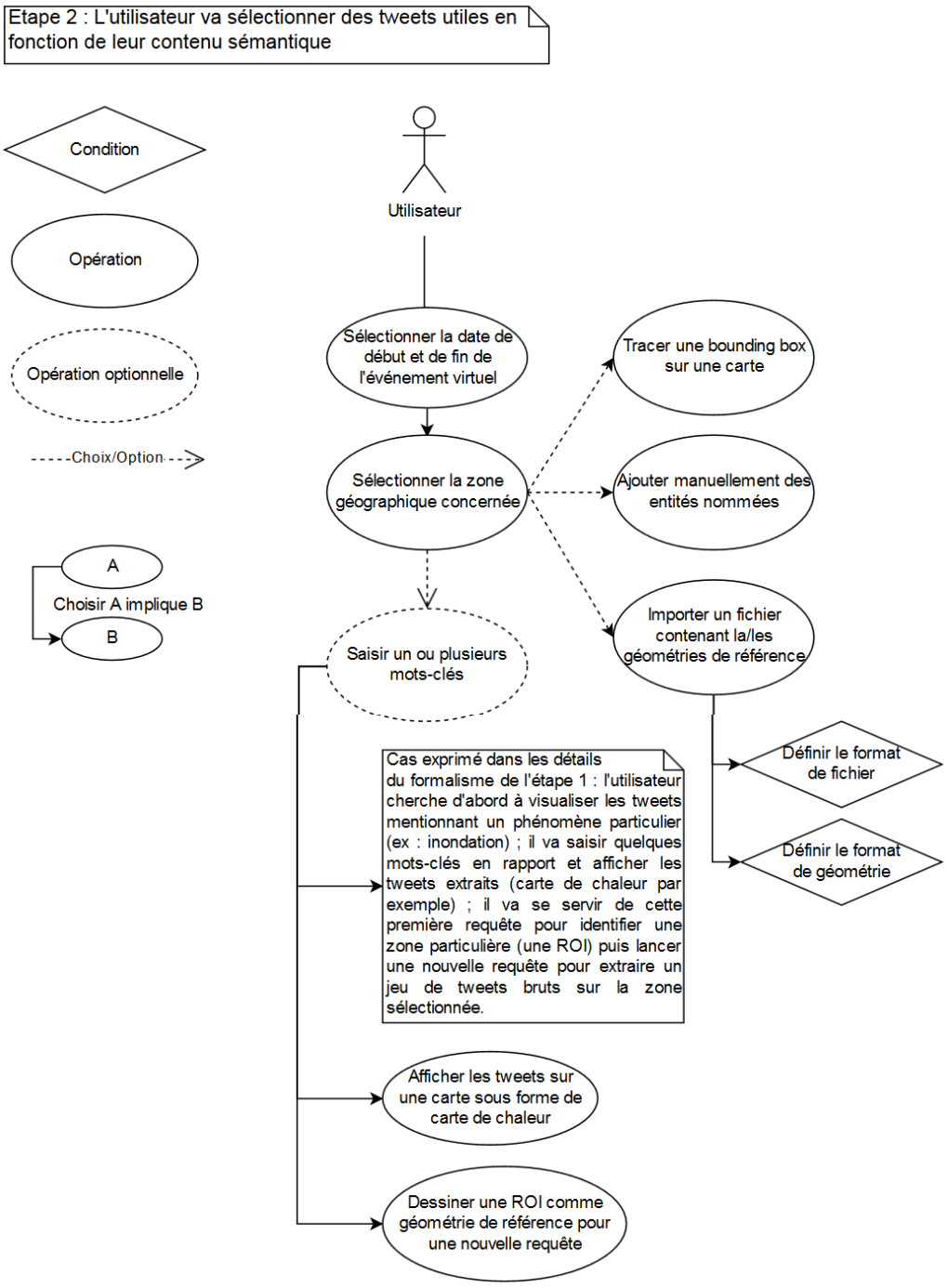


Figure 4.10 : Premières fonctionnalités de l'extraction lexicale (C.Cavalière)

Comme indiqué précédemment, l'application d'un premier filtrage lexical n'est pas une contrainte préalable ; dans tous les cas, les tweets sélectionnés par ces premiers paramètres sont affichés sur une carte interactive. Comme explicité dans les cas d'utilisation envisagés, l'interface doit permettre un basculement en mode d'extraction spatiale si l'extraction lexicale ne constituait qu'une *primo-recherche*, pour l'utilisateur, afin d'identifier un territoire d'étude plus précis.

La figure 4.11 présente la forme que peut prendre cette première fenêtre. Elle est constituée d'un panneau inférieur de saisie des paramètres de sélection des tweets géolocalisés et d'une carte interactive (zoom, navigation à la main, choix du mode de représentation des tweets et des couches à afficher) restituant les résultats de la requête soumise par l'utilisateur (dont le nombre de tweets retournés). Comme indiqué précédemment, elle intègre également un accès à des données externes et la possibilité de tracer des ROI. Ici, étant donné que la maquette est réalisée à partir du logiciel QGIS, les outils représentant ces fonctionnalités correspondent respectivement, dans la figure 4.11, à  et  . Les dates sont saisies à la main au format standard *année-mois-jour*, qui remplace les curseurs temporels intégrés aux interfaces de géovisualisation présentées dans le chapitre 3, pour deux raisons : la plage temporelle sélectionnée par l'utilisateur peut être plus ou moins grande (une seule journée, quelques mois, une année entière voire plusieurs années) ; par ailleurs, la vocation exploratoire de l'interface est articulée autour de la composante sémantique, et non autour de l'étude de la résolution spatiale de la réactivité des utilisateurs face à un phénomène.

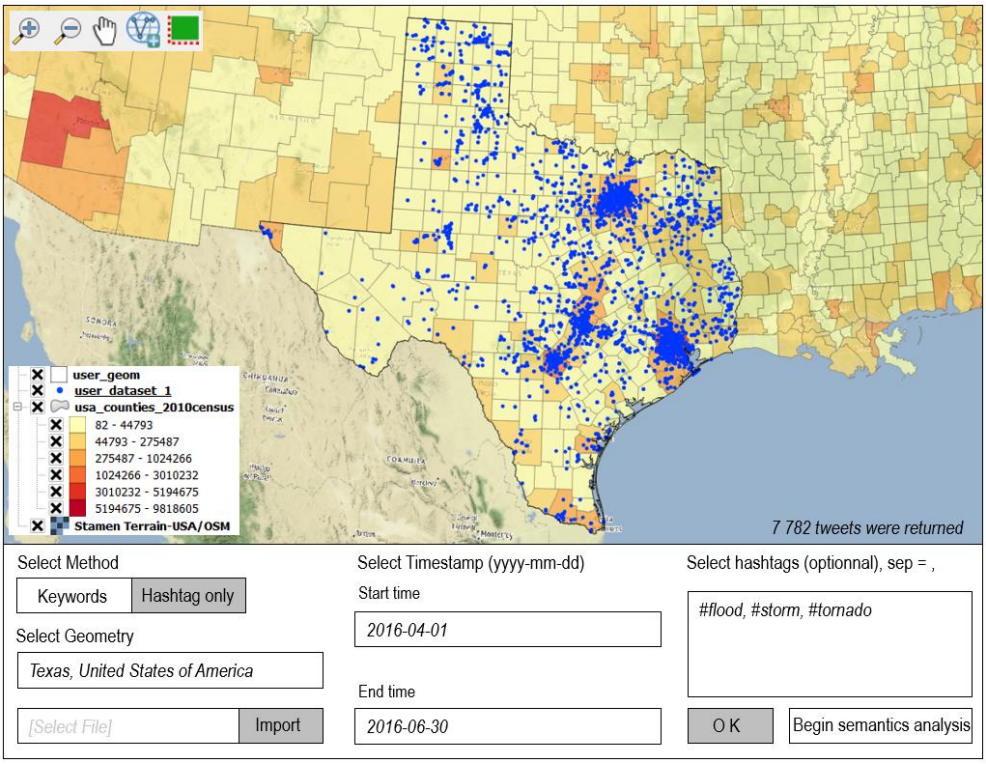


Figure 4.11 : Fenêtre d'accueil de l'extraction de tweets de crise par méthode lexicale (C.Cavalière)

Etapes de l'extraction lexicale par recherche de hashtags.

La recherche lexicale pouvant être focalisée soit sur les hashtags, soit sur les mots-clés toutes catégories confondues, nous intégrons ces deux approches à l'interface (cf. figure 4.8).

La figure 4.12 détaille alors les fonctionnalités de recherche et de visualisation associées à la méthode d'extraction lexicale par hashtag.

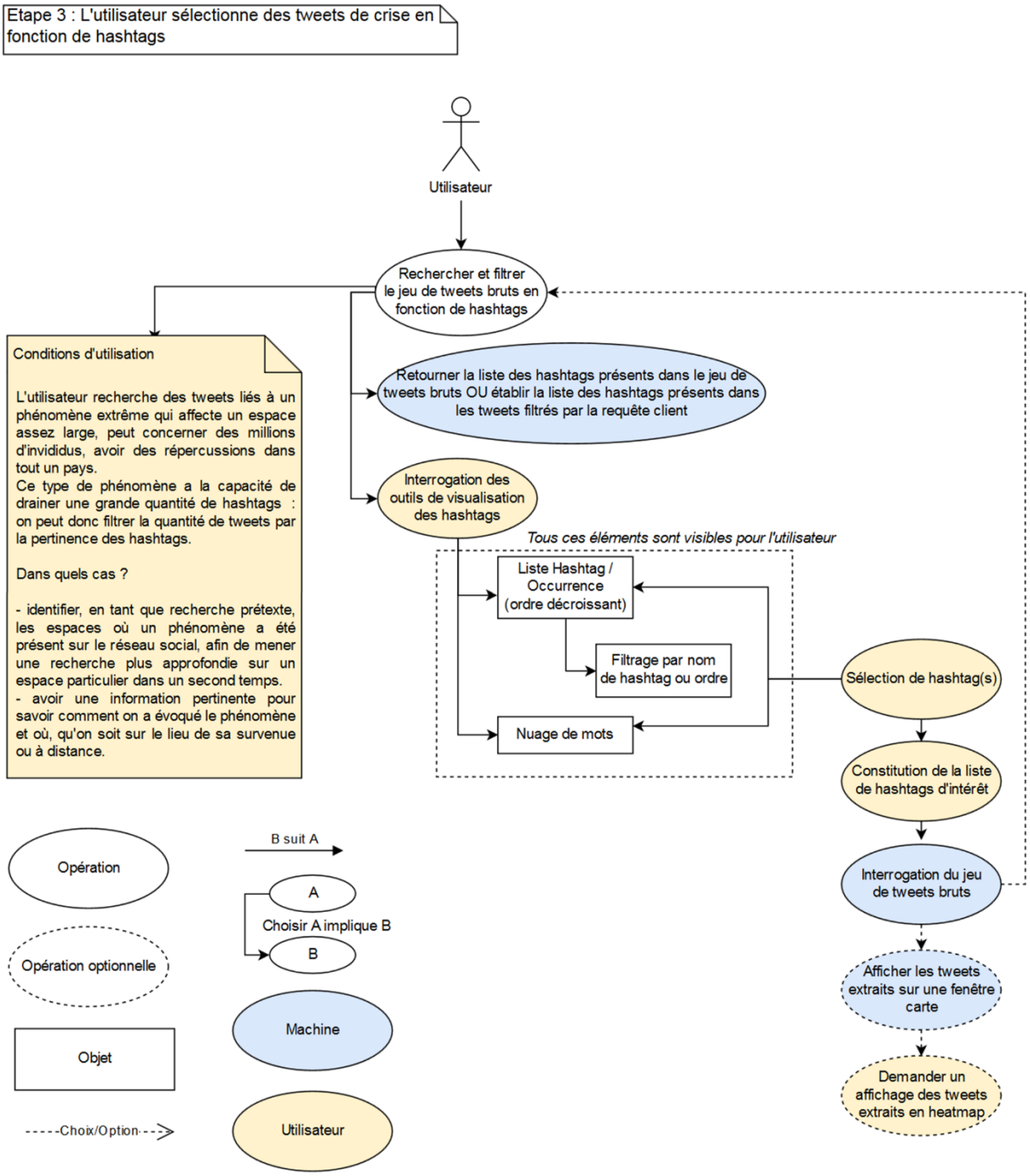


Figure 4.12 : Etapes de recherche et d'extraction lexicale par hashtags (C.Cavalière)

Dès que la première requête de filtrage des tweets, avec ou sans mention de hashtags recherchés, est soumise, le serveur doit retourner la liste de l'ensemble des hashtags présents dans le jeu de tweets filtrés. Comment ? La base de données *twitterdb* contient déjà deux tables d'identification des hashtags présents dans les tweets géolocalisés (figure 4.13) : la

table *tweet_hashtag* (qui contient les champs *tweet_id* et *hashtag_id* pour les numéros d'identifiant unique des tweets et des hashtags) et la table *hashtags* (*hashtag_id* et *name* pour le nom du hashtag). En effectuant une jointure attributaire, on peut créer une table intermédiaire de deux champs (*tweet_id* et *name*) contenant l'ensemble des hashtags présents dans un tweet et ensuite créer une table finale dans laquelle on concatène les caractères du champ *name* qu'on regroupe en fonction des valeurs de *tweet_id*[26].

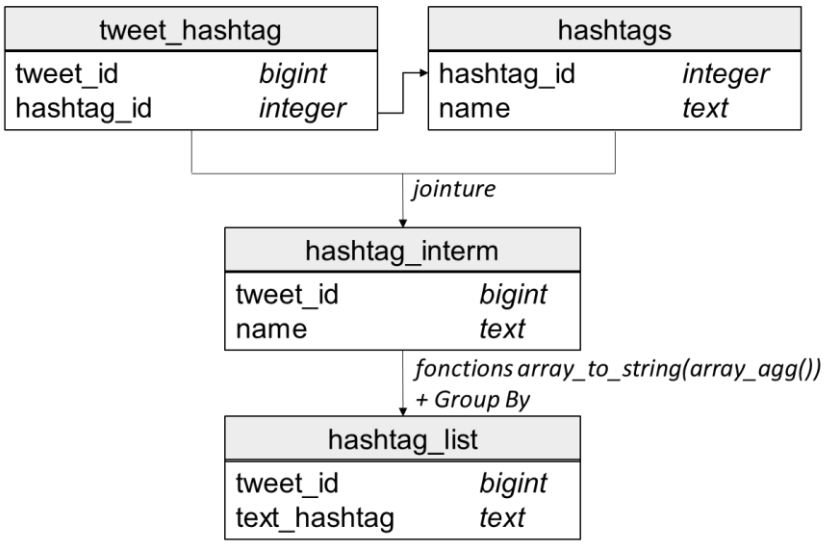


Figure 4.13 : Modélisation de la construction de la liste des hashtags, côté serveur

Lorsque l'échelle et le territoire d'étude sont fixés, les processus de l'analyse lexicale peuvent être lancés (cf. bouton *Begin semantics analysis* de la figure 4.11). L'interface doit ainsi embarquer l'ensemble des modules et fonctionnalités de fouille de texte de R décrits dans le paragraphe 4.2.2.2[27]. La visualisation des hashtags présents dans le jeu de tweets renvoyés par une requête est envisagée sous deux formes (cf. figure 4.12) :

- une liste de hashtags : on peut restituer une simple liste de l'ensemble des hashtags présents dans le jeu de tweets filtrés par la requête initiale, accompagnée de leur occurrence. Si l'on réalise des listes en *bi-grammes*, on pourra alors retourner cette même liste en fonction des collocations identifiées et de leur fréquence. Pour faciliter la navigation, la liste peut être filtrée en fonction de la recherche d'un radical précis ou en fonction de l'occurrence des hashtags. C'est ce que montre la figure 4.14. Ici, on a généré une simple liste de l'ensemble des hashtags présents dans la première requête exprimée dans la figure 4.11, et de leur fréquence. L'ajout de nouveaux hashtags de filtrage de tweets de crise peut simplement

[26] La requête exécutée par le serveur est alors de ce type : SELECT tweet_id, array_to_string(array_agg(name), ' ') as text_hashtag FROM table_interm GROUP BY tweet_id ;

[27] Dans le cas de l'extraction par hashtag uniquement, les fonctions de nettoyage de documents s'avèrent inutiles, étant donné que la représentation sémantique est effectuée à partir de la concaténation des différents hashtags présents dans chaque tweet et non sur le texte entier du tweet, cf. figure 4.13.

s'effectuer par sélection de hashtags d'intérêt dans la liste, et dépôt dans le cadre intitulé *Add hashtags to query list*.

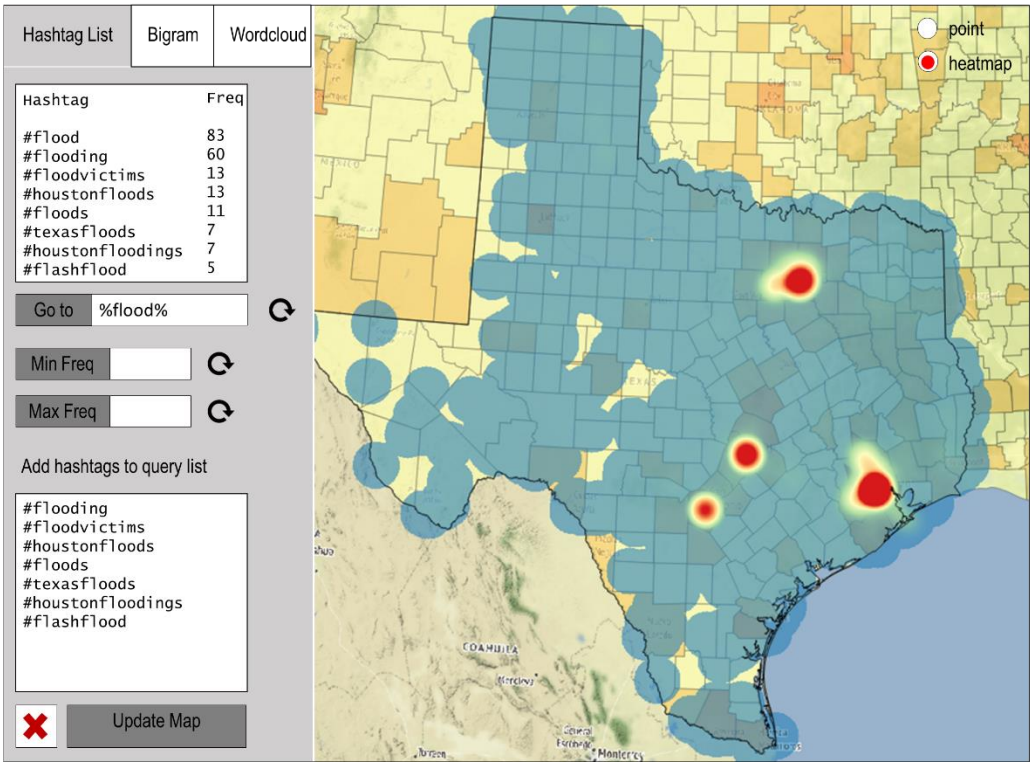


Figure 4.14 : Volet de visualisation des résultats de la recherche lexicale par hashtags (C.Cavalière)

- un nuage de hashtags : il peut afficher soit des hashtags simples dont la taille varie en fonction de la fréquence dans le jeu de tweets filtrés, soit des collocations entre hashtags si le client a effectué une analyse lexicale en *bi-grammes*. De la même manière que pour la liste de hashtags, le nuage doit être interactif : on peut ainsi sélectionner un hashtag du nuage et l'ajouter à la liste de vocabulaire d'extraction de tweets de crise.

A la fin de l'étape d'analyse lexicale, la nouvelle requête est soumise : les clauses temporelles et spatiales restent identiques mais la clause sémantique prend en compte l'ensemble des hashtags qui ont été saisis depuis la première requête (par exemple, ici, la deuxième requête doit retourner les tweets émis dans l'Etat du Texas, entre le 1er avril et le 30 juin 2016 et qui contiennent au moins l'un des hashtags-clés suivants : *#flood*, *#storm, #tornado* [cf. figure 4.11, requête initiale]*, #flooding, #floodvictims, #houstonfloods, #floods, #texasfloods, #houstonfloodings* et *#flashflood* [cf. figure 4.14, deuxième requête]). L'opération de recherche et de sélection de hashtags est renouvelée, sauf si l'utilisateur change de méthode d'extraction ou s'il ne souhaite pas poursuivre les étapes de recherche et d'extraction de hashtags-clés.

Etapes de l'extraction lexicale par recherche de mots-clés, toutes catégories confondues.

Dans ce deuxième cas, on considère toujours les phénomènes de pluies-inondations du Texas au printemps 2016, mais on élargit l'extraction lexicale à l'ensemble des catégories de mots. Le principe de fonctionnement de la méthode de recherche et d'extraction par mots-clés reste identique : en ayant défini un territoire, une plage temporelle et éventuellement saisi des entrées lexicales, on enrichit progressivement le jeu de tweets filtrés par fouille de texte dans les tweets retournés par les requêtes soumises au serveur (figure 4.15).

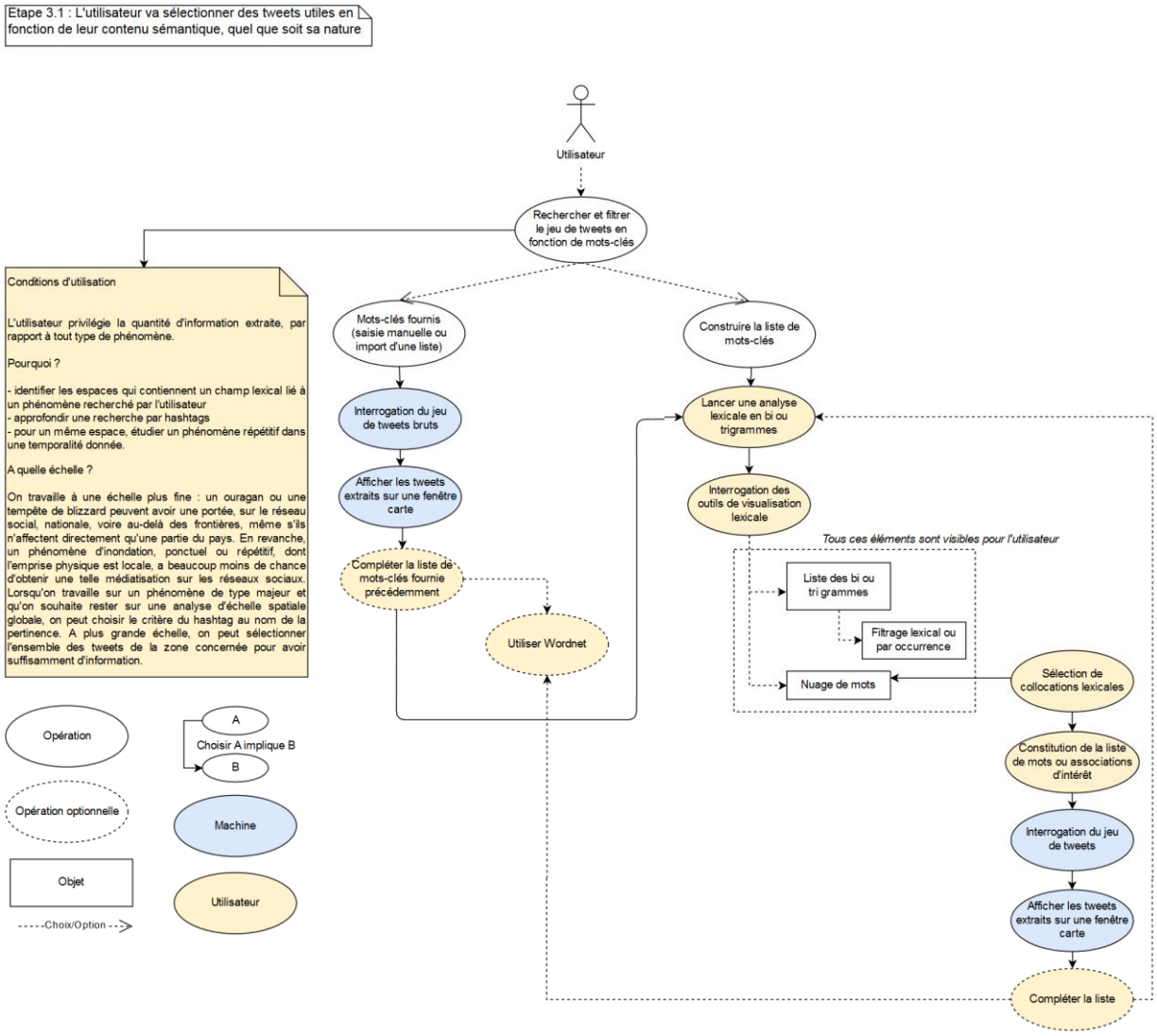
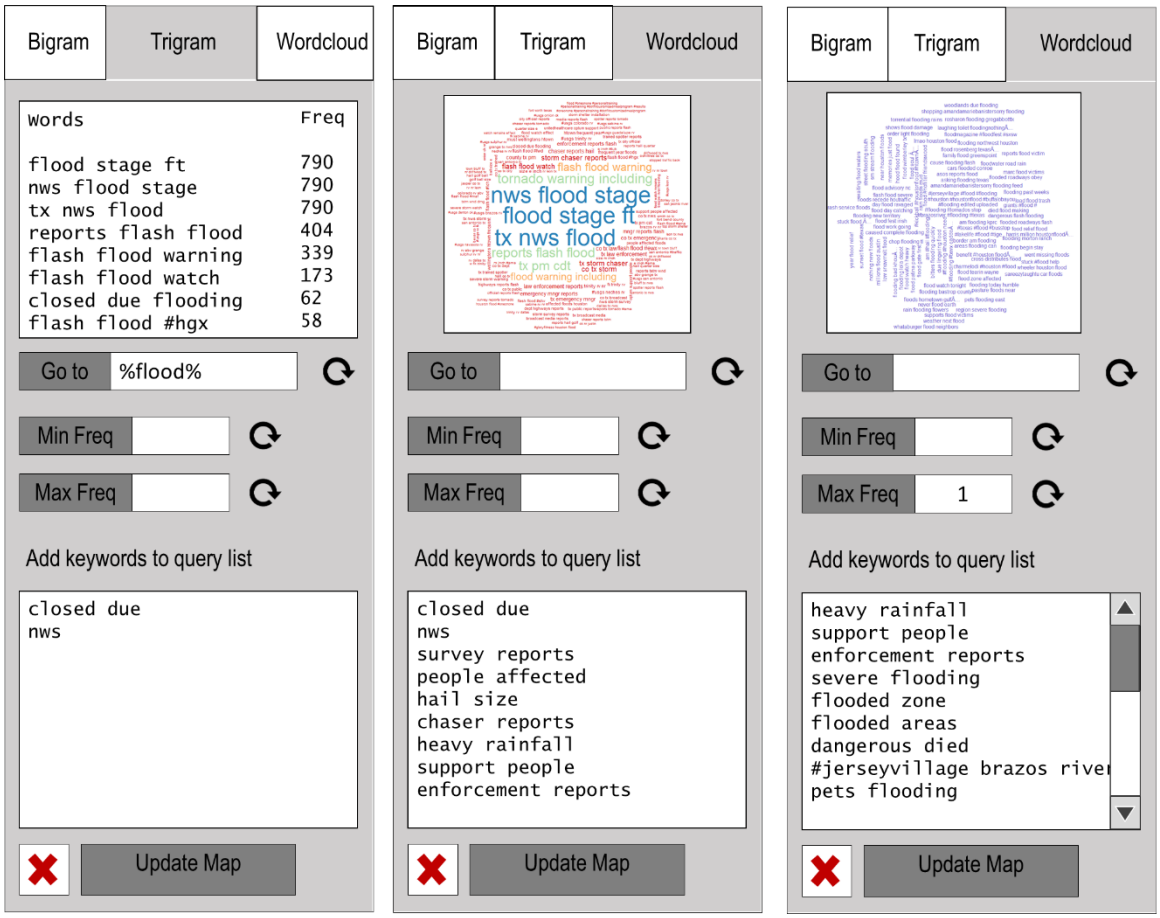


Figure 4.15 : Etapes de recherche et extraction lexicale par mots-clés, toutes catégories confondues (C.Cavalière)

L'accès au volet d'analyse lexicale s'effectue de nouveau après la soumission de la première requête (ses paramètres sont définis de la même manière qu'indiquée sur la figure 4.11 : sélection d'un territoire, sélection d'une plage temporelle et ajout optionnel de mots-clés). La liste de mots-clés peut alors être complétée ou constituée (dans le cas où aucun mot n'a été saisi dans la première requête). A la différence de l'approche focalisée sur les hashtags, l'analyse et la représentation lexicale sont ici envisagées par défaut sous forme de *bi-grammes*

ou de *tri-grammes* (et non en triant les mots un par un, comme on pouvait restituer les hashtags seuls). Alors que le hashtag permet d'identifier un contexte précis, le mot-clé peut être employé dans divers contextes ; par les *bi-* ou *tri-grammes*, on met en évidence des associations lexicales, autrement dit on peut vérifier qu'un mot susceptible d'être associé à divers contextes est ici employé avec des mots qu'on peut indubitablement relier au phénomène étudié. Par exemple, l'analyse simple peut mettre en évidence le mot *power* qui s'avère polysémique. En construisant un *tri-gramme*, on observe que le mot *power* est associé aux mots *cut* et *storm*. En lançant une recherche lexicale intégrant ces trois mots, on peut alors acquérir la certitude que le mot *power* sera employé dans le contexte approprié. La figure 4.16 présente trois exemples de résultats d'analyse lexicale, par liste d'associations en *tri-grammes* et nuages de mots construits selon ces mêmes *tri-grammes*. On considère, dans cet exemple, que la requête initiale contient des clauses identiques à celles que nous avons indiquées pour les hashtags : *Texas, United States of America* pour le lieu et du *2016-04-01* au *2016-06-30* pour l'intervalle temporel. Mais ici, la clause lexicale est focalisée sur les mots *flood*, *storm* et *tornado*.



(a). Liste des associations en tri-grammes

(b). Nuage des associations en tri-grammes

(c). Nuage des associations éparses en tri-grammes

Figure 4.16 : Volet de visualisation des résultats de l'analyse lexicale par mots-clés (C.Cavalière)

La liste affichée dans le volet de gauche (a) présente les associations mises en évidence contenant le radical *flood* ainsi que leur fréquence. Le volet central (b) représente cette même liste sous forme d'un nuage d'associations lexicales (on retrouve ainsi les associations *nws flood stage*, *flood stage ft* et *tx nws flood* comme dans la liste [a]). Le nuage de droite (c) se focalise sur l'exploration des associations peu fréquentes identifiées dans le jeu retourné par la requête initiale (fréquence maximale de l'association définie à 1). On attache ici un intérêt à ces associations marginales pour trois raisons :

- elles permettent d'identifier des objets ou lieux d'intérêts qui peuvent servir de support à une extraction spatiale : ici, on pouvait notamment relever des hashtags de noms de quartiers comme *#jerseyvillage*, *greenpoint* ou encore *rosenberg*, mais également des noms de rivières et fleuves (*#buffalobayou*, *#brazosriver*) ;
- les associations lexicales les plus fréquentes paraissent avoir une consonance officielle (par exemple, *nws flood stage* fait directement référence au *National Weather Service*, soit à une institution officielle inscrite et active sur le réseau virtuel). *A priori*, ce n'est pas par ce type de tweets qu'on pourra collecter de l'information sur les comportements des individus en situation de crise.
- la première étape de l'analyse lexicale est ici effectuée sur les tweets retournés par trois mots-clés. Or, on peut penser que des mots comme *rain*, *dangerous* ou encore *damaged* (qui apparaissent dans le nuage d'associations lexicales éparses, [c]) peuvent être employés dans de nombreux tweets sans pour autant être nécessairement associés aux trois mots-clés recherchés dans tous les tweets en rapport avec les phénomènes étudiés.

De la même manière que pour la recherche focalisée sur les hashtags, la sélection de mots-clés s'effectue par glisser-déposer (notons que l'utilisateur n'est pas contraint à indiquer exclusivement des associations de deux ou trois mots à rechercher ensemble dans les tweets). Par ailleurs, l'interface proposera l'accès à un nouvel outil de recherche lexicale, non exploré dans les sections précédentes : il s'agit de la base de données lexicales *Wordnet*[28] qui permet, entre autres, de rechercher les synonymes et collocations d'une entrée lexicale donnée (elle n'est cependant disponible que pour la langue anglaise). Le package *wordnet* connecte l'interface de R à la base de données *Wordnet* : la figure 4.17 affiche les résultats donnés par l'application de la fonction *synonyms()* sur les entrées lexicales *flood*, *storm* et *tornado* et en recherchant des noms. Ici, on pourrait par exemple ajouter, en rapport à l'entrée lexicale *flood*, les mots *deluge*, *overflow*, et *outpouring*.

```
> synonyms("storm", "NOUN")
[1] "storm"         "tempest"       "violent storm"
> synonyms("flood", "NOUN")
 [1] "alluvion"    "deluge"      "flood"       "flood lamp"  "flood tide"  "floodlight"  "flowage"     "inundation"  "outpouring"  "overflow"    "photoflood"
[12] "rising tide" "torrent"
> synonyms("tornado", "NOUN")
[1] "crack"   "tornado" "twister"
```

Figure 4.17 : Recherche de synonymes avec la base de données lexicales Wordnet

[28] Source : https://en.wikipedia.org/wiki/WordNet

Lorsque la liste de mots est complétée, une nouvelle requête est lancée et interroge les tweets bruts ; l'étape de recherche et de sélection de mots-clés peut alors être renouvelée ou interrompue.

Etapes de mise en œuvre de l'extraction spatiale.

La dernière méthode proposée et décrite correspond à l'extraction spatiale. Pour rappel, elle s'effectue plutôt à grande échelle géographique en ayant identifié des objets ou régions d'intérêt du territoire affecté et en construisant une région d'extraction des tweets plus précise qu'une simple mention de lieu. La figure 4.18 reprend cette étape de délimitation de la région d'intérêt. L'extraction spatiale peut être utilisée dans les deux cas de figure suivants : avant l'exécution de toute analyse lexicale si l'existence d'un phénomène localisé est déjà connue, ou après l'exécution d'une première étape d'analyse lexicale seule ayant souligné l'existence d'un objet ou territoire d'intérêt. Précédemment, nous avons identifié, par analyse lexicale de mots-clés, les hashtags *#brazosriver*, *#jerseyvillage* ou encore *#buffalobayou* : pour ce dernier cas d'étude, l'analyse peut alors être recentrée sur un lieu ou un objet particulier du territoire global texan.

Pour assurer l'extraction spatiale, le volet cartographique doit intégrer un certain nombre d'outils de tracé : la possibilité de numériser manuellement des polygones, de marquer un objet du territoire en ajoutant un point, de générer une zone tampon de *n* mètres autour d'une entité et éventuellement de sélectionner une entité afin de la supprimer en cas d'erreur. Le dessin des régions d'intérêt est effectué à partir de la consultation de données externes (du simple fond de carte permettant de marquer des objets d'intérêt sur le territoire aux données physiques mises à disposition sur les services Web). Au final, il sera alors possible de construire une ROI constituée d'un seul ou de plusieurs polygones (l'interface devra alors ajouter les polygones tracés par l'utilisateur dans la base de données, afin d'extraire les tweets croisant ces polygones). L'étape suivante consistera alors à lancer l'analyse lexicale par mots-clés (les fonctionnalités et la méthode mobilisées sont alors celles qui ont été présentées précédemment).

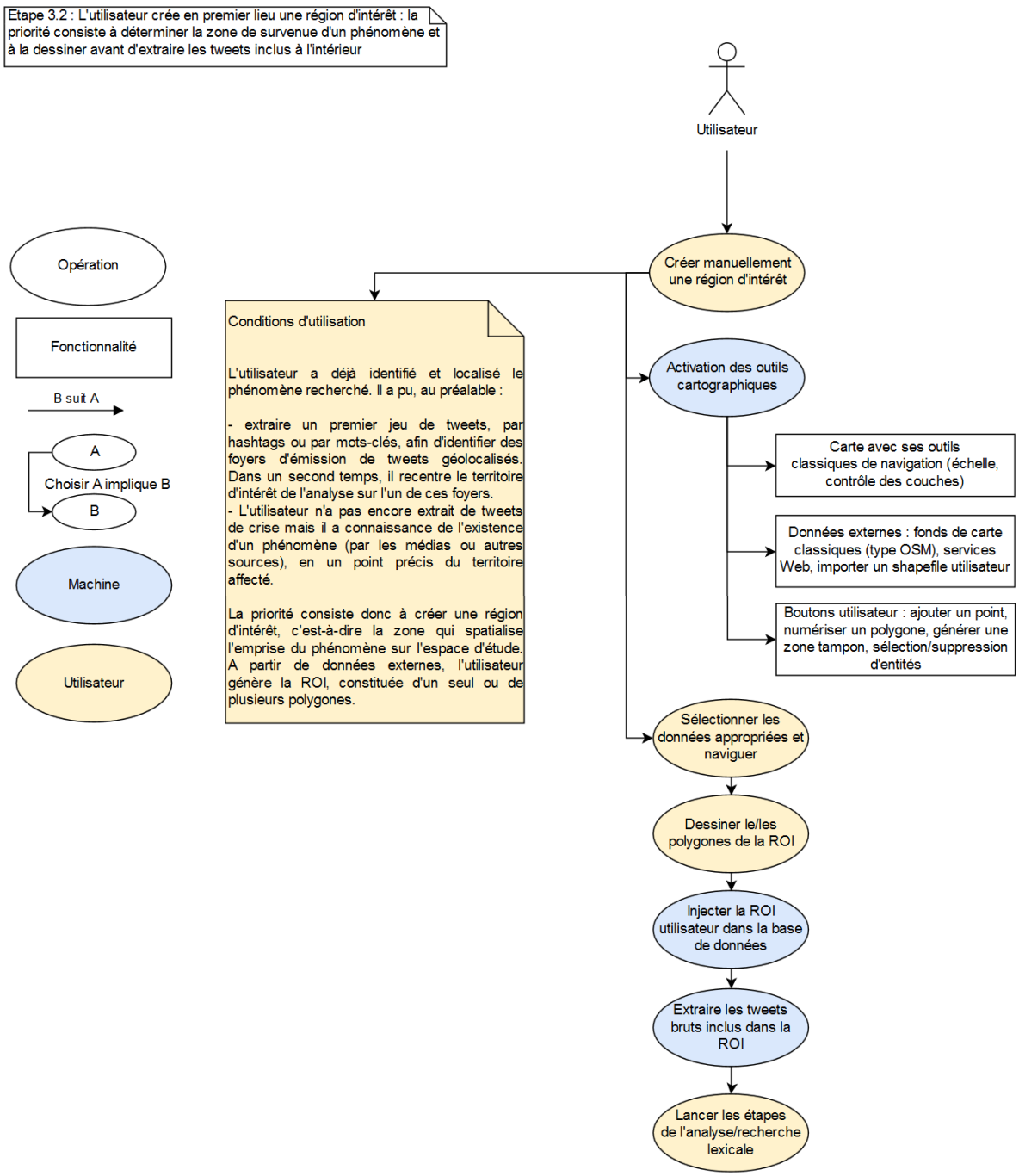


Figure 4.18 : Etapes de l'extraction lexicale - Construction de la région d'intérêt (C.Cavalière)

La figure 4.19 présente l'interface de la fenêtre spatiale, paramétrée selon le cas d'étude indiqué plus haut (l'objet d'intérêt local a été identifié par la première recherche lexicale effectuée sur les mots-clés lors de l'étape précédente). Dans l'exemple soumis par cette figure, un *shapefile* contenant les principaux cours d'eau de l'Etat du Texas est importé puis filtré par une simple requête SQL afin de ne sélectionner que la *Brazos River*. La ROI est définie en appliquant une zone tampon d'un kilomètre de chaque côté du tracé du cours d'eau. Ayant connaissance d'un phénomène de pluies-inondations survenu entre le 18 et le 19 avril 2016,

on indique ces deux dates de référence comme clauses temporelles dans la requête initiale. En revanche, on ne fournit ici aucun mot-clé de filtrage lexical.

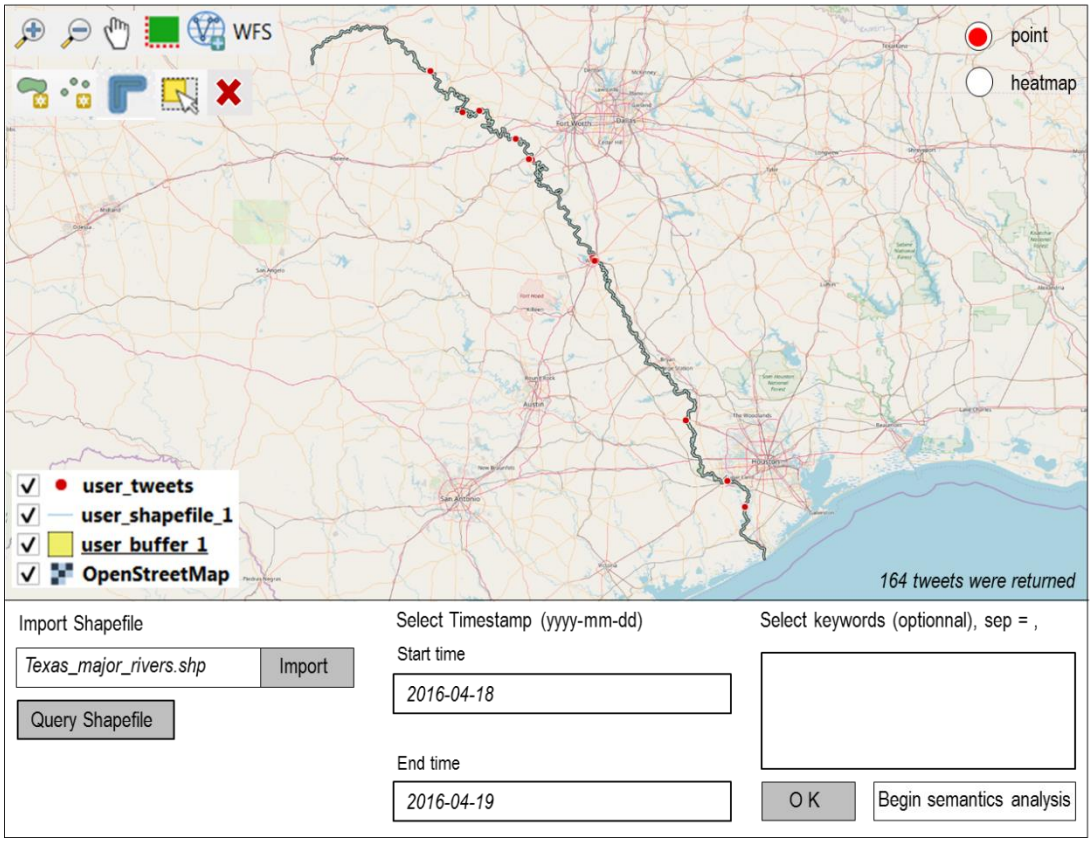
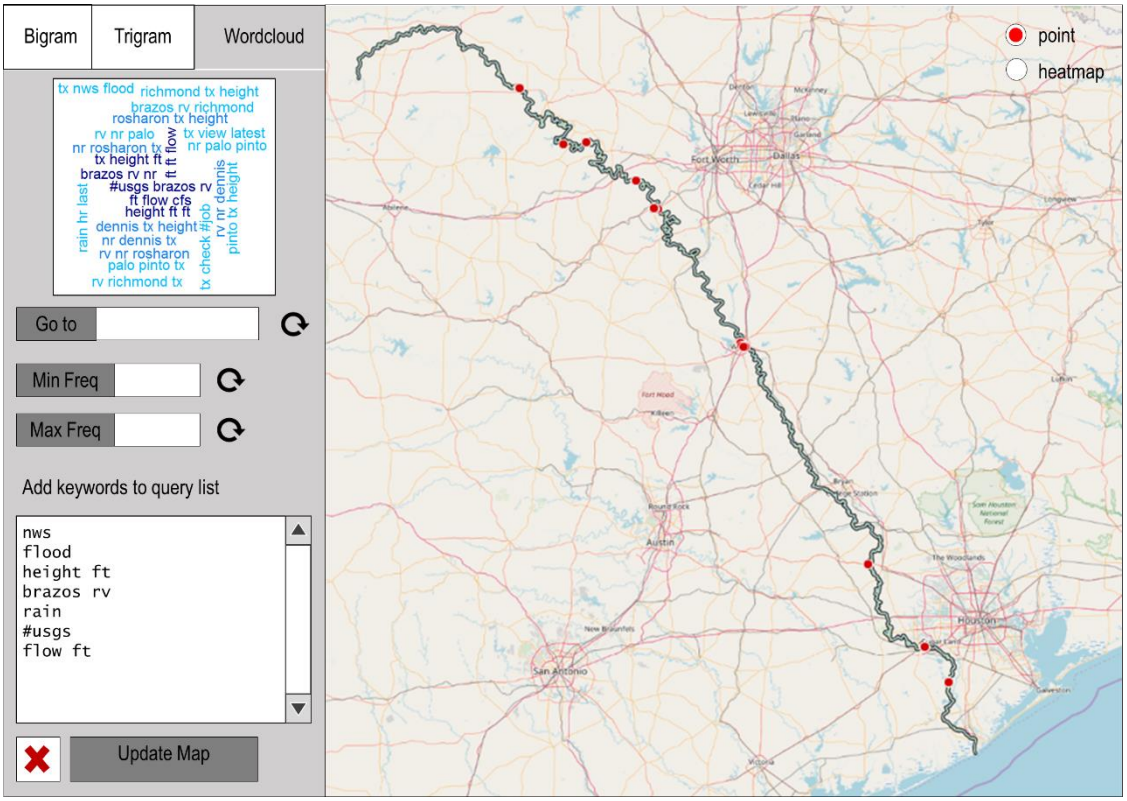


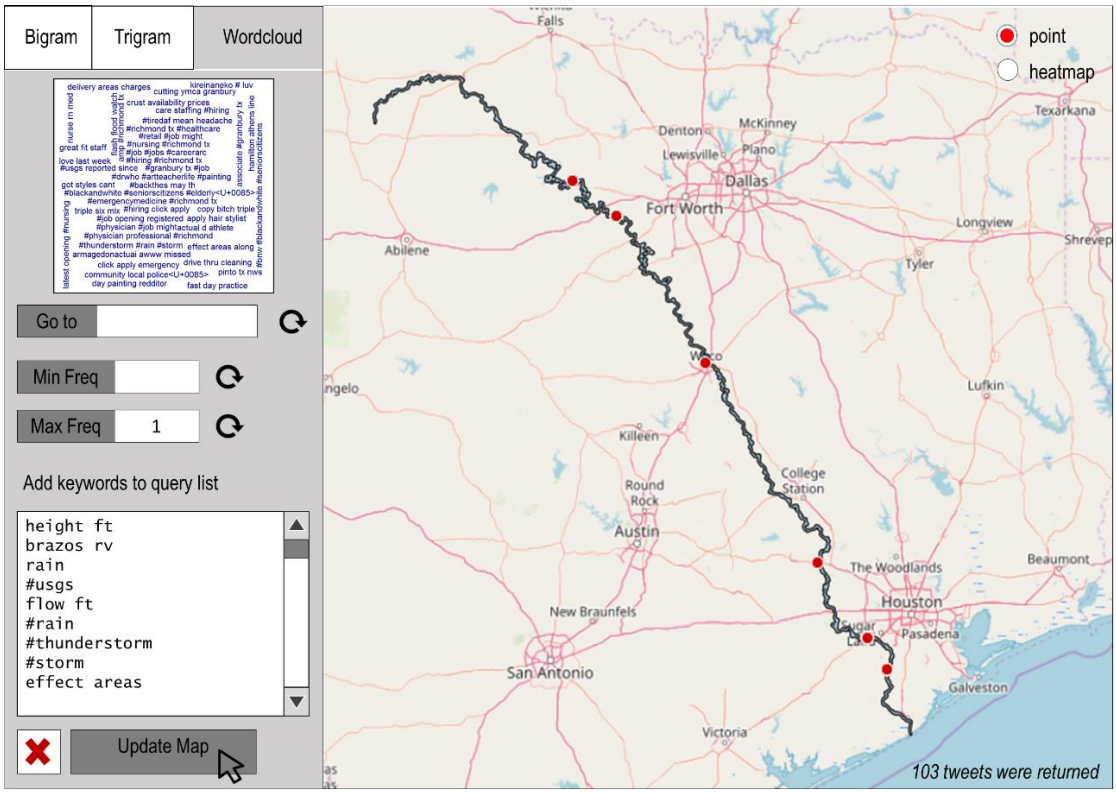
Figure 4.19 : Fenêtre cartographique de l'extraction de tweets de crise par méthode spatiale (C.Cavalière)

Ici, la requête spatiale initiale ne retourne que 164 tweets géolocalisés qui n'ont pas encore été triés lexicalement. On risque ainsi de se retrouver dans l'un des cas de figure exposés dans les mises en garde (cf. figure 4.9) : la *Brazos River* ne traversant aucun territoire urbain densément peuplé (malgré ses 2 060 km) en dehors de l'extrême périphérie occidentale de l'aire métropolitaine de Houston, on ne trouve que de faibles quantités de tweets géolocalisés. En conséquence, le territoire d'étude devra vraisemblablement être redéfini en fonction de cette contrainte.

Dans un second temps, le processus de recherche de mots-clés reste fondé sur le fonctionnement décrit dans le schéma explicatif de la figure 4.15. L'analyse lexicale est effectuée à partir des seuls tweets inclus dans la ROI. Si l'extraction spatiale suit une première extraction lexicale par mots-clés ou hashtags, on peut conserver le premier jeu de tweets de crise. Néanmoins, il est conseillé de relancer le processus de l'exploration lexicale à l'échelle de la ROI (figure 4.20), et ce afin d'éviter tout risque de biais de lexique introduit par des tweets collectés à une échelle spatiale différente de la ROI (par exemple, un phénomène local peut être enregistré dans les tweets dans la ville A et décrit par un vocabulaire particulier mais qui n'est pas perçu à l'échelle de l'Etat du Texas).



(a). Nuage de mots en tri-gramme



(b). Nuage de mots en tri-gramme et d'occurrence 1

Figure 4.20 : Volet de visualisation des résultats de l'analyse lexicale par l'approche spatiale (C.Cavalière)

Pour cette étape d'exploration lexicale, un ensemble de quatre mots-clés et de trois associations lexicales ont d'abord été sélectionnés depuis le nuage de mots en *tri-gramme* (figure 4.20.a) ; dans un second temps, on se focalise sur les associations lexicales éparses (clause *Max Freq* fixée à 1). La requête issue de cette première analyse lexicale en *tri-gramme* retourne 103 tweets de crise mais vraisemblablement émis par des automates faisant état de la hauteur de l'eau et des cumuls de précipitations (figure 4.20.b). Bien que ces automates s'activent de l'amont vers l'aval du fleuve, le problème reste que la majorité des tweets géolocalisés émis dans la ROI, qui ne sont pas associés à l'automate, correspondent à la diffusion d'offres d'emploi : dans les nuages de mots, on pouvait par exemple détecter la présence du mot *emergency* mais employé avec les termes *click*, *apply* et les hashtags *#hiring* et *#job*. Ce terme faisait donc référence à une offre d'emploi à pourvoir d'urgence et non à une situation d'urgence consécutive à un phénomène d'inondation.

A l'échelle d'un objet du territoire (le fleuve Brazos), le dernier cas d'étude se retrouve bien confronté à l'impasse soulignée dans les informations de mise en garde (cf. figure 4.9) : on peut effectivement extraire des tweets de crise mais la ROI construite ne permet pas d'identifier de tweet de crise géolocalisé relatif à l'expression des individus dans leur environnement perturbé.

Exporter le jeu de tweet final.

Pour terminer et quelle que soit la méthodologie d'extraction employée, il est nécessaire de paramétrer le format final d'export des tweets filtrés par l'ensemble des clauses appliquées (figure 4.21, en page suivante). En outre, la base de données du serveur hébergeant l'interface devra assurer l'anonymisation des données si le champ *user_id* (numéro d'identifiant attribué à tout utilisateur de Twitter) est sélectionné pour figurer dans le fichier final[29]. De la même manière, il est possible d'exporter, le cas échéant, toute région d'intérêt tracée sur l'interface.

[29] L'anonymisation des données peut être réalisée en suivant ces étapes : (1) extraire la liste de l'ensemble des valeurs présentes dans le champ *user_id* ; (2) affecter à chaque valeur un nouveau numéro d'identifiant ; (3) effectuer une jointure pour associer ce nouveau numéro d'identifiant à la table des tweets à exporter.

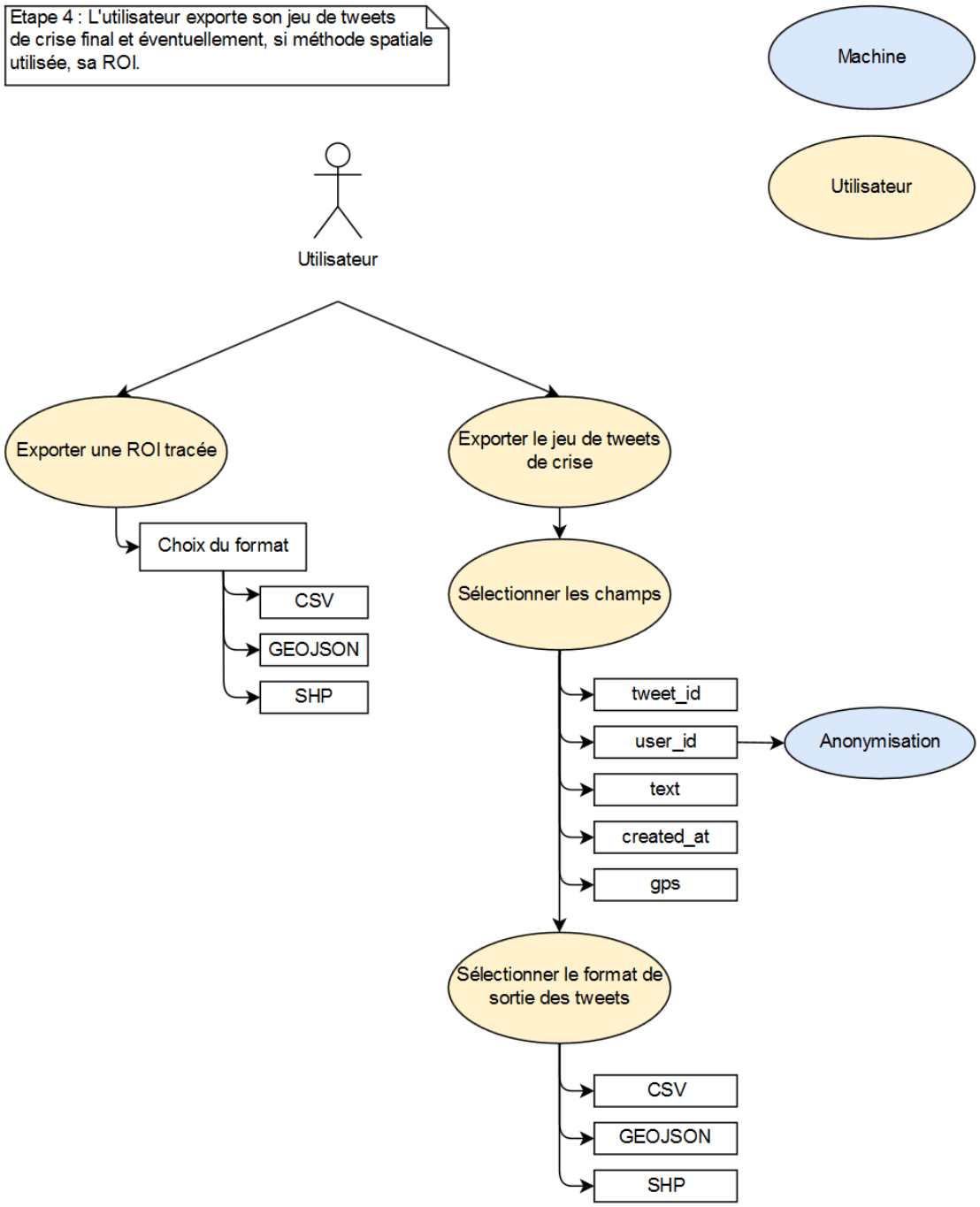


Figure 4.21 : Export des tweets filtrés et des géométries de l'utilisateur (C.Cavalière)

4.3. Données complémentaires, outils et méthodologie d'analyse

Ce dernier axe établit, dans un premier temps, l'inventaire des données officielles rassemblées afin d'apporter des éléments de recontextualisation des conditions d'émission des tweets géolocalisés[30]. Il présente ensuite les différentes étapes d'analyse à mettre en œuvre pour aborder les questions thématiques énoncées dans les tableaux 4.1 et 4.2 (cf. paragraphe 4.1.2) et enfin, notre approche des outils mobilisés.

4.3.1. Données officielles collectées en tant qu'éléments de contexte

4.3.1.1. Données phénoménologiques et événementielles

Compte-tenu de la thématique de cette recherche, nous avons tout d'abord collecté des données officielles liées aux phénomènes hydrométéorologiques physiques et à l'événementiel social qui en résulte. Les jeux de données collectées sont présentés dans le tableau 4.4 :

Tableau 4.4 : Jeux de données phénoménologiques et événementielles collectées

Type	Organisme	Description	Format	URL
Cours d'eau	Texas Waterboard Development	Simple tracé linéaire, pour la cartographie, des cours d'eau du Texas, à diverses résolutions	SHP (Lignes)	https://www.twdb.texas.gov/mapping/data-services.asp
Précipitations	National Weather Service	Grille carroyée de 4x4 km de côté	SHP (Points)	http://water.weather.gov/precip/download.php
Hauteurs d'eau	USGS National Water Information System	Mesures quoditiennes enregistrées par les stations de jaugeage déployées par l'USGS (pas de temps le plus précis = 15 min)	CSV	https://waterdata.usgs.gov/nwis
Refuges	Fulcrum community	Localisation des sites refuges ouverts et de leur profil pendant l'ouragan Harvey, données *crowdsourcées*	SHP (Points)	https://respond-harvey-geoplatform.opendata.arcgis.com/datasets/harveyrelief-crowdsourced-data-fulcrum
Dégâts aux habitations	FEMA	Ouragan Harvey : inventaire des dommages aux habitations aux dates 30-31/08/2017 et 01/09/2017	SHP (Points)	https://gis.fema.gov/DataFeeds.html
Storm Events Database	NOAA	Bulletins mensuels inventoriant les phénomènes naturels connus depuis 1950 jusqu'à mai 2019	PDF	https://www.ncdc.noaa.gov/IPS/sd/sd.html

[30] Conformément à l'hypothèse de recherche n°3 formulée dans l'introduction vis-à-vis de la non autosuffisance des traces numériques géolocalisées.

Nous donnons quelques précisions supplémentaires concernant les jeux de données suivants :

- les données de la *FEMA* faisant état des habitations ayant subi des dommages consécutifs à une catastrophe naturelle n'ont été disponibles que pour le phénomène extrême rare (ouragan Harvey). Les dommages aux habitations sont répartis en quatre catégories : *affordable* (peu de dégâts), *minor* (dégâts mineurs), *major* (dégâts majeurs), *destroyed* (complètement détruit). Nous nous servirons de ces données à titre indicatif car la *FEMA* ne fournit pas d'image permettant d'estimer le type de dégâts associés à chaque catégorie énoncée.

- les données de précipitations sont établies à l'échelle quotidienne uniquement : il s'agit d'un maillage spatial de 4km, autrement dit, on dispose d'un point de cumul pluviométrique tous les quatre kilomètres. En fait, ces points correspondent à des cumuls estimés à partir du croisement entre données issues des pluviomètres de stations météorologiques et d'images radar. Qu'en est-il de la fiabilité de ces estimations ? Le *NWS* présente ce produit comme l'un des plus fiables en termes de données de précipitations à haute résolution spatiale. Néanmoins, les territoires montagneux peuvent introduire des erreurs dans les images radar (estimées à cinq kilomètres maximum)[31]. L'organisme énonce également une restriction d'usage : le produit est mis à disposition de l'ensemble des publics mais ne bénéficie pas d'un agrément officiel et ne peut donc pas être utilisé dans des procédures légales.

- *Fulcrum Community* est une plateforme de *crowdsourcing* assurant le développement d'applications et la collecte de données pour la prévention et la gestion des catastrophes naturelles en temps réel. Le jeu de données téléchargé ici est constitué par le *public général*[32] et inventorie l'ensemble des sites refuges ouverts et ressources disponibles pendant le passage de l'ouragan Harvey (hôpitaux, services de ravitaillement, refuges, refuges accueillant et soignant les animaux, pharmacies, *etc*.).

- La *Storm Events Database* se présente sous la forme de bulletins mensuels et inventorie trois types de phénomènes[33] : les tempêtes et autres phénomènes susceptibles de causer des dégâts, victimes, blessures ou de perturber les activités humaines ; les phénomènes rares ou inhabituels (chute de neige en Floride) ou tout autre phénomène météorologique significatif (températures ou précipitations records).

D'une manière générale, il apparaît plus facile de collecter des jeux de données complémentaires relatifs aux phénomènes extrêmes (plutôt qu'aux inondations saisonnières) : pendant le passage de l'ouragan, l'entreprise américaine ESRI a directement ouvert une plateforme *opendata* de jeux de données spatiales dont la mise à jour a été

[31] Les territoires du Texas qui sont étudiés dans cette recherche échappent à cette contrainte : l'Etat est majoritairement constitué de plaines vallonnées de faible altitude, de baies et d'estuaires. C'est dans l'extrême ouest de l'Etat (Trans-Pecos), à la frontière mexicaine, qu'on trouve des régions aux reliefs plus contrastés : vallées, hauts-plateaux et monts (dont le point culminant, le Pic Guadalupe, se situe à 2 667 mètres d'altitude). L'altitude moyenne de l'Etat reste de 520 mètres. Source : https://en.wikipedia.org/wiki/Geography_of_Texas

[32] Source identique au lien intégré dans le tableau 4.4 pour la ligne *Refuges*.

[33] Source : https://www.ncdc.noaa.gov/stormevents/ (Consulté pour la dernière fois le 23/08/2019)

continue dans l'année suivante (en l'occurrence, il s'agit de la plateforme *Respond Harvey Geoplatform,* mentionnée dans la colonne URL du tableau 4.4).

4.3.1.2. Données socio-démographiques

Par le recours aux données socio-démographiques, nous chercherons à recontextualiser les profils éventuels des lieux et des individus générateurs de traces numériques géolocalisées. Les jeux de données téléchargées sont décrits dans le tableau 4.5 :

Tableau 4.5 : Jeux de données socio-démographiques collectées

Type	Organisme	Description	Format	URL
Caractéristiques de la population	ACS	Ensemble de variables quantitatives/qualitatives descriptives des profils des entités géographiques à diverses échelles : Etat, comté, *census tract*, *block group*, etc. Exemple : nombre d'habitants recensés dans une entité, nombre d'habitants dans chaque classe d'âge, nombre d'habitants par type d'habitat, etc.	CSV	https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
Caractéristiques des foyers	ACS	Ensemble de variables quantitatives/qualitatives descriptives des profils des entités géographiques à diverses échelles : Etat, comté, *census tract*, *block group*, etc. Exemple : revenu moyen/médian par foyer, nombre de foyers disposant des divers types d'accès à Internet, montant moyen annuel alloué à la couverture maladie par foyer, etc.	CSV	https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
Indice de vulnérabilité sociale (Social Vulnerability Index, SVI)	Centers for Disease Control and prevention (CDC)	Indice statistique d'estimation de la vulnérabilité des territoires et de leurs populations face aux risques naturels, technologiques et sanitaires.	CSV	https://svi.cdc.gov/factsheet.html

De la même manière, nous listons les différentes remarques afférant aux jeux de données collectés :

- Aux Etats-Unis, les recensements globaux de la population sont décennaux. Par rapport à notre recherche, le recensement le plus proche date donc de 2010, le prochain étant prévu pour 2020. Pour autant, l'*American Community Survey* publie des estimations annuelles[34] pour la plupart des variables collectées, en constituant un échantillon représentatif des foyers américains. Chaque année, environ 3,5 M d'adresses de foyers sont échantillonnées (par quota) dans le pays et interrogées. Pour les années 2016 et 2017, les pourcentages de questionnaires retournés étaient respectivement de 94,7% et de 93,7%[35]. Les données fournies incluent donc la valeur théorique calculée sur l'échantillon ainsi que l'intervalle de

[34] A l'exception de l'année 2018, pour laquelle la publication des estimations annuelles mises à jour a été retardée en raison du *Shutdown*.
[35] Source : https://www.census.gov/programs-surveys/acs/methodology.html (Consulté pour la dernière fois le 21/08/2019)

confiance. Nos analyses étant focalisées sur des phénomènes ayant eu lieu entre 2016 et 2017, nous privilégierons, dans la mesure du possible, les estimations annuelles calculées pour ces deux années. En outre, les statistiques socio-démographiques sont ethnicisées : nous pourrons donc nous poser la question de l'équité devant l'accès et l'usage des plateformes numériques génératrices de traces géolocalisées, en fonction des origines ethniques des individus. Enfin, les données sont collectées et agrégées à différents niveaux de granularité : certains niveaux respectent les découpages administratifs (Etats, comtés, communautés, *etc.*). D'autres, d'échelle plus fine, correspondent à des découpages comprenant des populations statistiquement homogènes : il s'agit principalement, de la plus petite à la plus grande échelle de recensement, des *census tracts*, *block groups* et *census blocks*.

- Bien qu'il soit développé par le *Centre de prévention et de contrôle des maladies*, l'indice de vulnérabilité sociale nous intéresse pour deux raisons principales, la première étant les variables prises en compte dans son calcul. Le *CDC* présente le *SVI* comme une aide indispensable à la localisation des communautés vulnérables en cas de catastrophe naturelle, technologique ou d'épidémie[36]. Il est déterminé à partir d'une quinzaine de variables réparties selon quatre thèmes (figure 4.22) : le statut socio-économique, la composition des foyers ainsi que le handicap, l'appartenance à une minorité ethnique et le niveau de langue, le type de logement et l'accès aux transports.

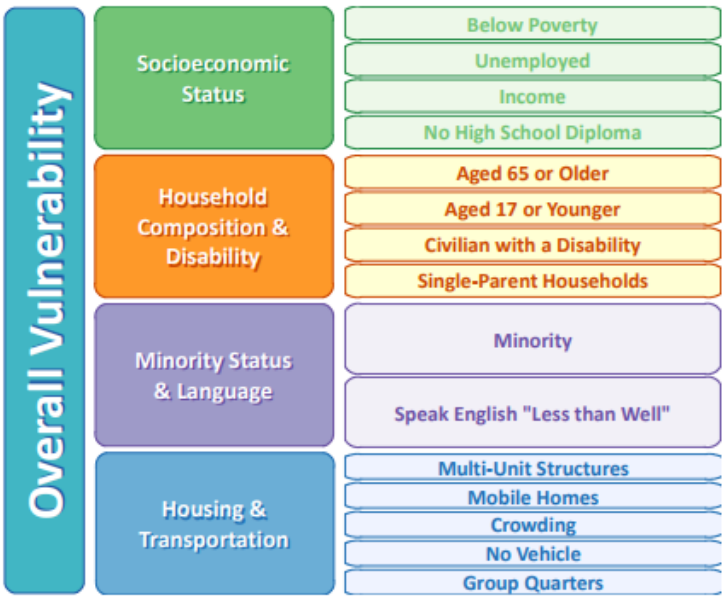


Figure 4.22 : Variables prises en compte dans le calcul de l'indice de vulnérabilité sociale (Source : CDC)

[36] Source : https://svi.cdc.gov/publications.html (Consulté pour la dernière fois le 21/08/2019)

L'indice de vulnérabilité sociale correspond en fait à l'attribution d'un rang percentile à chaque entité géographique considérée, selon la méthode suivante[37] : dans un premier temps, un rang percentile est calculé pour chaque entité et pour chaque variable. Ces rangs sont ensuite agrégés en fonction des thèmes : par exemple, pour chaque entité, le rang percentile du thème *Socioeconomic Status* correspond à la somme des rangs de l'entité pour les variables *Below Poverty*, *Unemployed*, *Income* et *No High School Diploma*. Le score final, qui constitue l'indice de vulnérabilité sociale, est obtenu à partir de la somme des rangs obtenus dans chaque thème. Il varie ainsi entre 0 (indice de vulnérabilité faible) et 1 (indice de vulnérabilité forte) : une entité géographique présentant alors un indice de 0,8 signifie que 80% des autres entités de la population affichent un meilleur score (inférieur à 0,8 et donc moins vulnérables) et que les 20% d'entités autres présentent une vulnérabilité encore plus forte. Afin d'éviter les biais introduits par les effets d'échelle et la taille de la population, les rangs percentiles sont calculés à diverses résolutions (*comté*, *census tract*) et à diverses échelles géographiques : Etat, comté. En revanche, l'indice n'est pas mis à jour chaque année : les données utilisées ici datent en conséquence de 2014 (pour des phénomènes naturels survenus entre 2016 et 2017). La seconde raison pour laquelle nous avons recours à ces données est directement liée à la thématique de la recherche : dans la problématique des risques naturels, il est logique d'avoir recours à la vulnérabilité des populations comme indicateur socio-spatial et il sera intéressant de comparer la localisation des tweets de crise géolocalisés avec les rangs de cet indice.

4.3.1.3. Données d'habillage

Pour les besoins cartographiques et pour assurer la préparation des données, nous avons également collecté des données géographiques d'habillage et de délimitation du maillage administratif américain (tableau 4.6) :

[37] Source : https://www.youtube.com/watch?v=REKFHOryflA (Consulté pour la dernière fois le 23/08/2019)

Tableau 4.6 : Jeux de données d'habillage et de préparation aux traitements

Type	Organisme	Description	Format	URL
Maillage administratif	GADM	Différents niveaux du maillage administratif américain : ici, nous utilisons les couches des Etats et des comtés	SHP (Polygones)	https://gadm.org/
Aires urbaines/ métropolitaines	Census Bureau	Fichiers SIG indiquant la localisation des aires urbaines et métropolitaines (points) ou délimitant leur emprise (polygones)	SHP (Points, Polygones)	https://www.census.gov/programs-surveys/geography.html
Unités de recensement (Census Tracts)	Census Bureau	Fichier SIG délimitant l'emprise spatiale des unités fines de recensement ; elles peuvent être associées, par jointure attributaire, aux données statistiques de l'*American Community Survey*	SHP (Polygones)	https://www.census.gov/programs-surveys/geography.html
Houston - Quartiers	Zillow Neighborhoods	Fichier SIG délimitant l'emprise spatiale des différents quartiers du centre de la métropole de Houston	SHP (Polygones)	https://data.opendatasoft.com/explore/dataset/zillow-neighborhoods@public/information
Houston - Voirie	Census Bureau	Fichier SIG de lignes stockant l'ensemble du réseau de voirie du comté de Harris (sur lequel s'étend en majeure partie l'aire métropolitaine de Houston)	SHP (Lignes)	https://catalog.data.gov/dataset/tiger-line-shapefile-2013-county-harris-county-tx-all-roads-county-based-shapefile

Le recours à ces données est indispensable afin d'effectuer les requêtes qui nous permettront de sélectionner les tweets inclus dans diverses entités spatiales ou de les associer à ces mêmes entités ; mais encore afin de pouvoir leur joindre les données statistiques de l'*American Community Survey*. En outre, elles constituent les fonds de carte permettant de spatialiser discours et analyses.

4.3.2. Méthodologie d'analyse des tweets géolocalisés et des données complémentaires

4.3.2.1. Traitements préparatoires appliqués aux tweets géolocalisés

Depuis l'infrastructure de collecte et de stockage du LIG, présentée dans le paragraphe 4.2.2.1, nous avons extrait deux jeux de tweets bruts (filtrés uniquement en fonction du critère spatial et du critère temporel et non en fonction de la sémantique) pour les deux périodes suivantes :

- du 1er mars au 30 juin 2016 sur l'Etat du Texas, soit sur l'ensemble de la période des épisodes de pluies et inondations printanières ;
- du 23 août au 1er septembre 2017 sur les Etats du Texas et de la Louisiane, soit deux jours avant que l'ouragan ne touche terre une première fois au Texas et deux jours après qu'il touche terre (alors rétrogradé au stade de tempête tropicale) une seconde fois à la frontière entre les deux Etats.

Pour ces deux jeux de données, nous sélectionnons les champs suivants : *tweet_id* (numéro unique d'identifiant de tweet), *user_id* (numéro d'identifiant de chaque utilisateur), *created_at* (horodatage du tweet au fuseau horaire GMT+01), *text* (texte du tweet) et *gps* (coordonnées GPS du tweet au format *Well-Known Text*). Pour rappel, ces deux jeux de données sont ensuite importés dans deux bases de données locales. Néanmoins, avant toute procédure d'analyse, nous devons appliquer à ces jeux de tweets géolocalisés bruts deux étapes de transformation :

- la conversion du champ *text* (contenu sémantique du tweet) en minuscules afin de faciliter les futurs traitements sémantiques (suppression du problème de la sensibilité à la casse sur R) ;
- la conversion du champ *created_at* dans un nouveau champ, *local_timestamp* : comme indiqué précédemment, les tweets sont capturés au fuseau horaire de l'Europe Centrale. Sur notre terrain d'étude, nous sommes en présence de deux fuseaux : Texas et Louisiane se trouvent dans le fuseau *central* des Etats-Unis (soit GMT-06), à l'exception de la pointe extrême ouest du Texas (El Paso) qui se trouve dans le fuseau *mountain* (GMT-07). Nous avons ainsi téléchargé un *shapefile* de polygones délimitant l'emprise spatiale des quatre fuseaux traversant les *CONUS*, c'est-à-dire les quarante-huit Etats contigus des Etats-Unis (hors Alaska, Hawaï et territoires ultramarins comme Porto Rico) et associé, par jointure spatiale, chaque tweet au fuseau approprié. On peut alors, par requête SQL, soustraire sept heures aux tweets localisés dans le fuseau *central* et huit heures aux tweets localisés dans la partie *mountains*[38]. Enfin, nous ajoutons deux dernières colonnes : *date* et *time* qui correspondent au nouveau champ *local_timestamp* respectivement tronqué à la date et à l'heure[39].

4.3.2.2. Démarches de l'analyse exploratoire des tweets de crise et données complémentaires

Nous détaillons ici les méthodologies d'analyse envisagées et les outils mobilisés dans le cadre des premières questions énoncées dans les tableaux 4.1 et 4.2. Les démarches exposées dans les paragraphes qui suivent constituent une base modulable de questionnements et d'expériences qui, conformément au positionnement épistémologique exprimé dans la section 4.1.3.1, pourront être réorientés en fonction des résultats que nous observerons (auquel cas, les démarches appliquées pour ces réorientations seront décrites dans les sections concernées des deux chapitres suivants de résultats). Dans cette section, chaque démarche est représentée par un schéma illustrant les étapes de réflexion, de préparation des données ainsi que les phases d'analyse et de test, dans le but de rendre reproductible tout traitement effectué en réponse à une question particulière.

[38] La requête utilisée est alors de ce type : *UPDATE tweets_bruts_2016 SET local_timestamp = created_at – INTERVAL '7 hours' WHERE time_zone = 'central'* ;
[39] La requête correspondante est alors : *UPDATE tweets_bruts_2016 SET date = local_timestamp::DATE ;*

Question des biais socio-spatiaux liés à l'intégration et à l'utilisation du numérique.

Dans le tableau 4.1, nous nous interrogions sur l'éventuelle existence d'une fracture socio-spatiale en termes d'usage des plateformes numériques et de production de traces géolocalisées. Le schéma proposé dans la figure 4.23 soumet alors la démarche adoptée au regard de cette question de représentativité des populations dans l'activité virtuelle.

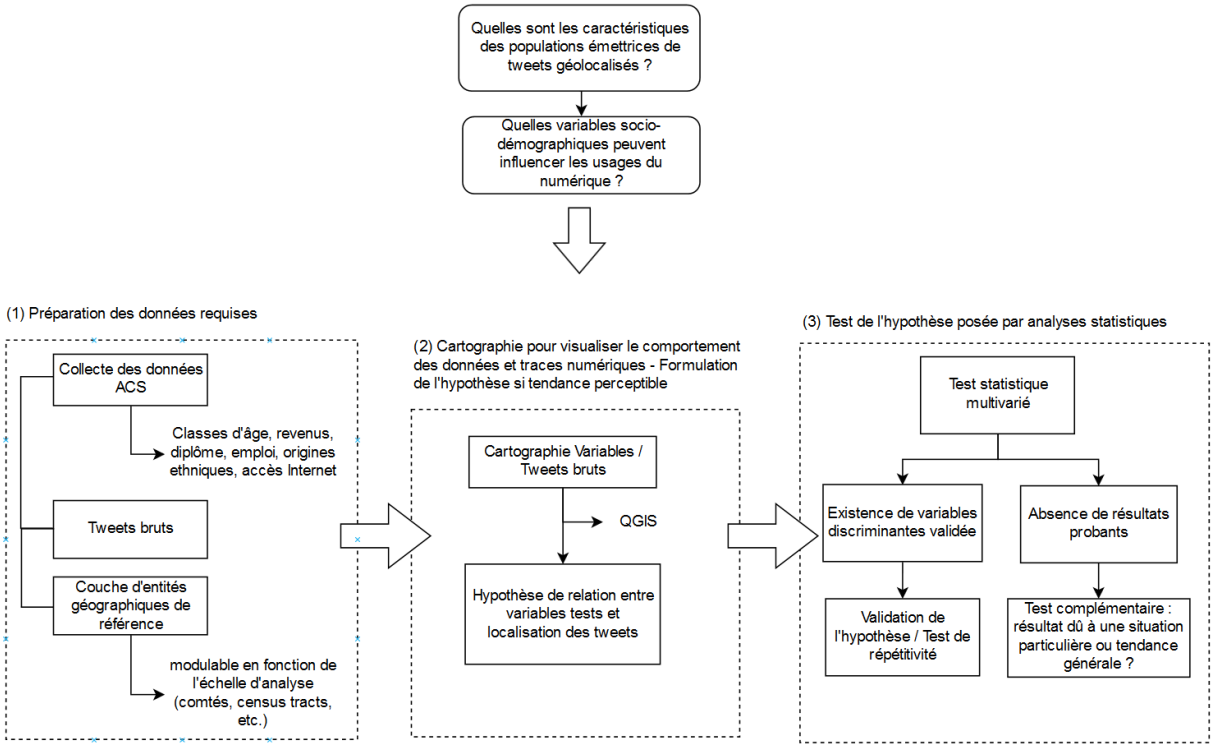


Figure 4.23 : Démarche théorique appliquée à la question des populations émettrices de traces géolocalisées (C.Cavalière)

Cette question revient tout d'abord à considérer les variables socio-démographiques susceptibles d'introduire une variabilité dans l'accessibilité aux technologies connectées et à leur usage. Comme évoqué dans l'introduction du manuscrit, (Eskenazi *et al*., 2017) avaient mis en évidence que les individus diplômés étaient plus aptes à utiliser les applications smartphone génératrices de traces. Observera-t-on des tendances analogues dans le cas des tweets géolocalisés au Texas ? Nous envisageons donc une première étape de collecte et de préparation des données relatives aux variables disponibles sur le site de l'*ACS* (niveau d'éducation, emploi, revenus, origines ethniques, équipement numérique du foyer, classes d'âges), jointes aux entités géographiques d'étude et cartographiées en parallèle des tweets géolocalisés. La cartographie n'est alors pas le résultat démontrant l'existence d'une hypothèse mais le support de formulation de l'hypothèse en comparant distribution spatiale des variables considérées et distribution spatiale des tweets et des utilisateurs. L'hypothèse formulée doit enfin être statistiquement testée pour validation ou réfutation de l'influence éventuelle de certaines variables sur la contribution des individus à l'activité tweeting

géolocalisée ; en outre, on préconise de renouveler le test de l'hypothèse afin de dissiper l'existence possible de résultats observés liés au hasard du premier terrain d'étude.

Dans un second temps, nous avons posé la question des lieux de l'activité tweeting géolocalisée : peut-on identifier des facteurs explicatifs de la localisation des tweets ? De la même manière que précédemment, la figure 4.24 illustre la démarche appliquée.

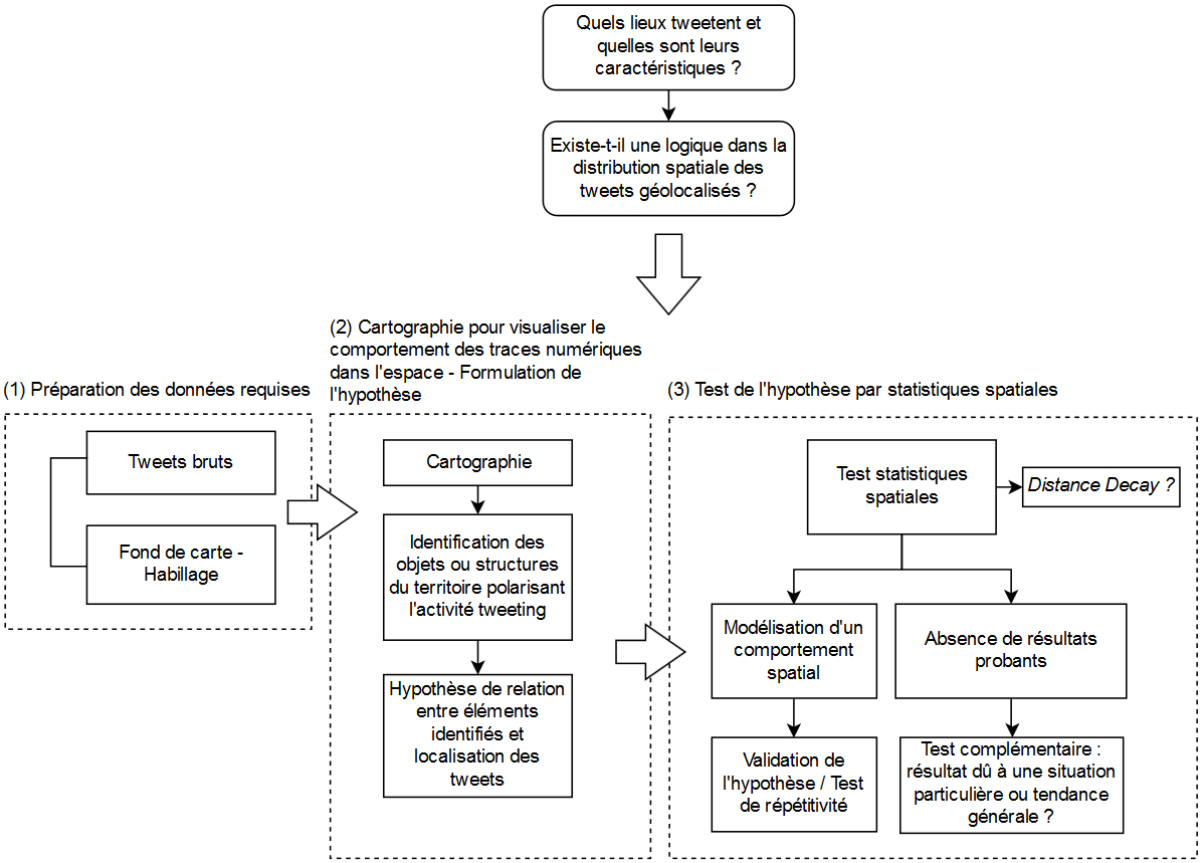


Figure 4.24 : Démarche théorique appliquée à la question des lieux de l'activité virtuelle (C.Cavalière)

Les tweets peuvent être affichés sous forme de carte de chaleur et superposés à des données d'habillage (par exemple, un fond de carte *OpenStreetMap*) afin d'identifier la localisation des foyers de l'activité tweeting géolocalisée et les éventuels objets du territoire qui les polarisent. Si de telles tendances spatiales émergent depuis la carte, on pourra alors tenter de mesurer l'évolution de l'activité tweeting en fonction de la distance à l'objet/au foyer et de la modéliser (*distance decay*)[40]. Si la carte n'affiche aucune tendance spatiale tangible ou si la modélisation échoue, on pourra alors s'interroger sur la probabilité d'être confronté à une distribution spatiale aléatoire qu'on ne peut pas prédire (par un test d'autocorrélation spatiale). Comme indiqué précédemment, il conviendra également de

[40] La méthode reste identique pour la question 2 posée dans le cadre de l'activité virtuelle de crise dans le tableau 4.2 : *Observe-t-on une décroissance de l'activité tweeting de crise en fonction de la distance aux lieux de survenue des phénomènes ?*

vérifier que les résultats constatés dans une étude de cas ne sont pas liés au hasard de la situation et peuvent se retrouver dans d'autres lieux.

Questions relatives aux événements virtuels en réponse à la survenue et aux manifestations d'un phénomène physique dommageable.

En ce qui concerne le questionnement spécifique aux risques naturels, nous avons posé la première piste de recherche générale dans le tableau 4.2 (*Quels sont les lieux de l'activité tweeting de crise ? Ces lieux s'agitent-ils en fonction de la dynamique spatio-temporelle et de l'intensité des phénomènes et événements du réel ?*). Le schéma illustrant la démarche d'exploration de cette première question est présenté en figure 4.25.

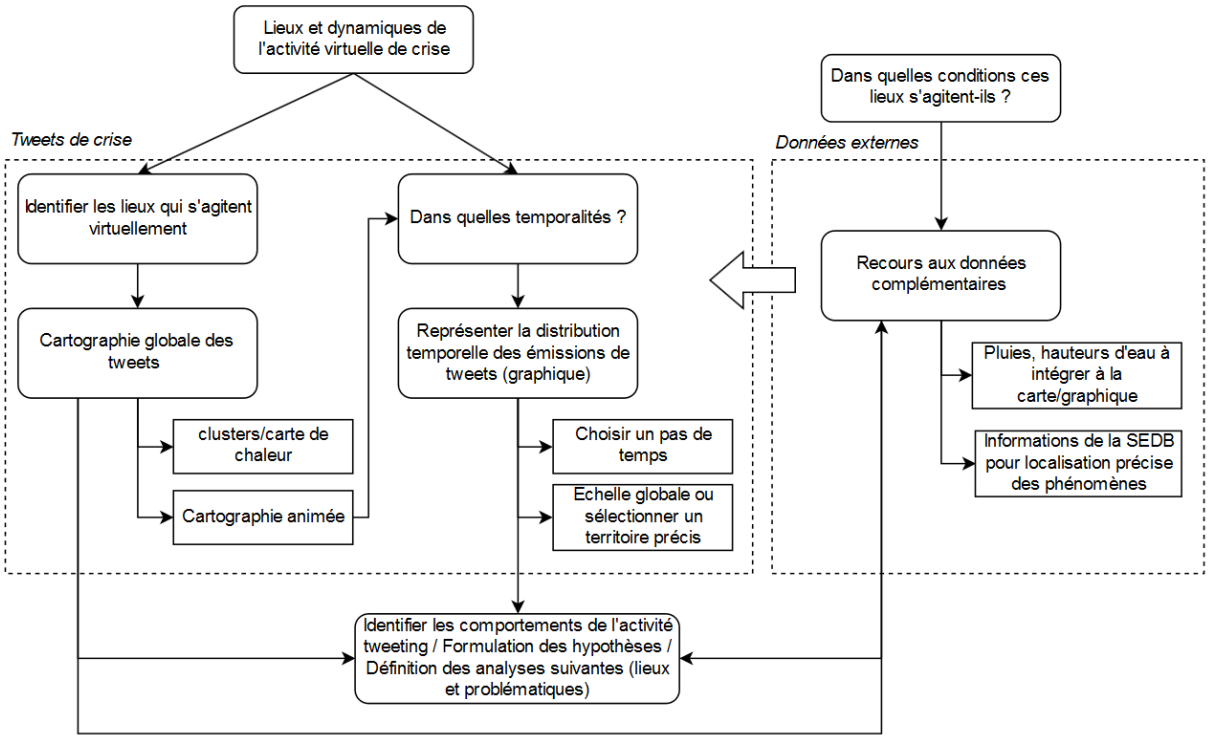


Figure 4.25 : Démarche théorique appliquée à la question de la correspondance des dynamiques réelle et virtuelle (C.Cavalière)

En fait, cette première piste de recherche s'avère générale : elle consiste à valider, réfuter ou nuancer l'existence d'une correspondance entre réel et virtuel, en d'autres termes entre lieux affectés et lieux inscrits dans l'événement virtuel, mais également entre conditions environnementales du phénomène et propriétés de la réponse virtuelle. Les lieux d'émission de tweets de crise géolocalisés peuvent ainsi être cartographiés en parallèle des données de précipitations ; de la même manière, on peut représenter les flux temporels d'émissions de tweets aux côtés de l'évolution des hauteurs d'eau mesurées sur les rivières. L'état de l'art aurait tendance à valider l'existence de correspondances dans les dynamiques réelles et virtuelles mais au regard de la complexité des usages socio-spatiaux actuels du numérique, nous faisons le pari que les résultats de ce premier lot d'explorations mettront en évidence des structures non attendues (qui permettront alors de nuancer les constats établis dans l'état

de l'art donné dans le chapitre précédent). En conséquence, cette première piste exploratoire sert de prétexte pour la définition d'un deuxième lot d'analyses qui seront construites en fonction des résultats observés (identification de lieux à explorer et formulation d'un nouveau questionnement).

La deuxième piste exploratoire formulée dans le tableau 4.2 (Q2 : *Observe-t-on une décroissance de l'activité tweeting de crise en fonction de la distance aux lieux de survenue des phénomènes ?* Q3 : *Des tweets géolocalisés émis dans une certaine proximité spatio-temporelle sont-ils lexicalement cohérents ?*) se focalise sur la vérification de la première loi de Tobler avec les tweets de crise géolocalisés. La première partie du questionnement peut être abordée en ayant recours à la démarche évoquée par la figure 4.24 (il s'agira alors d'identifier des objets ou lieux du territoire susceptibles de polariser l'activité virtuelle de crise). Dans un second temps, la démarche spécifique à la question de la cohérence des tweets entre eux est résumée par la figure 4.26.

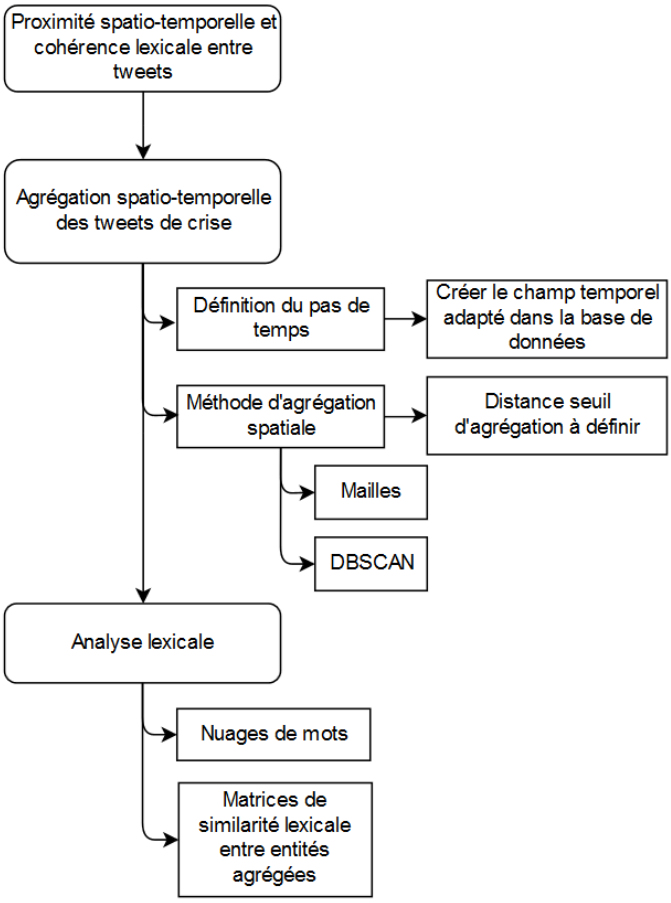


Figure 4.26 : Démarche théorique appliquée à la question de la cohérence entre proximité spatio-temporelle et discours des tweets de crise (C.Cavalière)

Comme précisé ci-dessus, nous cherchons à savoir si la première loi de Tobler est applicable aux tweets de l'événement virtuel ; ce type d'analyse est ainsi envisagé à l'échelle locale. Il s'agit en premier lieu d'agréger les tweets spatialement proches dans une temporalité donnée, en ayant recours à un carroyage régulier ou à l'algorithme DBSCAN. La seconde étape de cette question fait intervenir l'analyse sémantique pour l'exploration lexicale de chaque entité formée ; nous pourrons alors rendre compte de la variabilité ou de la similarité sémantique entre les agrégats ou dans un même agrégat, par nuages de mots et par construction de matrices de similarité (indiquant le pourcentage de mots communs entre deux agrégats).

La troisième piste évoquée explore les paramètres de la dynamique de l'événement virtuel seul et s'attache également à la problématique de la résolution spatio-temporelle des tweets de crise géolocalisés (cf. Q4 : *Quelles dynamiques spatio-temporelles peut-on mettre en évidence dans les tweets seuls ? Peut-on observer la diffusion d'un message d'alerte ? Les utilisateurs sont-ils mobiles dans l'espace en crise et si oui, que peut-on apprendre de leur parcours dans l'espace en crise ?* Q5 : *Quel est le degré d'implication individuelle dans l'activité tweeting de crise ?* Q6 : *La quantité de traces disponibles est-elle compatible avec les changements d'échelle et le passage à une résolution fine ?* du tableau 4.2). Le schéma de la figure 4.27 illustre les premiers traitements envisagés :

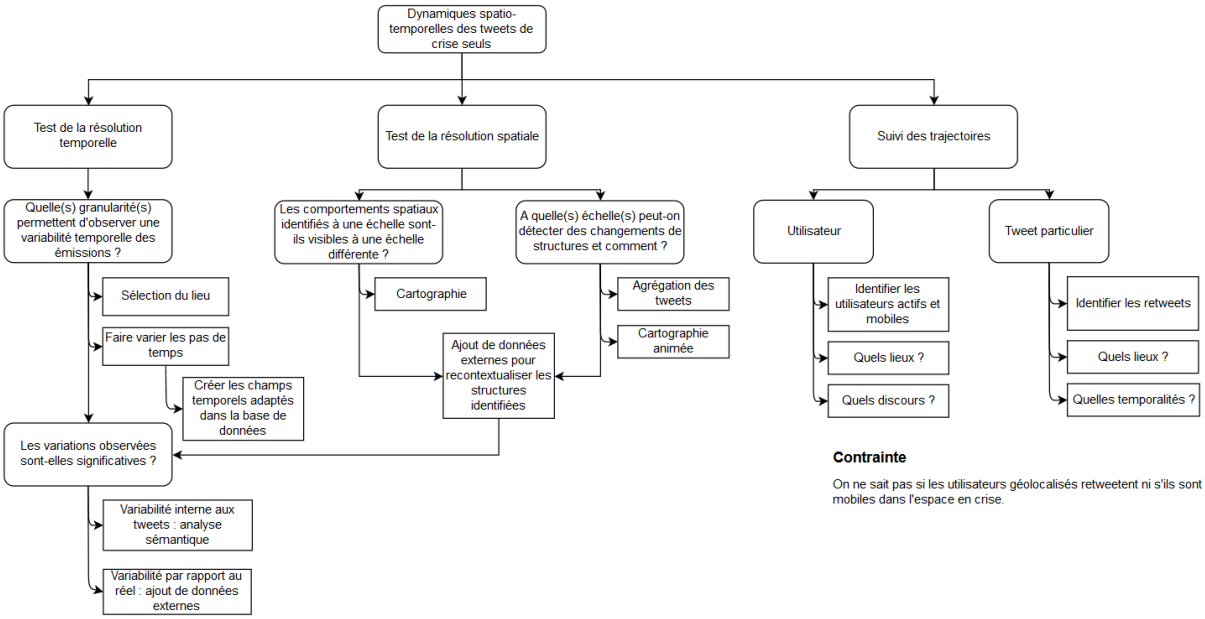


Figure 4.27 : Démarche théorique appliquée à la question des dynamiques spatio-temporelles des tweets seuls (C.Cavalière)

Etudier la résolution temporelle des tweets de crise géolocalisés revient à résoudre la question des pas de temps auxquels on peut détecter une variabilité des flux (par exemple, un pic soudain suivi par une décroissance rapide ou plus ou moins lente). Dans l'introduction du manuscrit, nous avons pris pour exemple l'effet de réactivité massive et immédiate en

réponse à l'annonce des résultats des élections présidentielles : on pouvait alors détecter des variations d'émissions à l'échelle de la minute. La question qui se pose ici est la suivante : à quelle résolution temporelle peut-on détecter des variations de flux pertinentes[41] ? Ce critère de pertinence est vérifié par le recours aux données externes ou à l'analyse sémantique : par exemple, on étudie la distribution des tweets de crise à l'échelle de la demi-heure. Le flux reste quasi stable pendant quelques heures puis on observe une période de rupture marquée par un accroissement subit du volume de tweets émis, qui se stabilise tout en restant anormalement élevé dans les heures suivantes. On sélectionne alors les tweets séparément en fonction des deux périodes : la période des flux bas et la période des flux élevés. Après vérification, l'analyse sémantique témoigne en effet de la survenue soudaine d'une crue éclair entraînant une perturbation du réseau pendant quelques heures (présence d'un vocabulaire mentionnant des dégâts, des observations physiques, des demandes d'aide, *etc*. qui n'était pas visible dans la période précédente de faibles flux). Bien évidemment, les tweets étant géolocalisés, il faut croiser la possibilité d'associer une éventuelle résolution temporelle fine à une résolution spatiale fine : par exemple, si l'on parvient à identifier un pic de tweets à l'échelle du quart d'heure, à quoi ressemble leur cartographie ? Si l'on observe des îlots de tweets, on pourra alors approfondir l'analyse sémantique mais si l'on se retrouve face à des tweets isolés et dispersés dans l'ensemble du territoire étudié, il sera plus difficile de pousser les analyses.

Par ailleurs, la dimension spatiale peut encore être envisagée sous deux autres angles : il s'agit de voir si des comportements spatiaux détectés à une certaine échelle sont identiques à une autre échelle. Par exemple, considérons qu'on détecte, à l'échelle de l'Etat du Texas, une correspondance entre la dynamique spatio-temporelle des précipitations et celle de la réponse virtuelle (cf. piste de recherche évoquée dans la figure 4.25). Si tel est le cas, alors on pourra introduire l'hypothèse logique selon laquelle les utilisateurs s'agitent par temps perturbé : on observe donc des tweets de crise dans les territoires affectés par les épisodes pluvieux printaniers. Cette hypothèse sera-t-elle toujours valable si on la considère à l'échelle locale ? Autrement dit, à l'échelle d'une métropole, les lieux de l'activité tweeting de crise correspondent-ils aux lieux affectés par les pluies ?

Le dernier angle envisagé de l'étude des dynamiques spatio-temporelles du réseau seul s'intéresse au suivi des trajectoires sous deux angles :

- l'utilisateur[42] : peut-on suivre des utilisateurs mobiles dans l'espace en crise et l'évolution de leur discours en fonction du temps ou des lieux fréquentés ?

[41] En prenant en compte le fait que dans le domaine des réponses virtuelles face à un phénomène naturel, la dynamique virtuelle n'est pas nécessairement inscrite dans cette logique de masse.

[42] Sous réserve évidente de disposer d'une quantité suffisante de tweets par utilisateur et que ceux-ci soient mobiles (*a priori*, on ne devrait pas identifier beaucoup d'utilisateurs adoptant un tel comportement fortement déconseillé)

- la diffusion par retweet d'un même message[43] : peut-on tracer les territoires atteints (et dans quelles temporalités) par un même message (d'alerte émanant d'un compte officiel par exemple) ?

La quatrième piste de recherche évoquée par le tableau 4.2 s'intéresse à la question de l'identification de différentes phases de l'événement virtuel, faisant écho au réel, du passage des territoires dans ces différentes phases et d'une manière plus générale, à la variabilité spatiale du lexique employé par les tweets de crise (Q7 : *Dans l'événement virtuel, peut-on distinguer les différentes phases de la crise ? Existe-il une variabilité spatio-temporelle des territoires vis-à-vis du passage dans les différentes étapes de la crise ?* Q8 : *Observe-t-on des variations spatiales du lexique dans une même crise ? Si oui, quels facteurs explicatifs peuvent être introduits ?*). La figure 4.28 explicite la démarche envisagée :

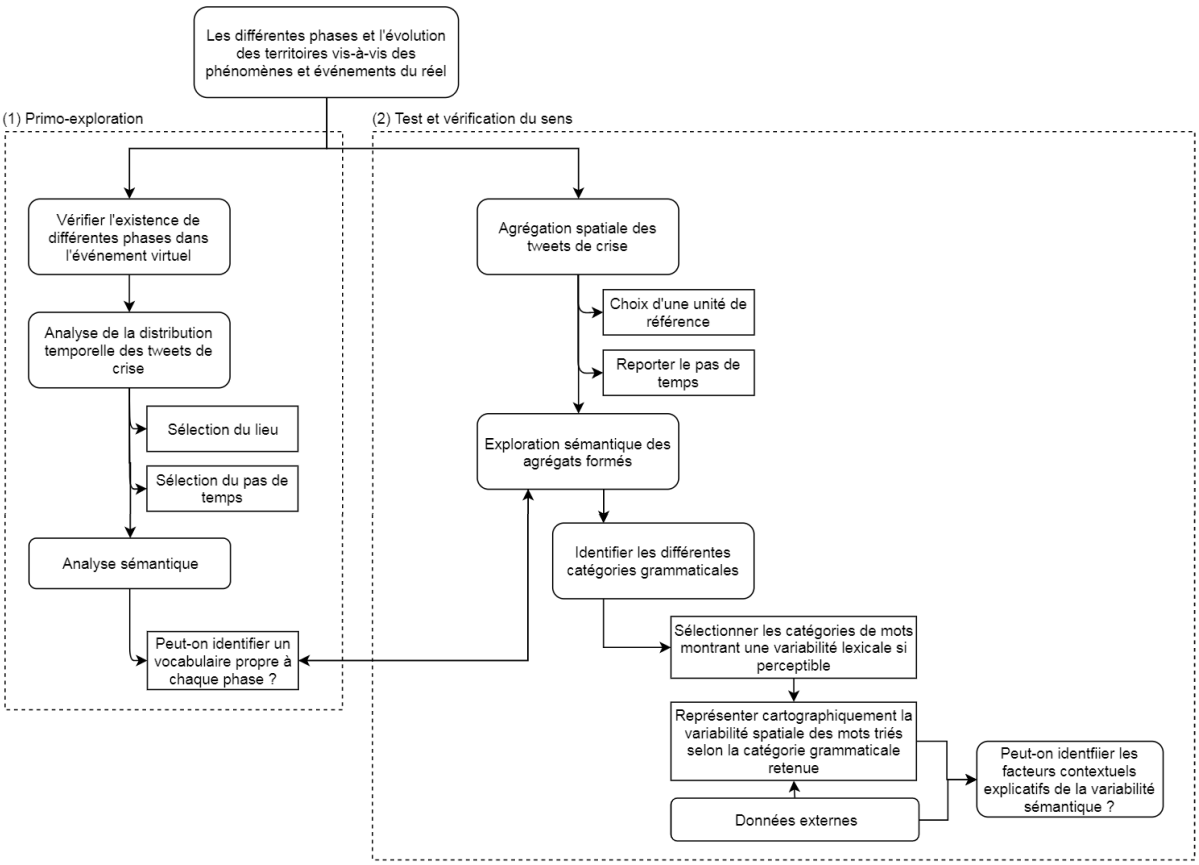


Figure 4.28 : Démarche théorique appliquée à la question des variations lexicales en fonction des territoires et du temps (C.Cavalière)

Pour vérifier l'existence de ces phases, on se propose d'abord d'avoir une nouvelle fois recours à la représentation des variations temporelles des émissions des tweets géolocalisés dont les pics, creux, stagnation, *etc*. peuvent révéler les impacts du phénomène réel sur les

[43] Là aussi, sous réserve que les utilisateurs géolocalisés utilisent les retweets.

individus et territoires, qui sont alors retranscrits dans l'événement virtuel. Une nouvelle fois, les tweets émis pendant ces différentes périodes doivent être analysés lexicalement afin de vérifier le sens des variations temporelles détectées et de mettre en évidence les mots de vocabulaire propre à chaque période. Pour la question des territoires, il conviendra en premier lieu d'agréger les tweets en fonction d'unités spatiales définies et de temporalités correspondant aux périodes mises en évidence. Dans un second temps, on pourra explorer la sémantique des tweets émis dans ces agrégats et caractériser sa variabilité spatiale selon deux critères :

- la présence des mots de vocabulaire marqueurs mis en évidence à l'étape précédente ;
- à partir de la catégorisation grammaticale du lexique (en fonction des noms, adjectifs, verbes ou adverbes), la présence d'un vocabulaire témoignant de la variabilité, d'un agrégat à l'autre, de l'intensité (*light*, *huge*, *very*, *low*, *high*, *etc*.), des actions (*help*, *go*, *evacuate*, *etc*.) ou des événements (*flooded*, *closed*, *postpone*, *etc*.).

Les résultats observés pourront être comparés aux données externes de recontextualisation de l'activité tweeting, et notamment à la *SEDB* qui a l'avantage de fournir des indications précises à l'échelle locale (jusqu'à l'échelle des quartiers des aires métropolitaines). L'enjeu consiste en effet à savoir, si l'on parvient à identifier une variabilité spatiale de la sémantique, si celle-ci peut être expliquée par des facteurs du réel.

Enfin, la dernière piste explorée concernait la possibilité d'identifier et de recontextualiser des lieux dans lesquels les utilisateurs se montrent plus actifs (cf. tableau 4.2 : Q9 : *Peut-on distinguer des lieux dans lesquels les utilisateurs sont plus actifs ? Si oui, peut-on en introduire les facteurs explicatifs ? Quels facteurs environnementaux ou sociaux font alors varier la mobilisation virtuelle des utilisateurs ?*). La démarche est explicitée dans la figure 4.29 :

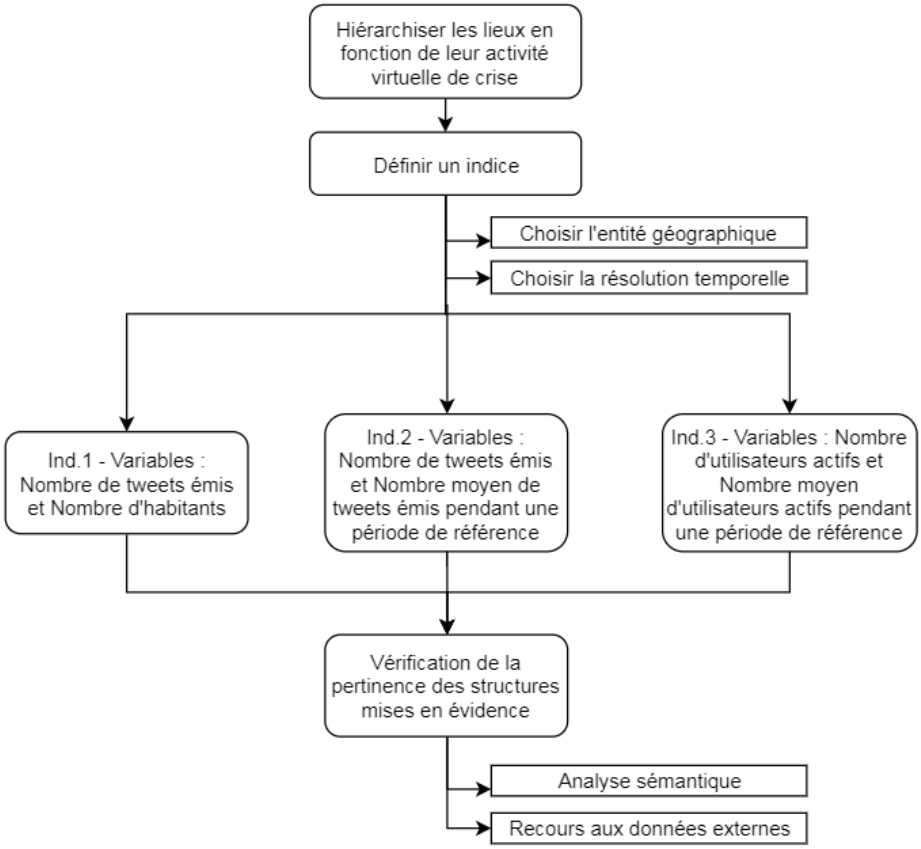


Figure 4.29 : Démarche théorique appliquée à la question de l'identification de lieux en fonction de l'intensité de l'activité virtuelle en situation de crise (C.Cavalière)

Même si l'on parvient à caractériser, en appliquant l'ensemble des démarches décrites ci-avant, des lieux et des périodes d'activités variables et à en identifier les causes plausibles, on peut néanmoins envisager un biais qui se surajoute aux conditions d'accès et d'usage du numérique. En temps normal, on observe déjà des disparités spatiales de l'activité géolocalisée de la plateforme en fonction de l'adhésion des individus, de la densité du réseau d'utilisateurs et des facteurs environnementaux qui incitent un utilisateur à tweeter. Cette variabilité spatiale des quantités de tweets émis en temps normal peut ainsi se reporter dans une période de crise. La question qu'il faut alors considérer est la suivante : quel comportement est significatif ? Le lieu dans lequel on trouve déjà, en temps normal, beaucoup d'utilisateurs actifs et de tweets, et qui reste inscrit dans cette dynamique en temps de crise ou au contraire, le lieu dans lequel on observe un signal d'agitation plus conséquent (ou qui s'active uniquement) en période perturbée ? Pour répondre à cette question, nous proposons le recours à la création d'indices permettant de comparer les lieux en gommant ce biais de l'activité virtuelle normale. Nous envisageons trois possibilités de calcul :

- un indice général calculant le ratio entre le nombre de tweets de crise émis et le nombre d'habitants recensés dans l'entité géographique de référence ;

- un indice calculant le ratio entre le nombre de tweets émis dans le pas de temps choisi et une valeur de référence normée : ici, il s'agit du nombre moyen de tweets émis sur le même pas de temps mais dans une période de référence plus longue (par exemple, en considérant que le pas de temps est l'heure, on peut calculer le ratio entre le nombre de tweets émis dans le créneau horaire sélectionné en situation de crise et le nombre moyen de tweets émis dans ce même créneau, sur une période de référence d'un mois).

- un indice calculant le ratio entre le nombre d'utilisateurs actifs enregistrés dans chaque entité géographique de référence dans l'intervalle de temps défini et une autre valeur de référence normée : dans ce cas, il s'agira du calcul du nombre moyen d'utilisateurs actifs dans l'entité géographique et à l'intérieur du pas de temps choisi, mais de nouveau dans une période de référence plus longue. L'avantage envisageable de cet indice consiste en ce qu'il efface les effets de la présence éventuelle d'un utilisateur hyperactif dans un lieu donné.

Comme indiqué dans les démarches précédentes et quel que soit l'indice utilisé, il conviendra de recontextualiser les valeurs observées, autrement dit de vérifier qu'elles soient significatives d'un phénomène ou d'un événement social consécutif local (par analyse lexicale ou en ayant recours aux jeux de données officielles complémentaires).

4.3.3. Opérationnalisation et paramétrage de la démarche d'analyse

4.3.3.1. Logiciels utilisés

Comme indiqué précédemment, l'ensemble des tweets collectés et triés pour les besoins des analyses (et selon les méthodes décrites dans la section 4.2.2.2) sont stockés sous une base de données locale *PostgreSQL*. Pour les besoins cartographiques, la base de données est connectée au logiciel SIG libre *QGIS* et en ce qui concerne le volet statistique et sémantique, elle est associée à l'interface *Rstudio*. L'annexe 1 récapitule l'ensemble des fonctions, extensions ou *package*s utilisés avec ces trois logiciels, pour l'ensemble des analyses.

Nous donnons néanmoins les précisions supplémentaires suivantes concernant l'utilisation de deux outils :

- QGIS reste principalement utilisé pour la visualisation et la navigation dans les divers jeux de données, et pour la mise en page des cartes qui sont présentées dans les deux prochains chapitres de résultats (et qui servent d'appui à la réflexion et à la prise de décisions quant aux pistes d'analyse à poursuivre ou à ouvrir). L'extension principalement utilisée ici est *TimeManager* : celle-ci permet d'animer, selon un pas de temps choisi, différentes couches de données SIG disposant d'un champ temporel de format *aaaa/mm/jj hh:mm:ss*. Certaines cartographies ont cependant été réalisées sous R, à l'aide du *package leaflet*, qui permet de

générer des cartographies interactives facilitant la rapidité de l'affichage et de l'exploration des traces et données, par rapport au SIG conventionnel.

- Les étapes d'analyse lexicale des tweets sont effectuées en ayant recours aux *packages* de R présentés précédemment dans le cadre de la recherche lexicale en vue de l'extraction de tweets de crise. Le *package tm* est utilisé dans le cadre du nettoyage textuel des tweets (suppression des émojis, des URL, des *stopwords*, réutilisation de la fonction supprimant la ponctuation en dehors des signes # permettant d'identifier les hashtags). Notons que nous n'avons pas pratiqué la *racinisation*, étape qui consiste à tronquer les mots à leur racine en supprimant l'ensemble des préfixes et suffixes : par exemple, si l'on trouve dans les tweets soumis à la fouille de texte, les mots *flood*, *flooded*, *flooding*, et que l'on applique la racinisation, alors ces trois mots apparaîtront tous sous le radical *flood*. Si nous n'appliquons pas cette fonction de racinisation, c'est parce que nous souhaitons identifier les différents mots employés par les utilisateurs pour qualifier et décrire les phénomènes et événements du réel. En outre, lorsque les tweets sont nombreux, nous avons recours à l'analyse en *n-grammes* afin d'identifier les collocations lexicales (et ainsi recontextualiser l'emploi de chaque mot de vocabulaire, plutôt que d'observer des mots isolés). De la même manière que pendant l'étape de recherche lexicale, les résultats de la fouille de texte peuvent être affichés sous la forme de nuages de mots. Le package *spacyr* est utilisé afin de mener l'étape de *part-of-speech tagging* (étiquetage morpho-syntaxique) : son bon fonctionnement nécessite au préalable d'installer le module *SpaCy* sur *Python*. Combinées avec le *package tm*, les fonctions de *spacyr* permettront de construire des tableaux inventoriant chaque mot présent dans un corpus de tweets nettoyé, sa fréquence ainsi que sa catégorie grammaticale. Par le *package sqldf*, il sera alors possible de créer de nouveaux tableaux dans lesquels on ne sélectionnera que les mots appartenant à une catégorie grammaticale précise.

Si toutefois les tableaux et nuages de mots ne se montrent pas suffisants pour mettre en évidence une variabilité sémantique en fonction des périodes ou des lieux, nous envisageons d'avoir recours à la LDA comme méthode d'exploration alternative. Le *package topicmodels* de R dispose des fonctions nécessaires à la constitution de groupes de mots (les prétraitements des corpus de tweets restent identiques et exécutés via les fonctions du *package tm*). L'utilisateur doit choisir deux paramètres : le nombre de groupes de mots à former et le nombre de mots inclus dans chaque groupe. En testant la LDA, nous pourrons également mesurer les réserves énoncées par (Ghosh et Guha, 2013) quant à l'applicabilité de cette méthode d'analyse lexicale sur les tweets : "*They* [les topics] *do not all make a clear sense. Indeed, some topics may appear to be a random collection of terms, whereas other topics can contain a more cohesive collection of terms*". Il serait alors en effet intéressant de savoir si, dans la problématique des risques naturels, on observe des résultats identiques aux observations effectuées par ces auteurs ou si la sémantique des événements virtuels en réponse à une crise d'origine naturelle s'avère plus construite et cohérente.

4.3.3.2. Paramétrage des critères modulables

Les démarches et outils exposés précédemment requièrent, pour la plupart, de déterminer, à un moment précis de l'analyse, des critères de paramétrage définis par l'utilisateur : choisir une distance seuil pour agréger des tweets par l'algorithme DBSCAN, choisir une distance afin de créer un carroyage ou des zones tampon autour des tweets géolocalisés, choisir une résolution temporelle d'analyse ou encore le nombre de *topics* à former pour une LDA. Comment s'assurer que ces choix soient les moins hasardeux et surtout qu'ils ne biaisent pas, en cascade, les résultats d'une analyse, les observations faites et hypothèses posées sur ces résultats et au final, les tests définis à partir de ces hypothèses ? Ainsi, pour nous guider dans la détermination de ces divers seuils, nous appréhenderons les éléments suivants :

- étudie-t-on le tweet en dehors de tout contexte social ? si l'activité tweeting est analysée indépendamment de toute variable socio-démographique capturée à l'échelle d'une entité administrative ou statistique, on peut alors envisager de s'affranchir de ces unités spatiales. Dans le cas contraire, il serait préférable de se conformer à ces entités, et notamment lorsqu'il s'agit d'entités spatiales délimitées en fonction d'une population aux caractéristiques socio-démographiques homogènes.

- En ce qui concerne la construction de carroyages, le choix de la longueur du côté de la maille dépend en premier lieu de l'échelle considérée. Si l'on travaille à l'échelle de l'Etat du Texas (superficie de 696 000 km² [et donc supérieur à la France métropolitaine], environ 1 200 km de distance du nord au sud et d'est en ouest) et qu'on construit des mailles de 100 km de côté, alors la résolution spatiale d'agrégation des tweets est petite et on risque finalement de limiter les observations aux mailles vides ou pleines de tweets géolocalisés. A l'inverse, si l'on établit des mailles de 10 km de côté, alors on risque d'obtenir une partition du territoire trop complexe pour être lisible sur la carte à cette échelle. L'enjeu consiste ainsi à trouver la maille qui permette d'établir, à une échelle d'analyse considérée, la typologie des lieux en fonction de leur niveau d'activité virtuelle. Dans les démarches, on se propose alors d'appliquer différents niveaux de mailles, de la plus petite échelle à la plus fine, d'observer la variabilité des structures mise en évidence (continuité, transition, rupture, *etc*.) et de conserver le niveau de résolution qui conciliera lisibilité graphique et observations pertinentes.

- Le choix de la distance lors de l'utilisation de l'algorithme DBSCAN dépend également de l'échelle spatiale d'étude, à laquelle on pourra ajouter deux critères. Dans un premier temps, il sera nécessaire de mesurer, pour des groupes de tweets proches sur une cartographie d'échelle fine, la distance moyenne qui les sépare (on pourra éventuellement utiliser cette distance comme critère d'agrégation spatiale. Ensuite, la distance seuil peut également être appréhendée selon nos propres critères relatifs aux tweets ou objets du territoire que nous considérons comme proches dans un environnement perturbé par un phénomène naturel. Autrement dit, quelle est la distance maximale que nous tolérons pour

considérer la proximité entre objets du territoire local en crise ? Par exemple, si le niveau d'un cours d'eau augmente en un point donné, il n'est probablement pas pertinent d'associer un tweet situé à 100 mètres de la rive à un groupe de tweets situés à une dizaine de mètres (soit la marge d'incertitude du *GPS*) de cette même rive.

- Comment gérer le choix des pas de temps ? Nous pouvons procéder par palier en affinant progressivement la résolution temporelle (jours, heures, demi-heures, quarts d'heure, *etc*.) d'analyse. L'objectif consiste à identifier la résolution la plus fine à laquelle on pourra observer une variabilité de la distribution des flux de tweets de crise tout en vérifiant si, à chaque pas de temps où l'on identifie cette variabilité, l'on détecte également un sens dans la sémantique (c'est-à-dire qu'on peut identifier des tendances dans l'événement virtuel, qui révèlent la survenue de phénomènes ou événements réels, et non pas un enchevêtrement d'informations hétéroclites).

- Enfin, si nous testons la LDA, il nous faudra définir le nombre de *topics* à créer. Si l'on se réfère aux propos précédemment cités de (Ghosh et Guha, 2013), alors nous pouvons souligner l'enjeu suivant pour ce type d'analyse sémantique : si la plupart des *topics* contiennent des ensembles de mots-clés sans cohérence particulière, faut-il alors créer une dizaine de *topics* voire davantage pour mettre en évidence le peu de cohérence lexicale ou un nombre restreint de *topics* se montre-t-il suffisant ? En outre, si l'on ne parvient pas à faire émerger des tendances lexicales de l'activité tweeting, ni dans les nuages de mots, ni dans la LDA, faut-il alors remettre en question la pertinence de la composante sémantique (qui reste la composante généralement la moins étudiée) ?

Conclusion du chapitre 4

Ce chapitre s'est attaché à la description de la contribution, en termes d'outils et de démarches méthodologiques, focalisée sur l'extraction de tweets de crise géolocalisés d'une part, et d'autre part, sur l'exploration des possibilités de valorisation de ces tweets ainsi que de leurs apports et limites à la problématique géographique des risques naturels. En effet, si le volet de la gestion de crise en temps réel enregistre des apports humains indubitables, qu'en est-il du tweet géolocalisé en tant que marqueur géographique des territoires perturbés et de leurs populations, qui ne sont pas affectés de manière homogène par un phénomène naturel.

Les outils mobilisés restent les outils traditionnellement employés par le géographe géomaticien (auxquels on aura cependant ajouté les outils d'analyse sémantique) : bases de données spatiales, analyse d'une distribution, tests statistiques, outils de cartographie applicables aux données de forme ponctuelle et outils d'analyse spatiale mais, conformément à l'approche épistémologique décrite, l'appréhension de la cartographie s'avère différente. Comme nous l'avons présentée dans le premier chapitre de ce manuscrit à travers quelques exemples, la production cartographique actuelle oscille entre le simple fait d'afficher des données sur un fond de carte d'habillage (apparence de la carte typique du Géoweb), l'analyse de diverses sources de données brutes et leur transformation en une information traitée, de nature qualitative (carte de vigilance de Météo France) ou quantitative (indice de pollution atmosphérique). Quelles que soient les formes de ces cartes, elles ne sont généralement accompagnées que de quelques lignes de commentaires descriptifs mais elles ne servent pas de support réflexif. De même, du côté académique, dans le cadre de l'enseignement cartographique, l'étudiant est amené à fixer un cadre autour d'un sujet choisi et à résoudre une question dans une démarche hypothético-déductive ; la carte alimente alors un discours descriptif mesurant la validité de l'hypothèse de départ sur les territoires (Nunez Moscoso, 2013). Ici, on espère que les propriétés des traces numériques géolocalisées permettent de dépasser cette approche. Le projet d'analyse décrit dans le chapitre s'en détache donc : la représentation cartographique n'est pas un résultat mais une étape transitoire entre une question initiale de recherche (qui suscite la cartographie) et l'observation de structures suggérant un questionnement plus approfondi. De plus, l'un des enjeux de le représentation cartographique des tweets de crise géolocalisés consiste à intégrer leur dimension qualitative hétérogène à la carte.

Enfin, signalons qu'avant toute conception d'une interface exploratoire de géovisualisation des tweets de crise, il nous paraît essentiel d'identifier les outils et démarches qui permettront de valoriser le matériau et de les séparer des éventuelles impasses. Par conséquent, il conviendrait en premier lieu d'établir un guide des recommandations d'analyse, de la même manière que les pistes proposées en ce qui concerne le choix d'une méthode d'extraction des tweets de crise géolocalisés.

5. Spatialisation, temporalités et sémantique de l'événement virtuel : le tweet géolocalisé, un témoin systématique des phénomènes et événements du réel ?

Ce chapitre présente les résultats des premières pistes explorées à partir de la trace spatio-temporelle que constitue le tweet de crise géolocalisé. Il s'articule autour de la question fondamentale du matériau comme marqueur de la survenue des phénomènes physiques et des événements résultants dans le monde réel ainsi que de leur dynamique, de l'échelle globale à l'échelle locale de la perturbation. Il a alors comme objectif final de répondre à la question formulée dans le titre ci-dessus : le tweet de crise géolocalisé constitue-t-il un témoin systématique des interactions entre phénomènes naturels et populations affectées dans le monde réel ?

Dans un premier temps, ce chapitre présente les résultats de la démarche méthodologique d'extraction de jeux de tweets utiles à la problématique de recherche, celle-ci constituant en effet la toute première étape d'exploration du contenu lexical des tweets. Le chapitre fournit ainsi les listes de vocabulaire propre aux différents phénomènes explorés, mis en évidence par la recherche lexicale, ainsi que les caractéristiques des jeux de tweets extraits.

Dans un second temps, le chapitre expose les résultats des analyses spatio-temporelles et sémantiques effectuées en croisant tweets de crise extraits sur notre terrain d'étude et données provenant d'acteurs reconnus et officiels. Notre premier objectif consiste à explorer, de manière expérimentale, les cellules de réactivité virtuelle détectées en réponse à un phénomène réel affectant différents lieux à des intensités variées. Ce chapitre apportera alors les premières précisions sur les questions suivantes qui émergent en début de recherche : sur le terrain texan, où tweete-t-on dans les périodes de crise ? Peut-on distinguer une succession de phases propres à la crise rapportées dans l'événement virtuel ? Quelle phase draine le plus de tweets géolocalisés ? Observe-t-on une variabilité des foyers d'émission en fonction des types de phénomènes et de leur violence ? Les observations dressées à partir de ces questions sont-elles répétées ou variables en fonction des lieux et phénomènes ?

Par cohérence avec la démarche de recherche présentée dans le chapitre précédent, l'organisation de la restitution de ces résultats (dans ce chapitre comme dans le prochain) suit la logique de progression des analyses : cette restitution retrace donc l'évolution des questionnements soulevés et des hypothèses émises quant à l'identification de facteurs explicatifs des résultats observés, au fur et à mesure de l'avancée des travaux.

5.1. Retour sur les résultats de la démarche méthodologique d'extraction de tweets de crise géolocalisés

Cette première partie soumet les résultats de la méthodologie d'extraction lexicale et spatiale, présentée dans le chapitre 4. Les extractions de tweets de crise ont été effectuées de manière empirique, en suivant les étapes décrites précédemment et en associant les outils base de données, SIG et R. Par ailleurs, elles ont bien évidemment été testées afin de construire les jeux de tweets de crise liés aux phénomènes d'étude annoncés dans l'introduction (phénomènes hydrométéorologiques printanniers du Texas et ouragan Harvey) mais ont également été appliquées sur d'autres terrains pour d'autres phénomènes. Ainsi, en nous renseignant régulièrement sur les phénomènes météorologiques ou hydrométéorologiques remarquables survenus entre 2016 et 2017, nous avons effectué un total de quatre extractions de tweets de crise : tempête Jonas en 2016 au Etats-Unis, crue de la Seine dans le bassin parisien en juin 2016, pluies intenses et inondations au Texas au printemps 2016 et ouragan Harvey en août 2017 (ces deux derniers types de phénomènes constituant donc les sujets d'étude).

5.1.1. Apports des approches d'extraction lexicale

5.1.1.1. L'extraction lexicale par hashtags

L'extraction lexicale focalisée sur les hashtags seuls a été appliquée sur un phénomène extrême rare dont nous souhaitions mesurer *l'emprise spatiale virtuelle globale*, la tempête de blizzard Jonas ayant affecté le nord-est des Etats-Unis entre le 21 et le 29 janvier 2016. Ce phénomène extrême ayant été largement relayé par les médias à l'échelle internationale d'une part, et ayant affecté des foyers de peuplement des Etats-Unis d'autre part, nous pouvons nous assurer que le volume de tweets émis en réponse ne sera pas un verrou quantitatif à l'extraction de tweets de crise.

Pour ce premier test, nous cherchons à visualiser l'emprise spatiale virtuelle du phénomène à l'échelle du pays entier (l'espace d'extraction de référence est donc constitué d'une entité administrative, à savoir les frontières nationales des Etats-Unis) ; en ce qui concerne les dates, nous fixons l'intervalle temporel d'extraction aux dates précises de survenue de la tempête (indiquées ci-dessus). Ces deux conditions correspondent à une requête d'extraction d'un jeu de tweets bruts constitué de 2 032 125 entités. Comme indiqué dans le chapitre 4, la première requête d'extraction lexicale de tweets de crise a été focalisée sur la base de trois hashtags cibles (correspondant aux surnoms couramment employés pour désigner une tempête de blizzard sur le continent nord-américain) : *#snowzilla*, *#snowpocalypse* et *#snowmageddon*. Cette première requête retourne un total de 4 986 tweets de crise. Nous avons ensuite effectué deux nouvelles étapes de recherche de hashtags

cibles afin d'enrichir cette liste : le graphique de la figure 5.1 ci-dessous indique ainsi le nombre de tweets retournés en fonction du nombre de hashtags cibles, mais également le nombre total de hashtags contenus dans chaque jeu de tweets de crise retournés en fonction du nombre de hashtags recherchés :

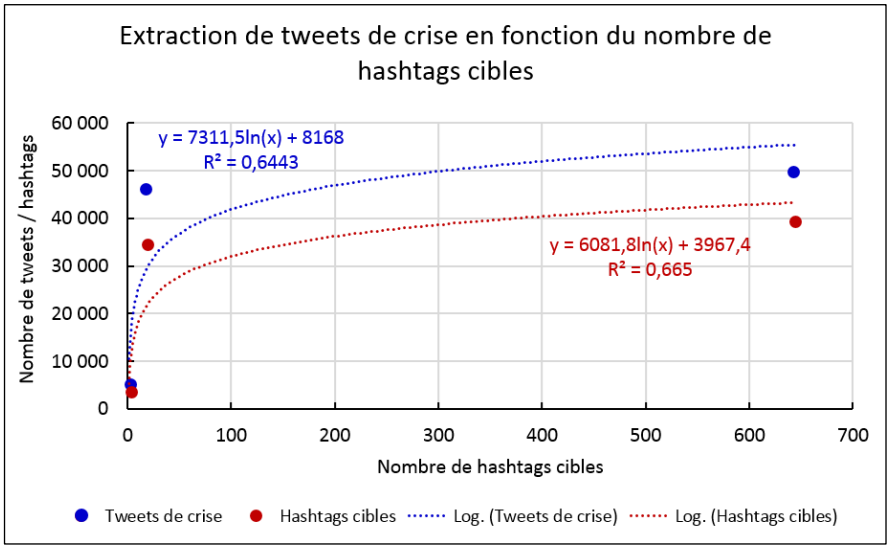


Figure 5.1 : Nombre de tweets et de hashtags retournés en fonction du nombre de hashtags cibles de recherche

On distingue bien le comportement que nous avions supposé dans le chapitre 4 : le nombre de tweets (ou de hashtags) inclus dans un jeu de crise en fonction du nombre de hashtags cibles, suit une relation d'ordre logarithmique plutôt fiable (les coefficients de détermination R^2 oscillent autour de 65% de variance prise en compte par les modèle de régression). En outre, on pourra remarquer que la première analyse isolant les hashtags contenus dans le jeu initial de 4 986 tweets de crise constitue l'étape déterminante fournissant les hashtags massivement diffusés (et qui en fait ne correspondent pas aux trois surnoms recherchés sur le Web avant la première extraction) : en recherchant 18 hashtags, on multiplie le nombre de tweets de crise extraits par 9 (et cette recherche fournit 92,6% du jeu de tweets de crise final, constitué de 49 826 entités). En revanche, lorsqu'on étend la recherche de hashtags, le volume de tweets de crise extraits ne témoigne plus du même ordre de grandeur : entre la deuxième et la troisième extractions, on ne multiplie la quantité de tweets de crise que par 1,08 et les derniers tweets extraits ne représentent que 7,4% du jeu de crise final. L'expérience menée ici semble alors confirmer les résultats de l'étude publiée par (Lin *et al.*, 2013). En effet, si l'on observe le comportement des hashtags dans le jeu de tweets de crise extraits à partir des 632 hashtags (figure 5.2), on pourra alors constater que 75% des hashtags contenus ne sont employés tout au plus que deux fois ; en fait, les principaux hashtags qui permettent d'extraire un maximum de tweets (ceux que [Lin *et al.*, 2013] qualifiaient de *winners*) correspondent aux points hors-normes du graphique de gauche : ainsi, 2% des 632 hashtags sont contenus dans 86% des tweets du jeu de crise final.

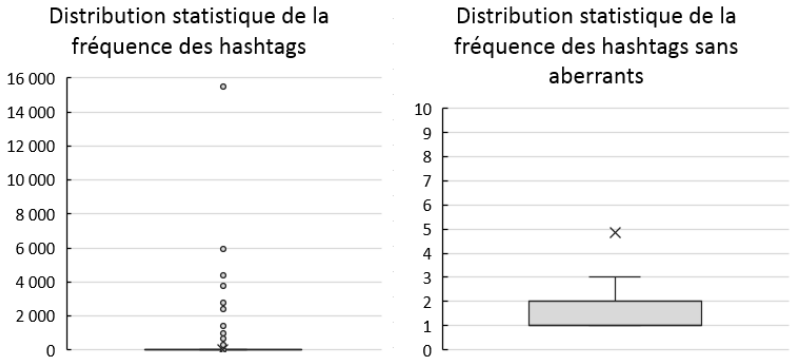


Figure 5.2 : Distribution statistique des hashtags dans le jeu de tweets de crise final

Face à ce comportement de distribution, comment avons-nous sélectionné les hashtags de filtrage ? Pour procéder à la deuxième extraction lexicale, nous avons sélectionné les hashtags dont l'occurrence, dans le premier jeu de 4 986 tweets de crise, était supérieure ou égale à 100 : *#snowzilla*, *#snowpocalypse*, *#snowmageddon*, *#blizzard2016*, *#snowmageddon2016*, *#jonas*, *#blizzard*, *#snow*, *#dc*, *#snowday*, *#snowzilla2016*, *#snowzilla16*, *#winterstorm*, *#washingtondc*, *#snowstorm*, *#dcsnow*. La liste des hashtags contenus dans le deuxième jeu de 46 135 tweets de crise se révélait trop conséquente pour être parcourue manuellement (34 540 hashtags isolés) ; cependant, les hashtags les plus fréquents de cette nouvelle liste correspondent à ceux que nous avons recherchés lors de l'étape précédente[1]. Nous avons alors ajouté l'ensemble des hashtags dérivés contenant les radicaux *#storm*, *#blizzard* et *#snow*, dont voici quelques exemples : *#snowblizzard*, *#snowing*, *#snowshovelling*, *#blizzardjonas*, *#blizzardwatch*, *#blizzardwarning*, *#blizzardready*, *#blizzardcoming*, *#stormjonas2016*, *#stormwatch*, *#stormyday*, *#stormiscoming*, *#stormageddon*, *etc*.

La figure 5.3 restitue la carte des densités de ces tweets par mailles de 100 km² : sans surprise, les foyers d'émission des tweets de crise géolocalisés se concentrent de manière continue sur la côte nord-est du pays, et principalement dans les deux aires métropolitaines de New-York et de la capitale Washington D.C. L'événement virtuel s'avère en revanche diffus dans l'ensemble du pays : on retrouve des foyers d'émission secondaires dans d'autres aires métropolitaines (Los Angeles, Seattle, Denver, *etc*.) mais également dans des territoires bien moins densément peuplés ; deux d'entre eux ont été encadrés sur la carte et correspondent aux villes de South Lake Tahoe (Californie et Nevada) et de Flagstatt (Arizona)[2], toutes deux entourées de territoires forestiers et montagneux.

[1] Les plus fréquents restent d'ailleurs les hashtags *#snow* (15 150 occurrences), *#blizzard2016* (5 888 occurrences, et enfin *#snowday* (4 255 occurrences).

[2] Si l'on se réfère à la méthodologie de questionnement énoncée dans le chapitre 4, ce sont ces lieux qu'il conviendrait d'explorer en priorité : il est intuitif de penser que les métropoles s'activent en réponse, par exemple, aux perturbations du trafic aérien. Pourquoi observe-t-on, dans ces territoires moins peuplés et moins

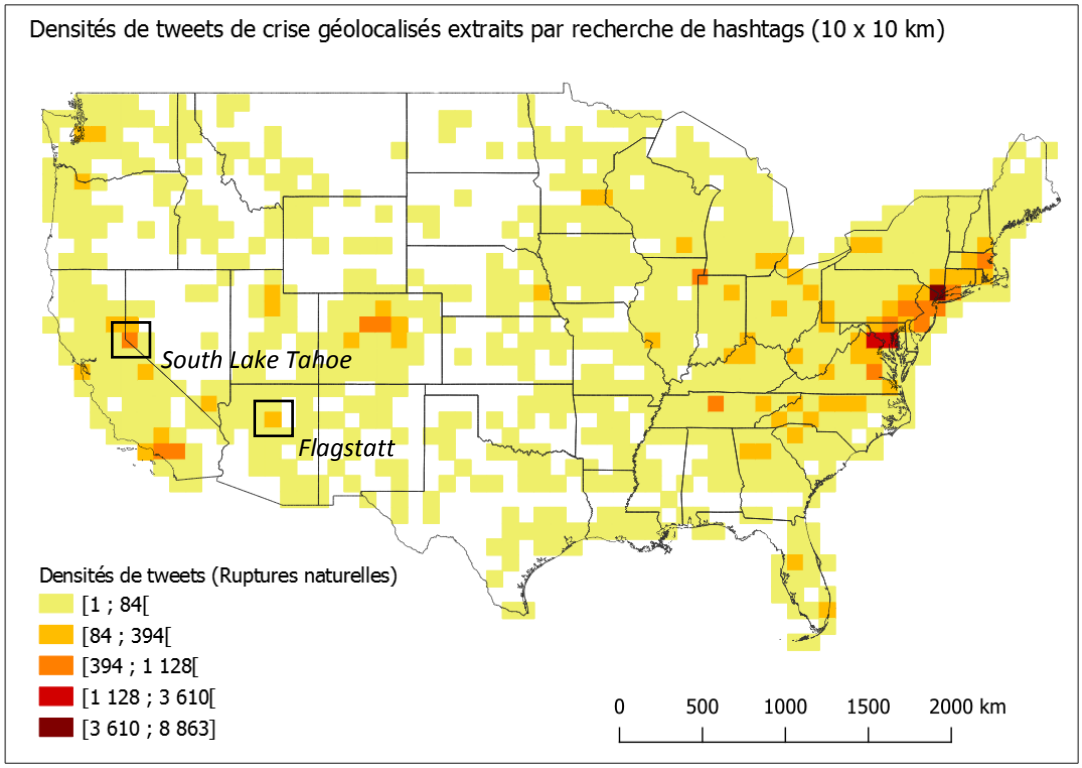


Figure 5.3 : Carte des densités de tweets de crise géolocalisés extraits sur la tempête Jonas dans l'ensemble des Etats-Unis (C.Cavalière)

5.1.1.2. L'extraction lexicale par un ensemble de mots-clés

La seconde méthode de recherche et d'extraction lexicale a été appliquée aux phénomènes saisonniers associant pluies intenses et inondations au Texas, survenus au cours du printemps 2016. Ici, l'enjeu est différent : ces phénomènes, bien que dommageables, n'ont pas la même portée médiatique que les tempêtes de blizzard ou encore les ouragans. Pour nous, il ne s'agit plus de visualiser l'ampleur de l'événement virtuel consécutif à ces phénomènes locaux, dans le pays entier, mais de constituer le jeu le plus complet possible afin de disposer de *suffisamment*[3] de matériau tweet à soumettre à des analyses plus poussées (toujours dans l'objectif de favoriser l'émergence du questionnement de recherche et la formulation d'hypothèses). C'est pourquoi, dans ce deuxième test, nous adoptons la méthode de recherche lexicale toutes catégories de mots-clés confondus (hashtags comme mots-clés simples ou associations lexicales).

Le terrain est alors fixé aux frontières administratives de l'Etat du Texas ; en ce qui concerne la sélection des dates de l'intervalle temporel, nous nous sommes référés à la presse en ligne américaine locale (indiquant la survenue de ces phénomènes de pluies-inondations à

urbains, une réponse supérieure à certaines métropoles ? Par exemple, Phoenix (capitale de l'Arizona), témoigne d'une réponse moins importante que la cellule d'activité détectée autour de Flagstatt, plus au nord de cet Etat.

[3] L'emploi de cet adverbe peut être discuté : à ce stade, nous ne savons pas si un foyer conséquent de tweets se montrera plus révélateur des dynamiques de terrain en cours qu'une poignée de tweets.

des intensités variables entre mars et juin 2016). Les dates de début et de fin d'extraction sont ainsi fixées au 1er mars 2016 et au 30 juin 2016. La requête d'extraction soumise avec ces deux conditions retourne un jeu brut constitué de 2 251 870 tweets géolocalisés. La première requête d'extraction lexicale de tweets de crise a été effectuée sur une base de recherche de trois mots-clés : *flood*, *storm*, et *tornado* ; elle retourne 7 782 tweets dans un premier jeu de tweets de crise géolocalisés. L'étape suivante, qui correspond à la première phase de recherche non supervisée de mots-clés et d'associations lexicales, témoigne du même comportement identifié précédemment avec les hashtags : à partir d'une poignée de mots-clés, on pourra mettre en exergue du vocabulaire de crise massivement adopté et diffusé sur le réseau. La deuxième liste d'extraction contient alors un ensemble de 53 mots-clés et de 23 associations lexicales (mots associés par l'esperluette), listés dans la figure 5.4 ci-dessous, et classifiés en fonction de cinq grands thème identifiés (vocabulaire lié aux inondations, cours d'eau, conditions météorologiques, diffusion d'alertes ou de vigilances ou encore effets sociaux des crises naturelles comme les recommandations comportementales, les évacuations et les secours) :

<table>
<tr><th>Inondations</th><th>Conditions météorologiques</th><th>Alertes</th><th>Effets sociaux</th></tr>
<tr><td rowspan="3">flooded, floodvictims, flood & victim, flood & victims, houstonflood, houstonfloods, houstonflooding, texasflood, texasfloods, texasflooding, flooddamaged, flooddamages, flasflood, floodstreet, floodravaged, floodroad, road & flooded, floodwater, floodwaters, flooding & waters houstonfloodwaters, floodmageddon, highwater, rising & water</td><td>rain, rainfall, rainfalls, rains, raining, lightning, lightnings, thunder, thunderstorm, thunderstorms, hail, stormchasing, stormy&day, stormyday, texasstorm, texasstorms, stormyweather, stormy&weather, severe&weather</td><td rowspan="3">stormwatch, storm & watch, floodwatch, flood & watch, tornadowatch, tornado & watch, stormwarning, storm & warning, floodwarning, flood & warning, tornadowarning, tornado & warning</td><td rowspan="3">disaster, disasters, rescue, safety, safetyfirst, turnarounddontdrown, chasers & report, people & affected, enforcement & reports, evacuation & rescue, evacuation & apartment, evacuation & home, emergency, shelter, shelters, vigilant, nationalguard, national & guard</td></tr>
<tr><td>Cours d'eau</td></tr>
<tr><td>buffalo & bayou, brazos & river</td></tr>
</table>

Figure 5.4 : Deuxième liste d'extraction pour l'extension du jeu de tweets de crise

La deuxième requête d'extraction de tweets de crise, tout comme avec les hashtags, s'avère de nouveau décisive, dans la mesure où elle augmente considérablement le volume de tweets de crise extraits : le deuxième jeu de tweets de crise contient ainsi un total de 43 535 tweets, il est donc 5,6 fois plus important que le précédent. L'étape combinant fouille sémantique et requête d'extraction lexicale a été accomplie une dernière fois d'une part, à partir de la recherche des mots à faible occurrence, contenus dans le deuxième jeu de 43 535 tweets de crise et d'autre part, via l'utilisation du package *wordnet* de R afin de rechercher des synonymes (au mot *flood* par exemple). Nous avons ainsi relevé une poignée de nouveaux mots-clés ajoutés à la liste précédente : *deluge*, *apocalypse*, *school & canceled*, *school & closed*, *rising & tide*, *overflow*, *overflowing*, *outpouring*, *torrent*, *torrential*, *power & cut*, *flooded & roadways*, *zone & affected*, *relief*, *stuck & help*, *dangerous & died*. Le troisième jeu de tweets de crise extraits en ajoutant ces nouveaux mots et associations contient au final 46 192 tweets de crise (soit 2% du jeu de tweets bruts collectés initialement). Un nouveau test de modélisation (du lien entre nombre de mots ou expressions cibles et nombre de tweets de

crise retournés) est lancé afin de vérifier si le comportement identifié précédemment se répète. Les résultats se montrent cette fois-ci plus contrastés (figure 5.5) :

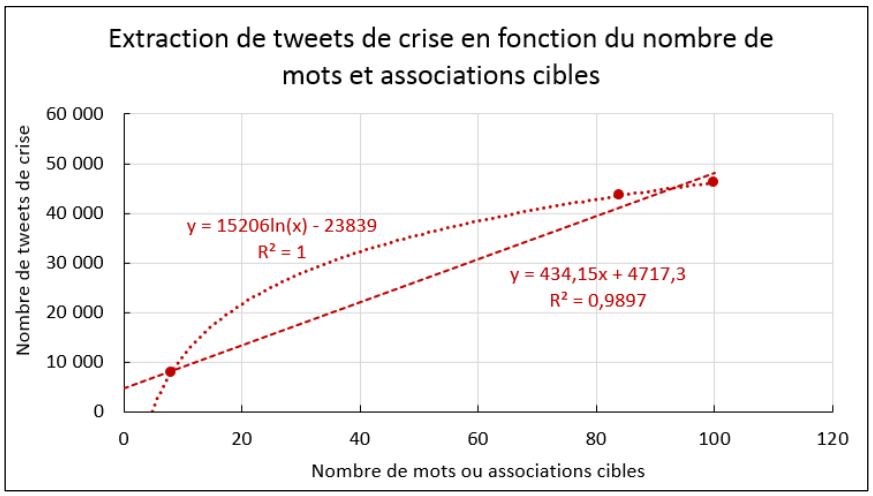


Figure 5.5 : Nombre de tweets retournés en fonction du nombre de mots-clés et d'associations lexicales cibles de recherche

Au regard de la situation des points, on peut tester deux modèles : le linéaire et le logarithmique, tous deux témoignant de forts coefficients de détermination (R^2 de 0,98 pour le modèle linéaire et modèle quasi ajusté pour la régression logarithmique). Doit-on alors considérer que les tweets de crise retournés recèlent encore un potentiel de nouveau vocabulaire massivement relayé ou que les résultats observés proviennent de la proximité des deux derniers points sur le graphique ? (dans le cas des hashtags, on passait de 18 à 632 cibles entre la deuxième et la troisième extractions alors que dans ce nouvel exemple, on passe seulement de 84 à 100 cibles lors de la même étape)[4]. En outre, il apparaît, dans les deux exemples testés jusqu'ici, que l'identification de l'effet de masse a lieu après la première extraction, lors de la première étape d'analyse lexicale non supervisée.

Dans le cas du Texas, nous nous sommes attachés à l'examen des profils d'utilisateurs en fonction de leur empreinte virtuelle sur Twitter, en période de crise (figure 5.6). L'allure de la distribution statistique se révèle similaire à celle que nous avons décrite pour l'emploi des hashtags dans les tweets de crise de la tempête Jonas.

[4] Dans tous les cas, les nuages n'étant constitués que de trois points, les résultats sont à considérer avec un certain recul, même si un comportement s'avère répété.

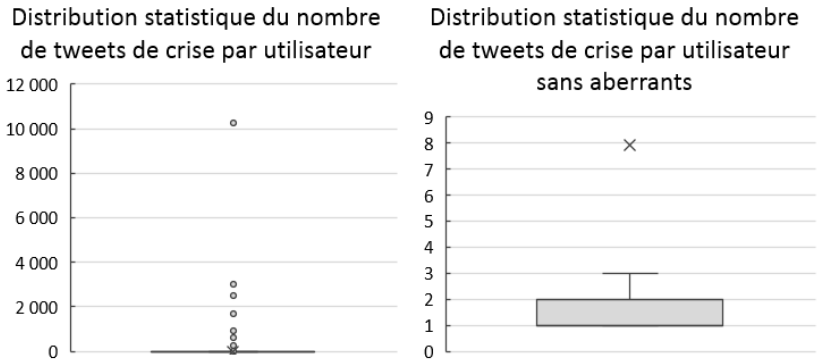


Figure 5.6 : Distribution statistique de la participation des utilisateurs dans le jeu de tweets de crise final

Le jeu final de tweets de crise contient 5 848 utilisateurs dont 75% n'ont tweeté tout au plus que deux fois pendant la période totale des quatre mois de collecte (de la même manière que 75% des hashtags n'étaient employés que deux fois dans le jeu de la tempête Jonas). En fait, ce sont les individus hors-normes, représentés par les points du graphique situé à gauche, qui contribuent à la création d'un jeu de tweets de crise : ici, 3,5% des utilisateurs enregistrés ont produit 82,13% des tweets extraits dans le jeu final (dont un utilisateur enregistre à lui seul 22% des tweets extraits). A qui correspondent ces utilisateurs au profil hors normes (puisque 75% des utilisateurs n'affichent qu'une participation sporadique à la création de tweets de crise) ? Les étapes d'exploration lexicale nous ont fourni les informations nécessaires : les comptes rassemblant plusieurs milliers de tweets de crise émis pendant la période de collecte correspondent à l'activité d'acteurs officiels. Nous en avons identifié deux principaux : les bulletins des centres du *NWS* qui émettent des alertes ou des vigilances face à des phénomènes prévus (crues éclair, orages violents, tornades) ainsi que les tweets envoyés par les automates associés aux stations météorologiques ou de jaugeage des cours d'eau, déployées par l'*USGS*. Le contenu de ces tweets se distingue facilement dans les nuages de mots utilisés pour les besoins de la visualisation de l'exploration lexicale (figure 5.7) : en effet, la syntaxe de ces messages reste identique et ceux-ci contiennent toujours l'acronyme de l'organisme d'origine. Dans la figure 5.7, on peut par exemple rapidement distinguer la forte empreinte virtuelle du *NWS* dans la diffusion de messages d'alerte aux inondations (ici, il s'agit de l'association lexicale *nws flood stage*) indiquant qu'un cours d'eau a dépassé son seuil d'alerte défini par le *NWS*.

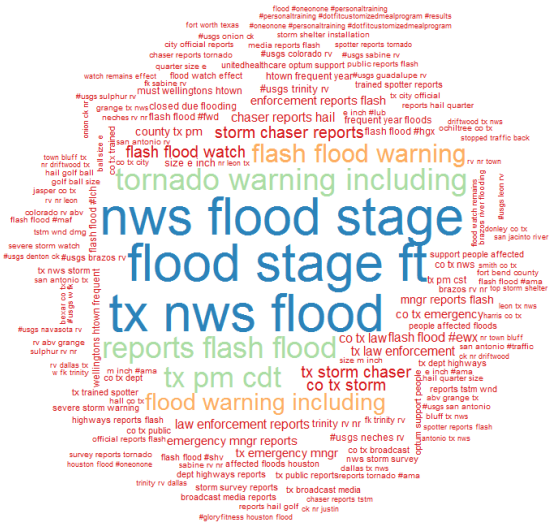


Figure 5.7 : Nuages de mots avec vocabulaire d'acteurs officiels

Dans les analyses, nous considérerons alors ces tweets géolocalisés comme des marqueurs de réalité terrain. En conséquence, nous avons identifié l'ensemble des comptes du *NWS* et de l'*USGS* présents dans notre jeu de tweets de crise et nous les avons isolés du jeu de tweets principal. Nous avons déjà évoqué ces types de tweets officiels dans l'introduction et le premier chapitre du manuscrit mais, pour rappel, ils prennent les formes suivantes :

- les tweets de l'*USGS* sont émis automatiquement dès qu'un seuil de cumul pluviométrique ou de hauteur d'eau (s'il s'agit d'une station de jaugeage) est dépassé. Ils se présentent sous cette forme : "*#usgs07301300* [code identifiant de la station] - *n fk red rv* [rivière] *nr shamrock, tx* [localisation] *is above nws flood stage (5ft)* [la hauteur d'eau a dépassé le seuil de mise en alerte de 5 pieds = 1,54 mètre]". Pour des précipitations violentes : *"#usgs07300000* [code de la station] - *salt fk red rv nr wellington, tx* [localisation] *heavy rain (0.27 in/hr) over the last 15 minutes* [survenue de précipitations intenses dans le quart d'heure écoulé, de 0,27 pouce par heure, soit 7mm/h]". Nous avons extrait 3 832 tweets émis par des automates de l'*USGS*.

- Les tweets du *NWS* diffusent un bulletin d'alerte ou de vigilance face à un phénomène prévu, sur un territoire donné : "*flash flood warning including cleburne tx, burleson tx, keene tx until 1:15 pm cdt @nwsfortworth* [tweet émis par le centre *NWS* de Dallas-Fort Worth diffusant une alerte crue éclair incluant les villes de Cleburne, Burleston et Keene, valable jusqu'à 13h15]". Nous avons extrait 5 427 tweets émis par des centres du *NWS*.

5.1.2. Apports de l'approche de l'extraction spatiale

5.1.2.1. L'extraction spatiale focalisée sur la définition d'une région et d'objets d'intérêt du territoire

Ce troisième test a été appliqué sur la France, et plus précisément dans le bassin parisien à l'occasion de la crue de la Seine, d'occurrence centenale, survenue en juin 2016. La région d'intérêt (ROI) initiale est ici définie par la création, sous SIG, d'une zone tampon d'une distance de cent mètres à chaque rive du fleuve ; nous avons alors collecté les tweets bruts contenus dans cette zone tampon et émis entre le 31 mai et le 15 juin 2016 (soit 1 200 tweets bruts). Cette fois, nous avons adapté le processus d'analyse lexicale : au lieu de lancer une extraction d'associations lexicales en bi- ou tri-grammes, nous avons testé la méthode de la *LDA* afin de mesurer la possibilité d'identifier la présence de thèmes cohérents dans le jeu de 1 200 tweets, dont le thème du fleuve en crue. Le résultat s'est montré positif (figure 5.8).

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	plus	change	champagne	pont	#creativeday	#ps4
2	île	#crueseine	bercy	alma	night	#pc
3	#job	#digital	euro	#paris	from	#xbox
4	here	#pointcrue	josselin	#crueparis	#tourdargent	trailer
5	#defrancelevalloisperret	better	amp	#quaideseine	oferattali	one
6	#pharmaceutical	are	jean	grenelle	popsiit	nouvelle
7	quintiles	you	you	#zouave	#tourdargentparis	xbox
8	recherche	#insolite	2016	concorde	bien	sortie
9	toujours	#pointdécrue	#bourgogne	#parisjetaime	for	jeu
10	the	arrive	ombiasypr	#cruedelaseine	chez	gameplay

Figure 5.8 : Résultats de la LDA paramétrée en six topics de 10 mots-clés

Dans les *topics* formés, on peut en effet distinguer plusieurs thèmes qui se manifestent par la cohérence des mots inclus dans le *topic* : offres d'emploi (*topic 1*), restaurant (*topic 5*), jeux vidéos (*topic 6*), *etc*. La crue s'avère présente parmi les sujets de discussion virtuels, dans les *topics* 2 et 4 : on identifie d'ores-et-déjà le hashtag *#crueseine*, mais également des lieux et objets du territoire intégrés dans les représentations collectives comme marqueurs d'intensité du phénomène (le Zouave du Pont de l'Alma, mentionné dans le *topic* 4). Le premier jeu de tweets de crise est alors extrait à partir des mots-clés et hashtags présentés dans la figure 5.9, auxquels on a ajouté du vocabulaire mis en évidence en faisant varier les paramètres de la *LDA* :

Hashtags	*Mots-clés*
#crueseine, #pointcrue, #pointdécrue, #crueparis, #cruedelaseine, #seineencrue, #crue, #laseine, #flooding, #flood, #parisfloods, #underwater, #innondation, #inondation, #quaideseine, #fluctuatnecmergitur	seine, siene (faute d'orthographe dans le nom), river, crue, inondation, flood

Figure 5.9 : Première liste d'extraction de tweets de crise parmi les tweets bruts émis à moins de cent mètres du fleuve

Après application du filtrage lexical, le premier jeu de tweets de crise extraits contient un total de 364 tweets de crise. Pour la suite des étapes d'extraction, nous modifions notre définition de la région d'intérêt. Chaque tweet de crise extrait devient un nouvel objet d'intérêt (OI) du territoire et nous définissons une nouvelle région d'intérêt par une zone tampon de 200 mètres autour de chaque OI. Dans cette deuxième ROI, nous extrayons un total de 8 353 tweets bruts que nous filtrons de nouveau par l'ensemble de hashtags et mots-clés présentés dans la figure 5.9. Ce nouveau jeu est ensuite analysé par les outils de fouille de texte afin d'identifier d'éventuels nouveaux mots ou hashtags : *sous & eau*, *under & water*, *flooded*, *river & banks, #parissousleseaux*, *#parisunderwater*, *#alertecrue* sont ainsi ajoutés à la liste de filtrage. Après la deuxième extraction, le jeu de crise contient un total de 554 tweets. On élargit alors la ROI à une distance de 400 mètres autour de chaque objet (extraction de 15 057 tweets bruts). Cette étape aura permis de mettre en évidence encore quelques hashtags qui n'étaient pas identifiés dans un périmètre plus proche du fleuve : *#pointcrueseine*, *#floodedparis*, *#parisisdrowning*, *#cruedeseine*, *#débordement*, *#grandecrue*, et *#crueseine2016*. Cette étape aura mis en évidence 120 nouveaux tweets, pour un total de 674 tweets de crise collectés. Nous avons interrompu le processus à cette étape et représenté le nombre de tweets extraits en fonction des distances de recherche, ainsi que le nombre de nouveaux tweets ajoutés au jeu de crise en fonction de ces mêmes distances (figure 5.10) :

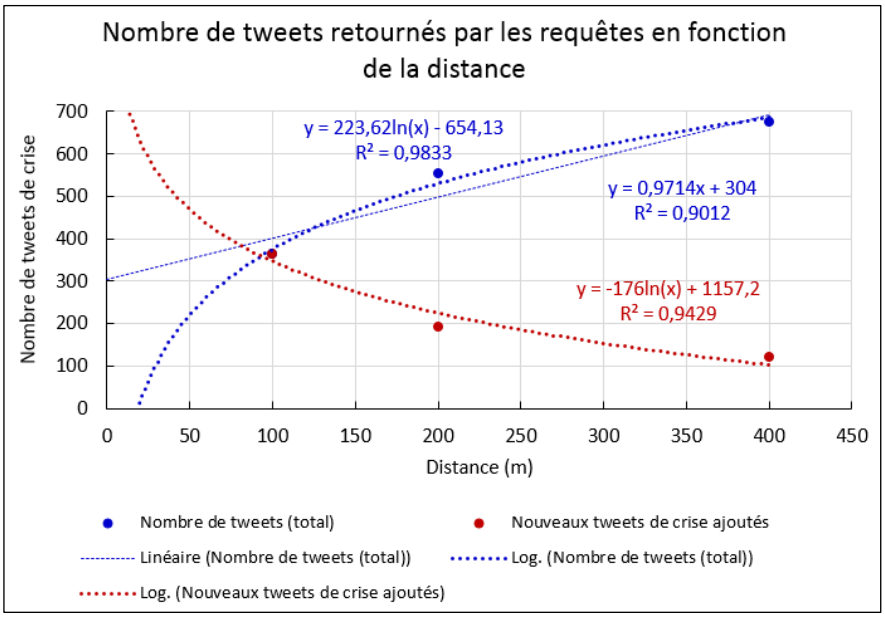


Figure 5.10 : Nombre total de tweets et nombre de nouveaux tweets retournés en fonction de la distance

A travers cet exemple, on peut de nouveau modéliser, à la fois pour le nombre total de tweets extraits et le nombre de nouveaux tweets mis en évidence en fonction de la distance à un objet d'intérêt, une relation d'ordre logarithmique qui reste la mieux ajustée aux points : on observe la formation de ce même palier (mais cette fois, avec un moindre contraste entre les quantités de tweets de crise retournés entre la première et la deuxième étape) ainsi qu'une décroissance effective du nombre de nouveaux tweets de crise ajoutés au corpus en s'éloignant progressivement de la région d'intérêt d'origine (zone tampon de cent mètres à chaque rive du fleuve). Pour terminer, la figure 5.11 affiche le corpus de tweets de crise constitué ainsi que l'ensemble des régions d'intérêt créées :

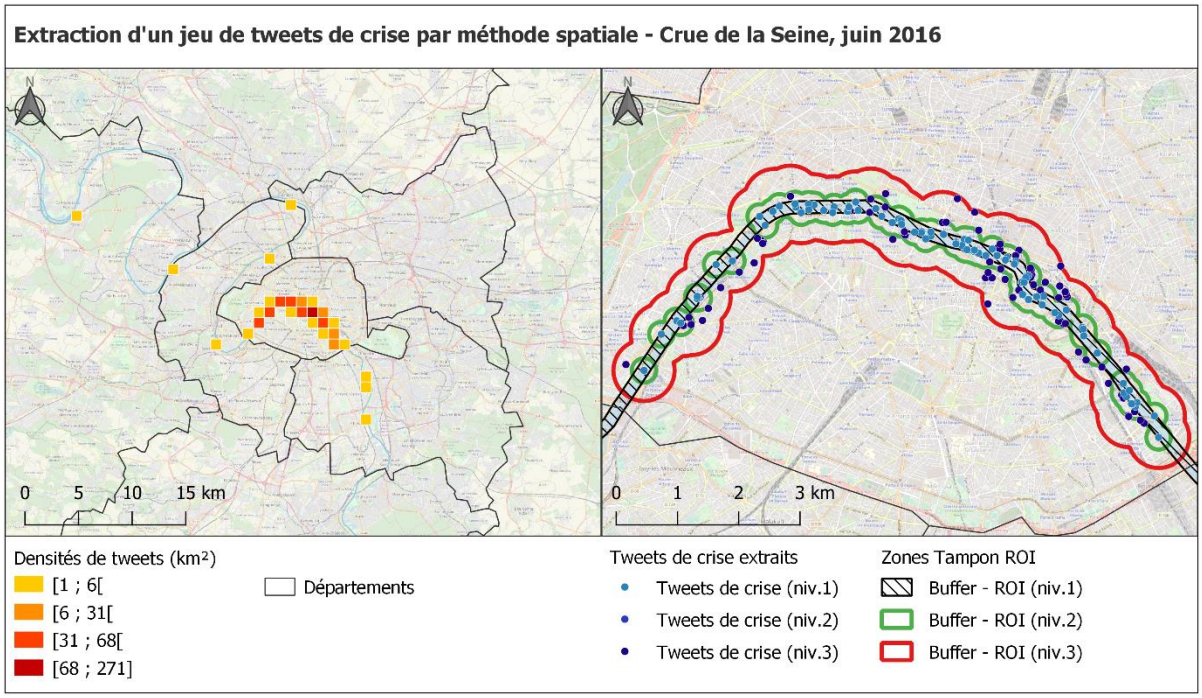


Figure 5.11 : Densités de tweets de crise extraits sur le terrain global et étapes de constitution du même jeu sur Paris (C.Cavalière)

Un premier constat est frappant sur cette carte finale : bien que nous n'ayons pas limité la recherche initiale de tweets bruts à l'Ile-de-France, les tweets de crise géolocalisés extraits sur le critère de la proximité au fleuve ne sont perceptibles que dans cette région et plus particulièrement dans la capitale (en Normandie, on n'a détecté aucune activité virtuelle géolocalisée fondée sur ce même critère de proximité). La maille concentrant la plus forte densité de tweets de crise géolocalisés correspond à l'Ile de la Cité. On peut introduire deux facteurs explicatifs à ce constat : en premier lieu, si l'on observe la distribution globale des tweets de crise dans la capitale en la comparant aux éléments du territoire, on remarque rapidement qu'une grande partie des tweets géolocalisés sont émis depuis les lieux d'observation idéale du fleuve : ponts et passerelles, pontons des ports et voies sur berges concentrent ainsi 51% du jeu de tweets de crise ; l'Ile de la Cité est connectée aux deux rives par neuf ponts dont le pont Neuf qui cumule 47 tweets de crise et le pont Notre-Dame qui en cumule 123. Dans un second temps, après avoir parcouru l'ensemble du jeu de tweets de crise, il s'avère que 36% des tweets émis sont rédigés dans une langue autre que le français, indiquant ainsi une participation virtuelle à la crise par le tourisme ; sur l'Ile de la Cité, cette participation issue des visites touristiques augmente à 45%. Ces deux facteurs peuvent ainsi expliquer la surreprésentation du lieu dans l'événement virtuel consécutif à la crue du fleuve.

5.1.2.2. L'extraction spatiale focalisée sur la définition d'une région d'intérêt unique

Le quatrième test effectué correspond à l'extraction d'un jeu de tweets de crise pour l'ouragan Harvey. Pour cette dernière application, nous avons dans un premier temps extrait un jeu de 362 336 tweets bruts émis entre le 20 août et le 3 septembre 2017 sur les Etats du Texas et de la Louisiane. Ce jeu brut a ensuite été filtré à l'aide de la recherche lexicale seule fondée sur cinq mots-clés : *hurricane*, *#harvey*, *flood*, *storm* et *rain.* Le premier jeu de tweets de crise extraits contient 14 759 entités. Les tweets de crise étant localisés à la fois dans des lieux directement affectés et en dehors de tout espace ayant subi l'ouragan de plein fouet, nous avons alors redéfini la région d'extraction des tweets de crise. Par conséquent, nous avons eu recours aux données physiques externes, en l'occurrence les grilles du *NWS*, de 4km de côté, modélisant les cumuls pluviométriques (cf. chapitre 4), afin de construire une nouvelle région d'intérêt, physique et non administrative, sur laquelle rechercher l'information lexicale contenue dans les tweets. Nous avons téléchargé la grille représentant, pour l'ensemble des Etats-Unis, les cumuls pluviométriques enregistrés entre le 23 et le 31 août 2017. Après conversion des cumuls en mm (les données d'origine figurant en pouces), nous avons mis en évidence les mailles ayant reçu plus de 322 mm d'eau pendant les neuf jours (afin de nous concentrer sur les territoires du Texas et de la Louisiane les plus affectés) ; les mailles sont ensuite converties sous format vectoriel afin de faciliter la sélection spatiale des tweets (figure 5.12).

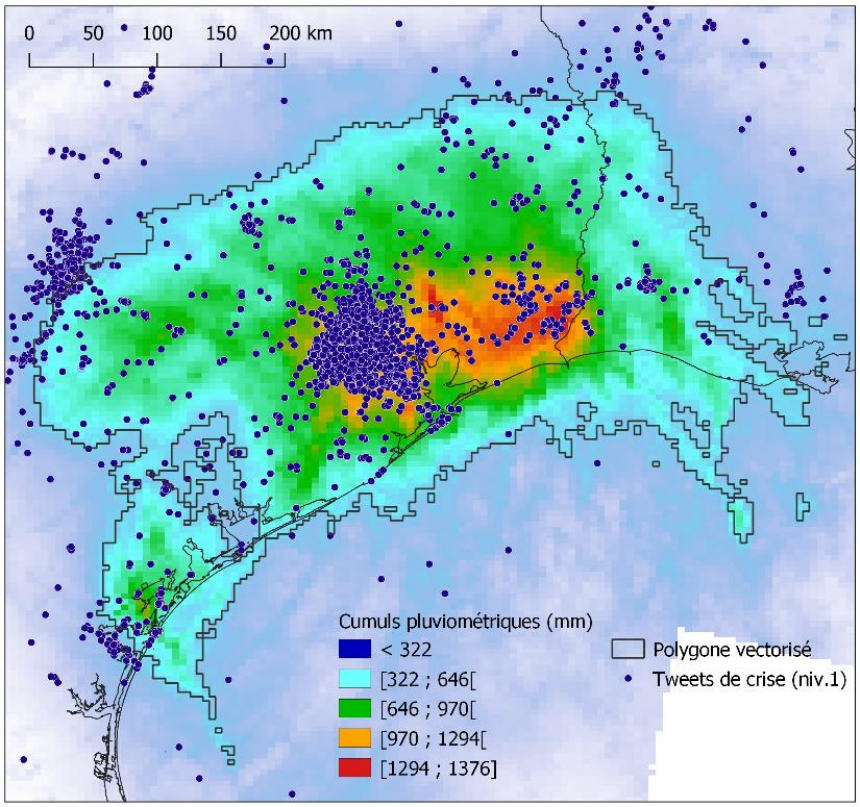


Figure 5.12 : Construction d'une région d'intérêt physique pour la sélection spatiale des tweets de crise géolocalisés

Après sélection spatiale, le premier jeu de tweets de crise est restreint à 8 555 entités incluses dans le polygone. Nous avons alors lancé une deuxième ainsi qu'une troisième étapes associant recherche et requête de filtrage lexical mais cette fois-ci, nous avons exploré des moyens alternatifs de représentation graphique du contenu lexical du jeu de 8 555 tweets de crise. La première solution consiste à simplement représenter graphiquement chaque mot ou hahstag isolé dans le corpus en fonction de son nombre d'occurrences : la figure 5.13 affiche ainsi les mots isolés dans le jeu de 8 555 tweets de crise, sous forme de points rouges, en fonction de leur occurrence (les deux axes du graphique représentent le nombre d'occurrences de chaque mot isolé). Dans ce cas, nous pouvons identifier les mots les plus fréquents (catégorie *topwords* sur la figure 5.13) mais également nous concentrer sur le vocabulaire épars (catégorie *scarce words* de la figure 5.13) :

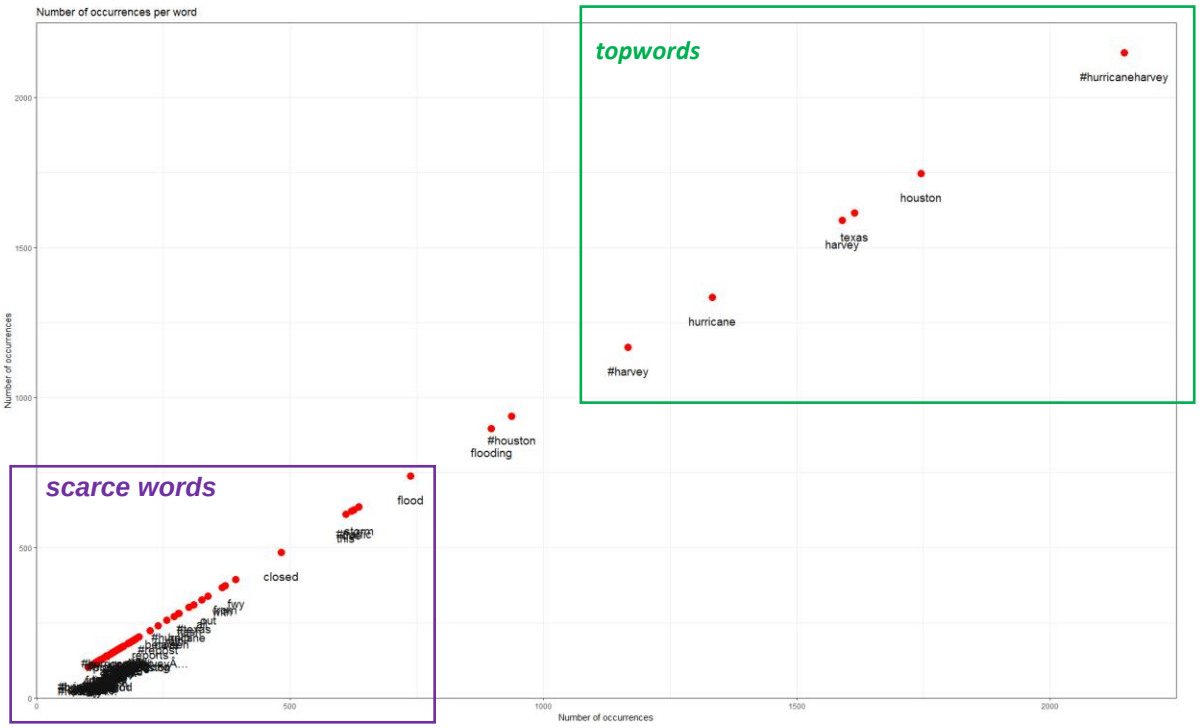


Figure 5.13 : Fréquence des mots isolés dans le premier jeu de tweets de crise

On peut distinguer le comportement qu'on avait précédemment mis en évidence par l'étude des hashtags de la tempête Jonas : même si l'on traite ici de l'ensemble des mots, on constate qu'une poignée de mots-clés (dont les hashtags *#hurricaneharvey* et *#harvey*, ainsi que le nom de la ville de Houston), se détachent de tout autre contenu lexical par leurs occurrences qui se comptent en milliers alors que la majorité des mots n'apparaissent que quelques dizaines de fois tout au plus. En paramétrant l'axe des abscisses, on peut alors se focaliser exclusivement sur la catégorie des *scarce words* : la figure 5.14 représente ainsi les mots dont l'occurrence est comprise entre 180 et 200.

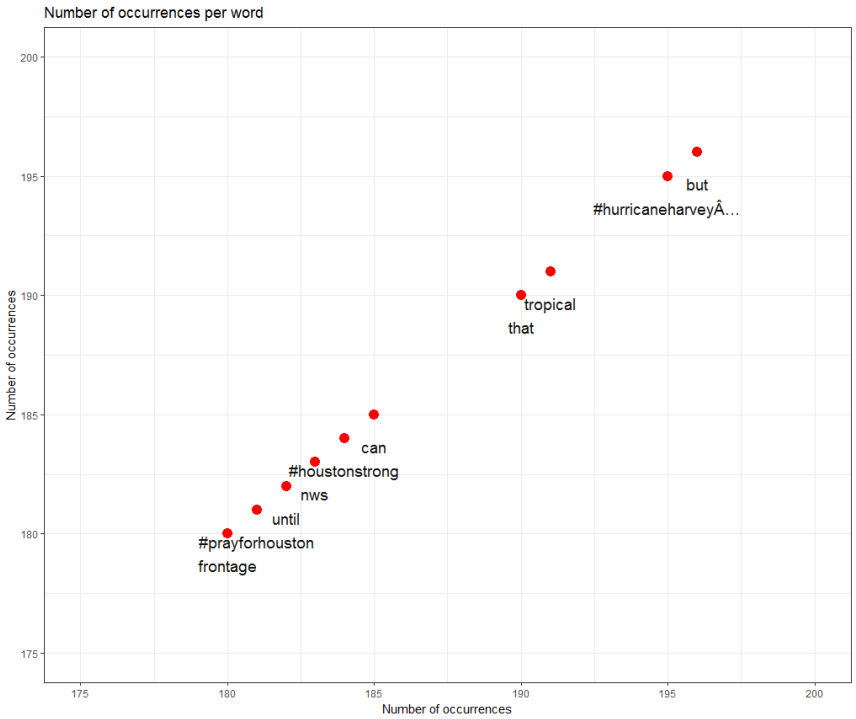


Figure 5.14 : Exemple d'exploration des mots épars – mots dont la fréquence est comprise entre 180 et 200

En procédant de cette manière, nous pouvons alors identifier un certain nombre de nouveaux mots-clés potentiellement utiles : *tropical, #houstonstrong, #prayforhouston, wind, frontage, blocked, help, evacuate, efforts, outbound, affected, closed, devasted, donate, pray*.

La deuxième solution alternative explorée consiste à afficher les collocations entre les mots sous forme réticulaire : nous avons eu recours au package *igraph* de R, qui permet de générer des graphes de réseaux facilitant la visualisation des collocations lexicales d'une part, et l'identification des *spams* d'autre part. Dans un premier temps, nous appliquons les étapes de nettoyage du corpus, décrites dans le chapitre précédent. La construction d'un graphique réticulaire nécessite ensuite la constitution d'une matrice en *n*-gramme (nous avons donc de nouveau recours au package *Rweka* afin de générer ici des matrices en bi-grammes). La figure 5.15 montre un premier résultat obtenu avec les deux hashtags-clés *#hurricane* et *#hurricaneharvey*, ainsi que le mot-clé *hurricane*.

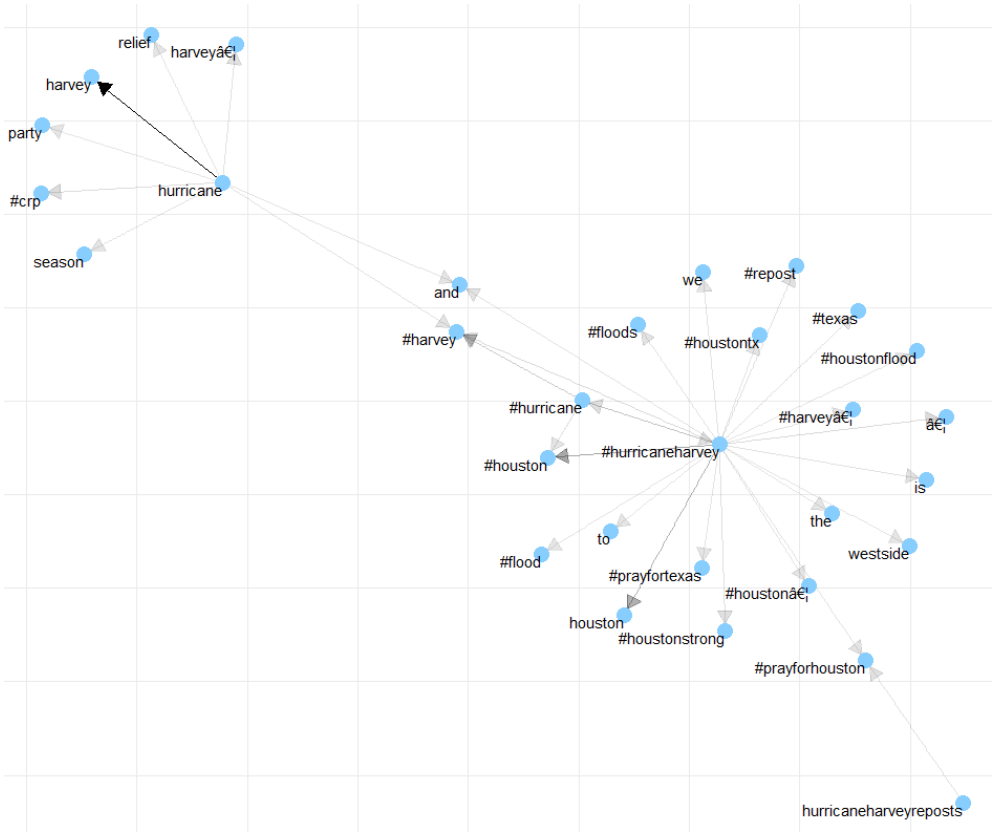


Figure 5.15 : Graphique des collocations lexicales à partir des mots-clés hurricane, #hurricane et #hurricaneharvey

Tout graphique de collocations est ainsi constitué de noeuds (chaque noeud représentant un mot-clé de la matrice en bi-grammes) et de flèches qui les relient. Ici, trois noeuds marquent les points de départ des flèches (il s'agit des points bleus associés au trois mots-clés sélectionnés) ; l'ensemble des autres points marquent les mots qui leur sont associés. Enfin, l'intensité de la couleur de la flèche indique la relation plus ou moins forte entre deux mots (dans la figure 5.15, l'association la plus fréquente est donc *hurricane* et *harvey*, suivie de *#hurricane* et *#harvey*, *#hurricaneharvey* et *#houston*, et enfin *#hurricaneharvey* et *houston*). Nous avons alors créé d'autres graphiques de ce type en sélectionnant des mots-clés déjà mis en évidence précédemment et en nous intéressant aux *scarce words* : la figure 5.16 ci-après affiche le graphe réseau construit à partir de la sélection de l'ensemble des mots-clés et hashtags contenant le radical *help :*

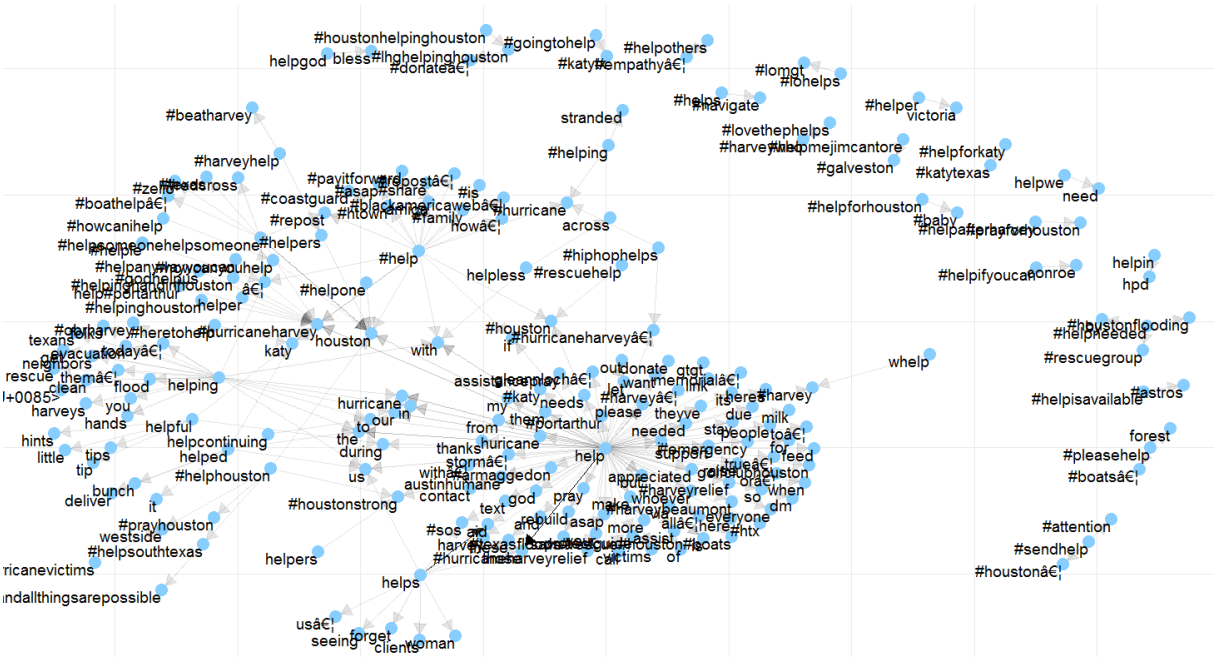


Figure 5.16 : Graphique des collocations lexicales à partir du radical "help"

On pourra ainsi mettre en évidence un certain nombre de mots-clés dérivés du radical : *helpers*, *helpless*, *helpful*, *helping*, mais également les nombreux hashtags qui leur sont associés (et que nous ajoutons à la liste de mots-clés recherchés pour le filtrage du jeu de tweets bruts) : *#beatharvey, #harveyhelp, #houstonhelping, #goingtohelp, #helpothers, #helpers, #helpforkaty, #helpforhouston, #helpifyoucan, #rescuegroup, #helpisavailable, #sendhelp, #pleasehelp, #rescuehelp, #houstonstrong, #helphouston, #helpsouthtexas, #prayhouston, victims, needed, emergency, support, #sos, #redcross*. L'expérience est répétée avec d'autres radicaux de mots et hashtags fréquemment employés, dont le radical *pray*, visible sur le graphique de la figure 5.14 (hashtag *#prayforhouston*), qui se décline également en de nombreux hashtags : *#stillpraying, #sendprayers, #praytogether, #prayersneeded, #prayfortexas, #prayforus, #hope, #sayamen, #stayblessed, #prayers, #texasstrong, #staystrong* ; d'autres sont probablement significatifs des lieux les plus affectés : *#prayforhouston, #prayforrockport, #prayforportarthur, #prayforgrandmission, #prayforbeaumont, #prayforwindsorvillage*. Enfin, les derniers mots ou hashtags associés au radical *pray* font référence à des actions ou conditions environnementales concrètres : *#evacuation, #currentlywatching, #thestormisover, #houstonflood, donate, flooding, raining, relief*.

Par ces graphiques, on peut également vérifier la pertinence d'un mot-clé et identifier l'éventuelle existence de messages *spams*. La figure 5.17 présente le graphe réseau construit à partir des mots dérivés du radical *home* :

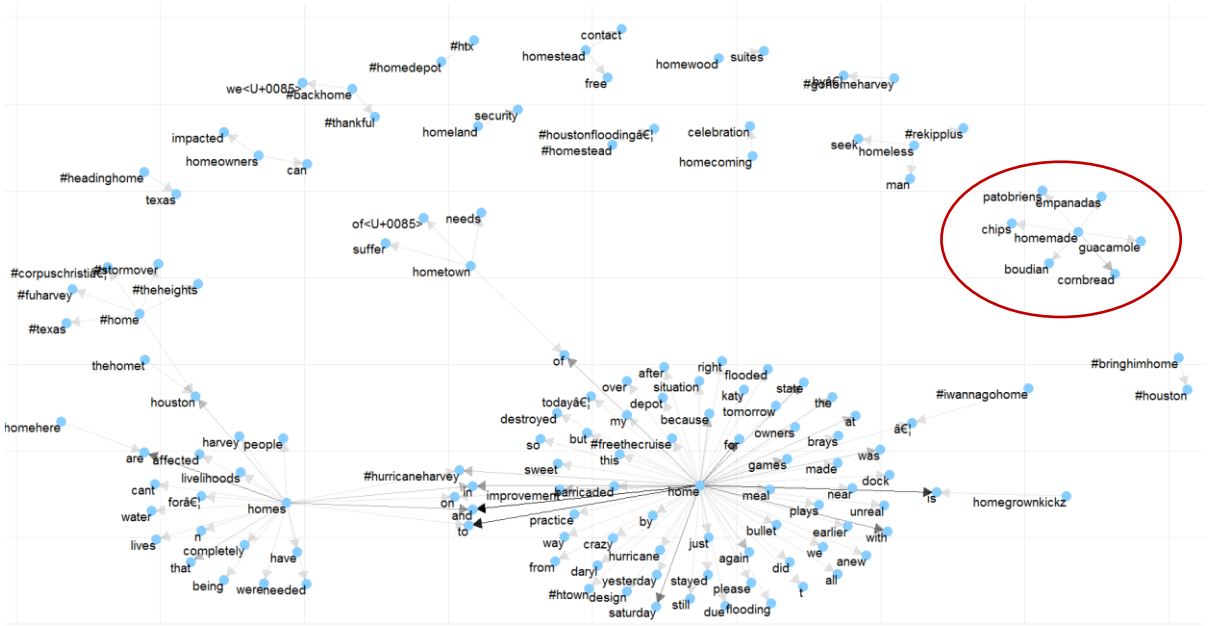


Figure 5.17 : Graphique des collocations lexicales à partir du radical "home"

Sur ce graphique, on pourra remarquer, dans un premier temps, l'emploi de termes marquant la mention du radical *home* dans le contexte de l'ouragan : *flooded, destroyed, barricaded, affected, impacted & homeowners* ; d'autres collocations se montrent plus ambiguës (dans la partie supérieure du graphique) : *a priori*, il serait risqué d'établir un lien entre des collocations comme *homecoming & celebration* ou encore *#headinghome & texas* et l'ouragan. En conséquence, la requête fondée à partir d'un mots-clé comme *home* doit rechercher les associations cohérentes mises en évidence (et non le mot-clé seul). Dans un second temps, on peut également détecter l'existence d'un *spam* fondé sur le mot *homemade* (entouré en rouge sur la figure 5.17), les messages proposant ici des idées de repas faits maison (*homemade guacamole*, *cornbread*, *chips*, *boudian*).

En procédant de cette manière (exploration des graphiques affichant la fréquence de chaque mot et des graphiques réticulaires basés sur l'exploration des collocations de mots-clés définis par l'utilisateur), nous avons mis en évidence, pour la deuxième extraction, un total de 68 nouveaux mots-clés, hashtags ou associations lexicales ayant permis de retourner un total de 46 334 tweets de crise géolocalisés[5]. Après soumission aux mêmes étapes d'exploration lexicale, nous pouvons encore identifier une poignée de mots-clés et associations lexicales : *urgent, bayou, disaster, national & guard, devasted, devastation, food & supplies, food & bank, food &stocked, food & necessities*. La dernière requête constitue ainsi un jeu de 47 297 tweets géolocalisés dans le polygone d'intérêt (soit 13% du jeu de tweets bruts initialement localisés dans le polygone).

[5] Et on pourra également noter que les mots comme *pray* ou *help* témoignent de déclinaisons et collocations bien plus nombreuses que les mots explorés à l'origine (*hurricane*, *harvey*) ; il en était de même pour le mot *flood*.

Enfin, si l'on tente de modéliser la relation entre le nombre de cibles lexicales recherchées et le nombre de tweets retournés par la requête (figure 5.18), on constate la répétition du comportement mis en évidence dans les exemples précédents : on peut ajuster une courbe de tendance linéaire ainsi qu'une courbe logarithmique (celle-ci demeurant la plus ajustée aux points).

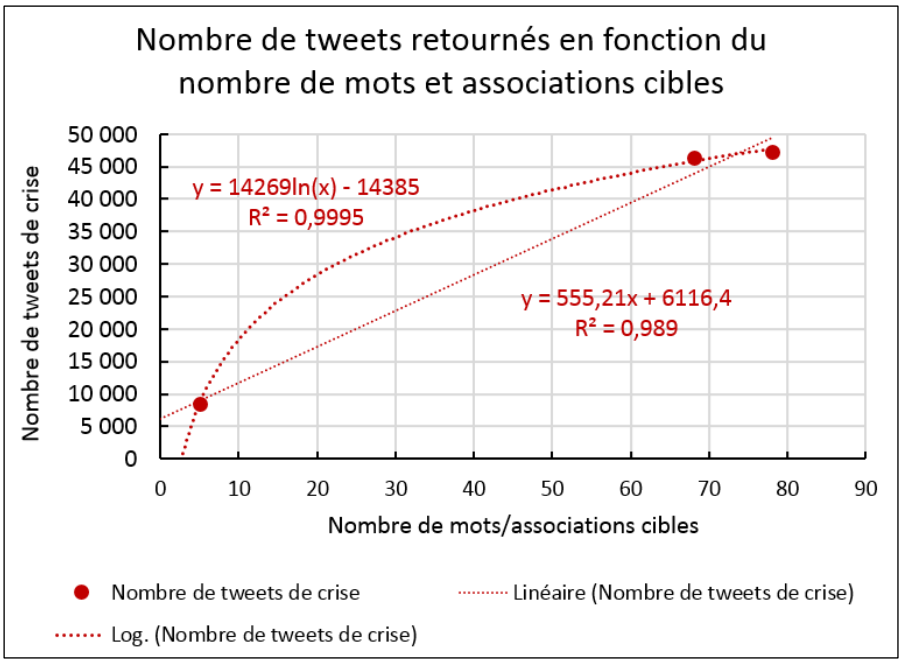


Figure 5.18 : Nombre total de tweets de crise retournés en fonction du nombre de mots et associations lexicales cibles

De la même manière, si l'on s'intéresse aux pratiques des utilisateurs, on constate le même comportement que nous avons mis en évidence dans le jeu de tweets de crise constitué pour l'exploration des phénomènes de pluies-inondations saisonniers du Texas (figure 5.19) : ici, le corpus de crise de l'ouragan Harvey enregistre 11 632 utilisateurs différents, soit presque deux fois plus que le corpus des phénomènes saisonniers, dont 65,8% n'ont envoyé qu'un seul tweet pendant la période de collecte. De nouveau, la majorité de l'activité tweeting de crise est liée aux émissions hors-normes d'une poignée de comptes (stations météorologiques, organismes officiels, *etc.*).

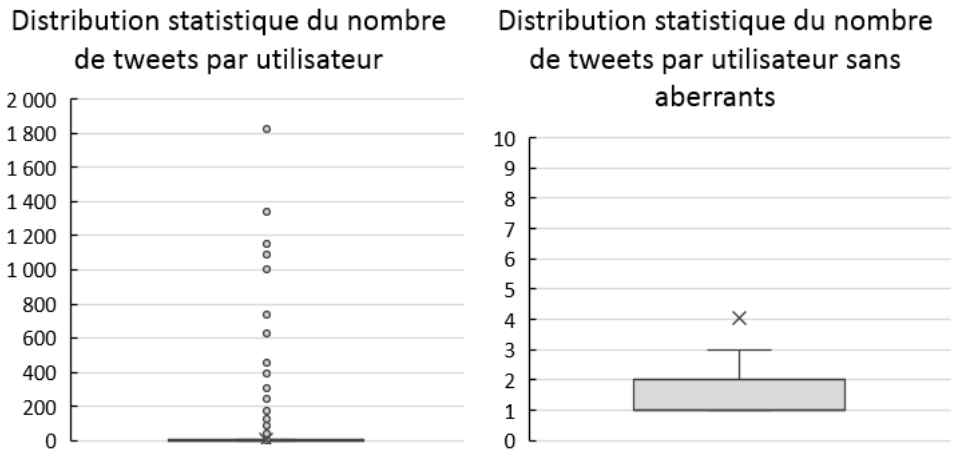


Figure 5.19 : Distribution statistique de la participation des utilisateurs dans le jeu de tweets de crise final de l'ouragan

Nous avons alors également isolé les tweets provenant des différents comptes actifs du *NWS* et des stations météorologiques et de jaugeage de l'*USGS* dans des jeux à part entière (ceux-inventorient un total de 7 408 tweets émanant de comptes associés à ces organismes officiels).

Pour terminer cette section relative aux retours de la démarche méthodologique d'extraction de tweets utiles, nous précisons que les fonctionnalités basiques de l'interface dont nous avons présenté la conception dans la section 4.2.3 du chapitre 4, ont été implémentées par Le Van Tuan[6] dans le cadre du projet *Sakura* du Laboratoire d'Informatique de Grenoble[7] : ce projet s'articule autour du développement d'une plateforme de capitalisation et de mutualisation des jeux de données spatio-temporelles ainsi que de leurs outils de traitement. Les travaux effectués pour le développement de l'interface destinée à automatiser la démarche d'extraction de tweets utiles proposée ici sont présentés dans l'annexe 2.

[6] dans le cadre d'un stage d'ingénieurie de Master 1.
[7] Référence du projet : https://github.com/sakura-team/sakura/wiki/Intro

Bilan des méthodes d'extraction des jeux de tweets de crise géolocalisés

A l'issue de ces différents tests, nous pouvons souligner l'existence de trois comportements répétés, quels que soient les territoires ou phénomènes considérés :

- à partir d'une poignée de mots-clés, on peut identifier de nouveaux mots cibles et constituer un jeu de tweets de crise plus exhaustif en adoptant alors une approche non supervisée. En outre, la première étape d'analyse lexicale, qui suit directement l'extraction du premier jeu de tweets de crise, s'avère la plus prolifique en termes d'identification d'un nouveau vocabulaire cible et de potentiel d'enrichissement du corpus de crise.

- Lorsqu'on tente de modéliser le nombre de tweets de crise retournés en fonction du nombre de mots cibles, nous identifions deux tendances qui s'ajustent aux points : une relation logarithmique qui indique une saturation du nombre de tweets à partir d'une certaine quantité de mots-clés recherchés, ainsi qu'une relation linéaire. Au regard du comportement des mots dans les tweets (une poignée de mots sont massivement adoptés alors que la majeure partie du vocabulaire de crise témoigne de quelques dizaines d'occurrences), nous pourrions valider la relation logarithmique : et en effet, c'est la requête prenant en compte ces mots massivement adoptés qui retourne le plus de tweets. Après cette étape, il faut s'intéresser aux mots et associations lexicales éparses pour identifier de nouveaux tweets.

- Les utilisateurs détectés dans un corpus de crise témoignent d'un comportement analogue quel que soit le type de phénomène considéré : 75% des utilisateurs n'enregistrent qu'une participation unique voire double dans l'événement virtuel. Les utilisateurs correspondant aux profils hors-normes, qui cumulent plusieurs millers de tweets à leur compte, correspondent en fait à des acteurs officiels. Nous avons cependant conservé et considéré comme traces de réalité terrain les tweets émis par les comptes associés au *NWS* ainsi qu'aux stations de l'*USGS*.

5.2. Le tweet de crise géolocalisé : exploration de l'événements virtuel comme marqueur des lieux et des temporalités du réseau dans les territoires en crise

Ce second axe est focalisé sur l'analyse et la pertinence du tweet comme marqueur de localisation de phénomènes naturels dans le temps et dans l'espace ainsi que des dynamiques d'émission. En effet, depuis les publications des premières études sur le matériau tweet géolocalisé, on sait que les dynamiques du réseau constituent l'écho des dynamiques événementielles survenues dans le monde réel, qu'elles soient d'origine sociale ou naturelle (De Longueville *et al*., 2008 ; Lee Hughes et Palen, 2009 ; Lin *et al*., 2013). Ici, nous portons davantage notre attention sur les profils des espaces de réactivité et, afin de ne pas restreindre le tweet géolocalisé à un simple point sur la carte, à l'étude de sa composante sémantique. On observe dans un premier temps l'échelle globale du phénomène physique pour appréhender la variabilité spatiale de la réponse numérique enregistrée sur le réseau virtuel. L'objectif général consiste ainsi à évaluer l'ensemble des composantes du tweet comme témoin spatio-temporel et sémantique des dynamiques d'un phénomène hydrométéorologique récurrent ou peu fréquent.

5.2.1. Spatialisation et temporalités d'un événement virtuel en réponse à des phénomènes physiques récurrents

5.2.1.1. Critères d'identification d'un phénomène physique réel par les événements virtuels

Temporalités de l'événement virtuel.

Comme l'état de l'art des apports des tweets géolocalisés à la question des risques et catastrophes naturels l'a montré, on peut rapidement identifier les temporalités de phénomènes physiques réels par les *pics* d'émission de tweets relatifs au thème d'étude. Les graphiques ci-après (figure 5.20) indiquent que notre terrain d'étude n'échappe pas à cette norme : ils représentent, pour chaque journée comprise entre les mois de mars et juin 2016, l'ensemble des tweets géolocalisés contenus dans le corpus de crise extrait sur l'Etat du Texas. En outre, ils mettent en évidence les différents acteurs du réseau identifiés dans ce corpus : en conséquence, les courbes distinguent les tweets associés aux acteurs officiels (*National Weather Service, NWS*) et émis par des comptes automates liés au réseau de stations météorologiques déployées par le *United States Geological Survey* (*USGS)*, fonctionnel à partir d'avril 2016.

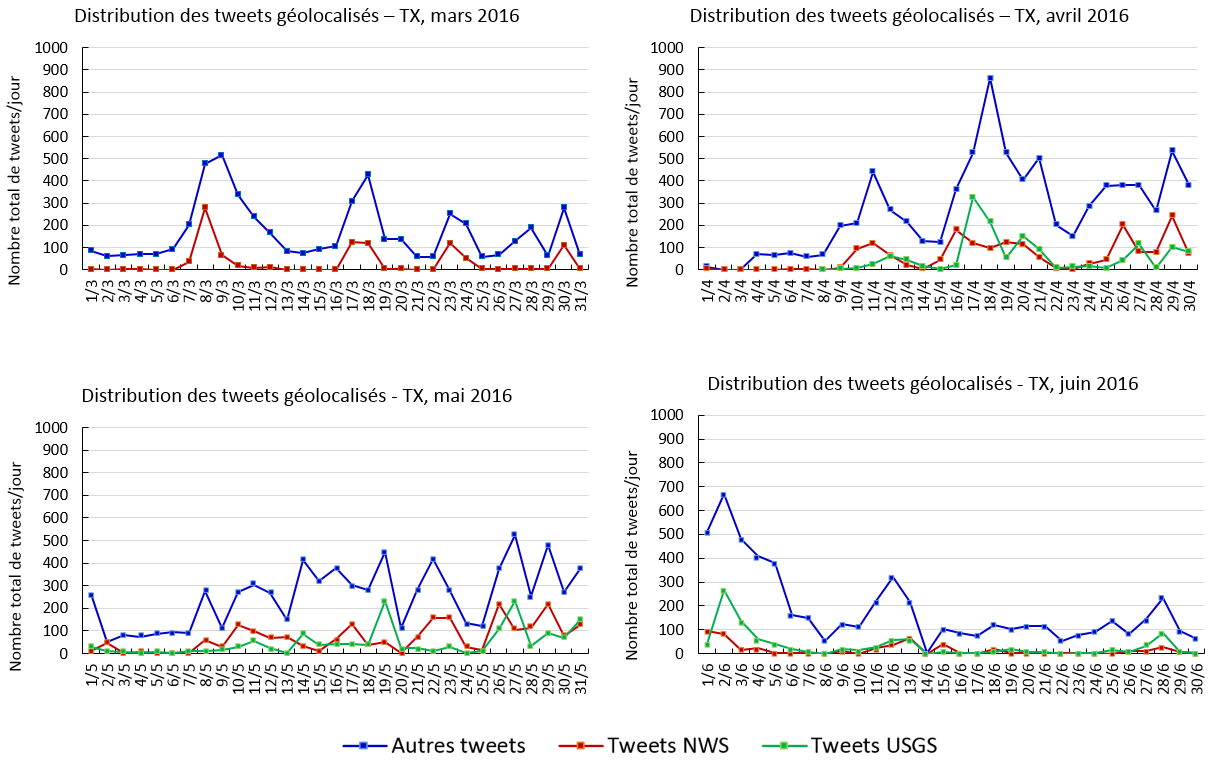


Figure 5.20 : Distribution temporelle quotidienne des tweets de crise géolocalisés émis au Texas entre mars et juin 2016

Le flux de tweets de crise géolocalisés non rattachés à l'activité du *NWS* ou des stations de l'*USGS* s'avère quasi continu sur l'ensemble de la période collectée, et ponctué de pics d'émission témoignant d'événements virtuels récurrents mettant en exergue les périodes suivantes : 8-9 mars, 17-18 mars, 23-24 mars, 30 mars ; 11 avril, 16-21 avril ; 25-30 avril ; 8 mai, 10-12 mai, 14-19 mai, 21-23 mai ; 26 mai - 6 juin, 12 et 28 juin 2016. Ces périodes correspondent par ailleurs aux pics d'activité enregistrés depuis les comptes des acteurs officiels. En revanche, quelques périodes enregistrent un événement virtuel sans pour autant qu'on observe un pic de tweets officiels, marqueurs de la survenue d'un phénomène physique : 28 mars, 1er mai : s'agit-il alors de comptes automatiques ponctuels, d'épiphénomènes, d'utilisateurs qui s'agitent par îlots (mais sans pour autant qu'un phénomène physique ne vienne perturber l'environnement) ou d'une simple sensibilité à la moindre pluie ?

Nous avons ici émis des hypothèses de manière relative quant à la survenue de perturbations physiques générant un événement virtuel, en observant exclusivement les pics d'émission d'acteurs divers : à partir de ces graphiques, notre attention se porte sur l'événement virtuel enregistré mi-avril 2016 (celui-ci ayant drainé les plus fortes quantités de tweets de crise) pour la suite des analyses. Néanmoins, ces pics s'avérant d'intensité variable, n'étant pas toujours concomitants en fonction des acteurs et s'inscrivant dans des durées différentes, ces graphiques soulèvent l'interrogation suivante : un état de veille et de

sensibilité aux phénomènes hydrométéorologiques semble quasi permanent mais à partir de quel seuil d'émission de tweets peut-on affirmer l'existence d'un événement virtuel ? Par exemple, si l'on cherche à délimiter le début et la fin de l'événement virtuel qui répond au phénomène physique survenu mi-avril par des paramètres statistiques (figure 5.21), il se dégage alors une tendance caractéristique des usages des réseaux sociaux numériques : l'influence des valeurs extrêmes sur la distribution générale a pour effet une moyenne supérieure à la médiane (voire parfois même au troisième quartile), quelle que soit la série considérée.

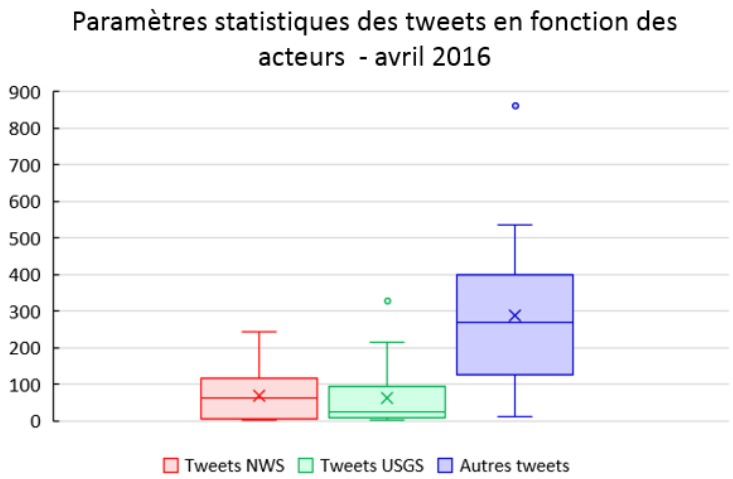


Figure 5.21 : Boîtes à moustaches - Paramètres statistiques quotidiens des tweets de crise en fonction des acteurs – avril 2016

En conséquence, nous sélectionnons la médiane comme paramètre indicateur de début et de fin d'événement virtuel pour cette perturbation de mi-avril : le tableau 5.1 ci-dessous récapitule les dates, en fonction des acteurs identifiés, pour lesquelles les émissions de tweets de crise dépassent les médianes respectives enregistrées pour le mois d'avril seul. Nous retenons ainsi un événement virtuel survenu du 16 au 21 avril 2016.

Tableau 5.1 : Périodes d'enregistrement d'un phénomène physique sur le réseau virtuel, par acteur

Acteur	Médiane avril 2016	Emissions de tweets > Q2
USGS	25	17/04 – 21/04
NWS	61	16/04 – 20/04
Autres	267,5	16/04 – 21/04

Spatialisation de l'événement virtuel.

Dans un second temps, la cartographie des précipitations et des tweets de crise géolocalisés émis en réponse à la survenue du phénomène physique permet de visualiser l'existence *supposée* d'une relation entre la localisation du phénomène et des foyers d'émission de tweets. Pour l'exemple de l'événement virtuel retenu, les résultats se démarquent des constats effectués dans l'état de l'art. Si la réactivité locale à la survenue d'un phénomène physique est perceptible (notamment sur les cartes représentant les tweets émis les 16 et 17 avril 2016), l'affirmation du tweet géolocalisé comme témoin systématique de survenue d'un phénomène physique réel s'avère discutable (figure 5.22).

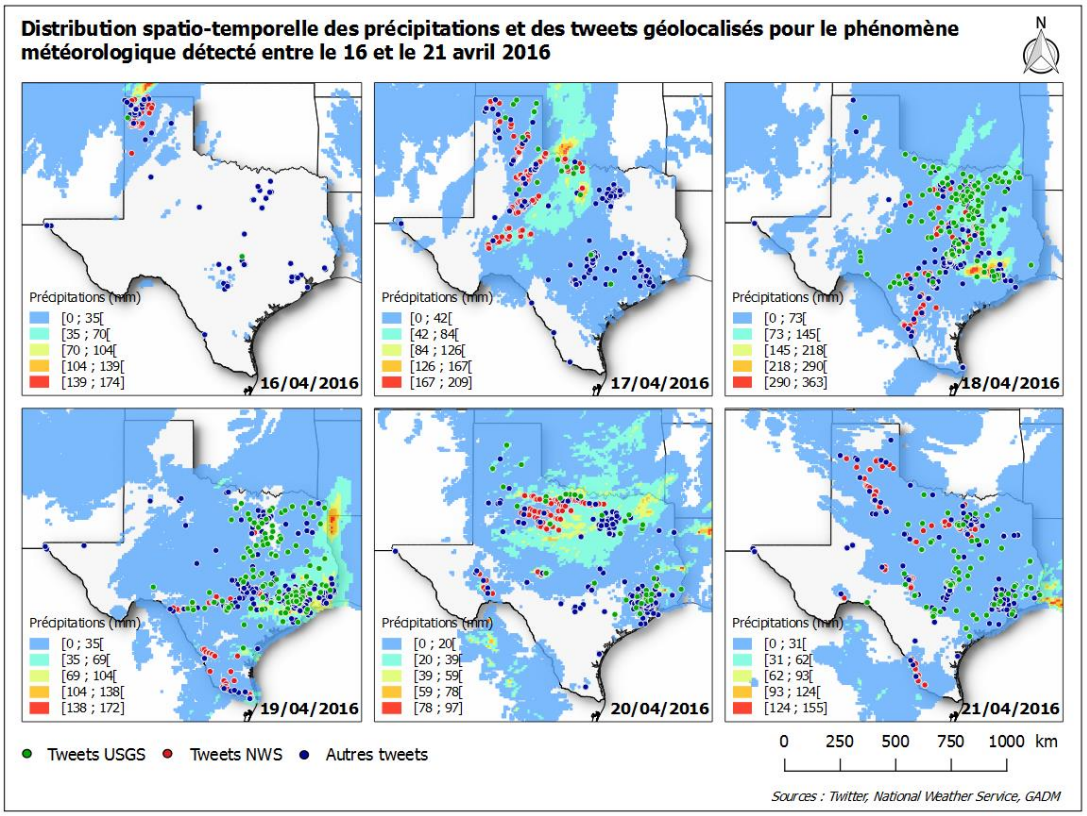


Figure 5.22 : Dynamique spatio-temporelle des précipitations et des émissions de tweets géolocalisés par acteur du 16/04/2016 au 21/04/2016 – Texas (C.Cavalière)

A l'échelle de l'Etat, la carte de la figure 5.22 permet de distinguer trois types de logique de spatialisation des tweets de l'événement virtuel :

- une logique de suivi des dynamiques spatio-temporelles du phénomène physique : le 16 avril, l'activité est concentrée, quels que soient les acteurs émetteurs, dans le nord de l'Etat alors frappé par la perturbation ; cette activité cesse totalement le 18 avril alors que la perturbation s'est déplacée sur l'est de l'Etat. Cette logique de suivi se trouve également perceptible les 17 et 18 avril : les tweets du *NWS* forment des structures linéaires qui s'inscrivent dans la logique d'intensité de la perturbation. Pour autant, le 17 avril, un foyer de

tweets *USGS* est enregistré au niveau de la métropole de Dallas alors que les données du *NWS* n'enregistrent pas de précipitations.

- Les tweets de crise sont également émis de manière sporadique dans les espaces affectés ou en marge des précipitations : une poignée de tweets est régulièrement émise depuis la pointe occidentale de l'Etat, à la frontière avec le Nouveau-Mexique et le Mexique (ce qui correspond à la ville d'El Paso) ;

- les émissions de tweets de crise sont concentrées en amas indiquant une poche d'activité ; ces foyers s'identifient aux grandes aires métropolitaines de l'Etat (Dallas, Houston) et, contrairement à l'activité de la partie septentrionale de l'Etat, ceux-ci semblent être actifs de manière continue, qu'ils soient concernés ou non par la perturbation et quelle que soit son intensité.

Par ailleurs, la carte met en évidence un fait non négligeable dans la problématique des risques naturels : en moyenne, sur les six journées collectées, 96% des mailles représentant les cumuls de précipitations quotidiennes sont vides de tweets de crise géolocalisés. L'activité *tweeting* enregistrée correspond donc à l'agitation d'une minorité de foyers. Or, si l'ouest du Texas n'est que très peu peuplé, certains territoires affectés à l'est entre le 17 et le 20 avril concentrent des espaces urbains d'un niveau hiérarchique moindre par rapport à Dallas ou Houston mais peu, voire pas visibles sur le réseau, notamment dans le nord et l'est de l'Etat.

Au final, l'identification d'un lieu de crise par les lieux d'émission virtuels se révèle ambiguë : si les lieux d'émission virtuels peuvent constituer les miroirs des lieux en crise (ce constat est nettement perceptible en début de perturbation lorsque le nord de l'Etat est d'abord frappé, puis lorsque la perturbation atteint Dallas et Houston les 18 et 19 avril), ils semblent également mettre en évidence les comportements particuliers de certains espaces qui constituent des foyers d'émissions ponctuelles ou conséquentes, mais continues et persistantes, quelles que soient les conditions environnementales du moment : le 16 avril, une activité sporadique est d'ores-et-déjà enregistrée sur Dallas et Houston alors que seul le nord du Texas est frappé. Les chiffres du tableau 5.2 ci-après présentent alors pour chaque journée, le pourcentage de tweets géolocalisés émis en dehors des mailles correspondant aux précipitations. Ils semblent confirmer le comportement particulier des espaces urbains actifs en permanence : dès que la perturbation affecte ces espaces, le pourcentage de tweets situés en dehors des mailles affectées par la pluie tend à diminuer.

Tableau 5.2 : Pourcentages quotidiens de tweets de crise géolocalisés situés en dehors des mailles de précipitations

Date	% Tweets en dehors d'une maille
16/04/2016	64,34
17/04/2016	15,87
18/04/2016	0,28
19/04/2016	9,4
20/04/2016	16,65
21/04/2016	2,1

En fait, si l'on agrège l'ensemble des tweets de crise géolocalisés émis pendant cette perturbation dans des mailles de dix kilomètres de côté et qu'on étudie leurs paramètres statistiques, on se retrouve confronté à la tendance soulignée précédemment (figure 5.23 ci-dessous) : 75% des mailles de tweets de crise géolocalisés sont concentrées dans les valeurs faibles (elles regroupent tout au plus trois tweets) ; la moyenne est ici supérieure au troisième quartile et une minorité de mailles se dégagent de la série par leurs valeurs extrêmes (cf. boîte à moustache de droite).

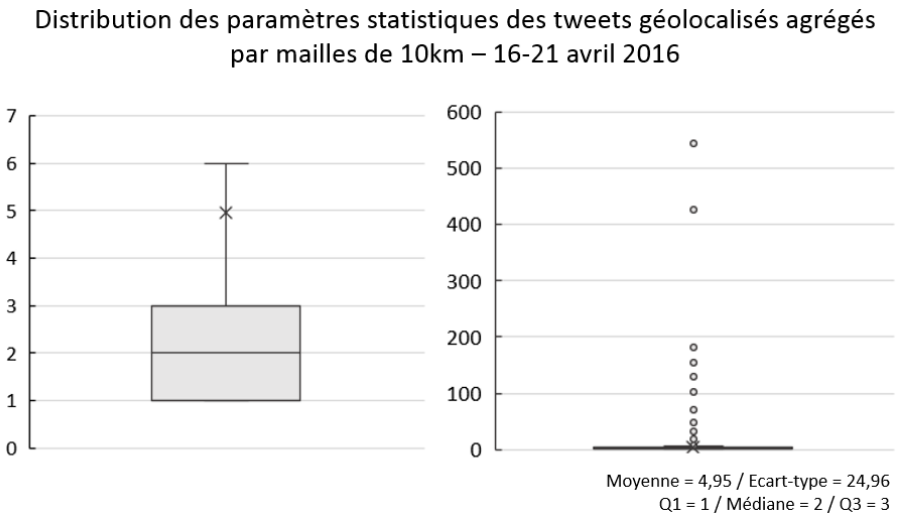


Figure 5.23 : Boîtes à moustaches - Distribution quotidienne des tweets de crise agrégés par 10km, tous acteurs confondus, du 16 au 21 avril 2016 - Texas

5.2.1.2. Le tweet de crise géolocalisé, marqueur des espaces urbains ?

Face aux résultats constatés et dans la problématique du risque de hiérarchisation des territoires par le numérique, nous souhaitons explorer le poids éventuel d'un premier facteur susceptible d'influencer la localisation des tweets de crise : la répartition de la population en fonction de son profil urbain ou rural. Si l'on sait déjà que la carte des tweets est le miroir de la carte des densités de population, la distribution spatiale des tweets géolocalisés en période de crise est-elle liée aux comportements numériques propres aux individus connectés en milieu urbain ?

Dans un premier temps, en ayant recours aux données de l'*ACS* issues du recensement de 2010[8], nous classons les profils des comtés du Texas en fonction du nombre d'habitants inventoriés comme *ruraux.* Les comtés dont moins de 25% de la population est rurale sont alors considérés comme urbains. Les comtés dont le pourcentage de population rurale est compris entre 25% et 50% sont considérés comme des territoires périurbains. Les comtés dont le pourcentage de population rurale est compris entre 50% et 80% sont considérés comme des territoires ruraux ; ces deux catégories correspondent à des territoires situés en périphérie des grandes aires métropolitaines du Texas ou à des territoires ruraux avec un pôle urbain ou un réseau de pôles urbains locaux. Les comtés dont plus de 80% d'habitants sont ruraux sont considérés comme des territoires hyper-ruraux, à l'habitat ponctuel et dispersé, sans pôle urbain. La carte de la figure 5.24 affiche le résultat de la typologie effectuée en fonction du nombre d'habitants *ruraux*.

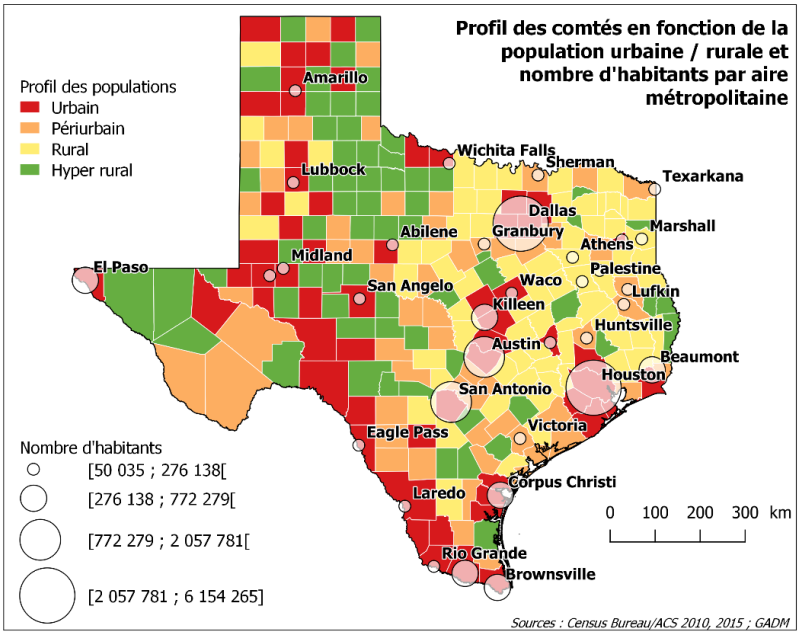
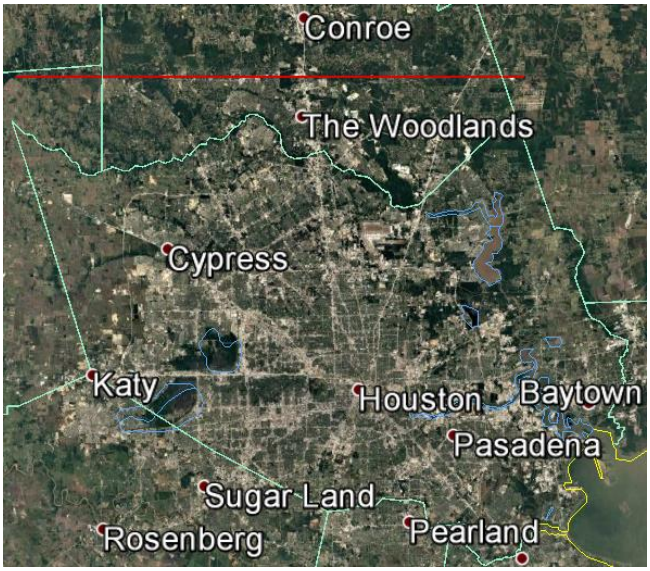


Figure 5.24 : Typologie des comtés en fonction du profil de la population

[8] Les séries de données classant les individus recensés comme urbains ou ruraux ne font pas l'objet d'estimations annuelles. Elles sont donc collectées lors des recensements décennaux ; en conséquence, les chiffres les plus récents datent de 2010 et ne seront pas mis à jour avant le prochain recensement qui aura lieu en 2020.

Notons cependant qu'un habitant d'un comté périphérique aux grandes aires métropolitaines n'est pas forcément considéré comme *rural* : en effet, dans certains comtés essentiellement agricoles ou désertiques, les populations peuvent se concentrer dans un pôle urbain local. Les habitants de tels pôles sont alors considérés comme des *urbains* ; c'est ce que montre la figure 5.25 : les populations des comtés de Harris et de Reeves sont toutes deux classées comme urbaines dans la mesure où elles se concentrent dans une aire métropolitaine (Houston pour le comté de Harris) et dans un pôle urbain local (Pecos pour le comté de Reeves).



Des comtés aux populations urbaines – Harris (aire métropolitaine de Houston) et Reeves (pôle urbain de Pecos)

100 km

Figure 5.25 : Deux profils de comtés aux populations urbaines : Harris et Reeves (Google Earth, 2016)

En conséquence, l'analyse ne va pas se focaliser sur le nombre d'habitants par comté mais sur leur profil, afin de déterminer si l'habitant du milieu urbain est plus enclin à tweeter dans une situation de crise que l'habitant du milieu rural.

Dans un second temps, nous analysons les paramètres statistiques liés aux émissions quotidiennes de tweets de crise collectés pendant l'ensemble de la période étudiée, et classés en fonction des profils de la population des comtés (figure 5.26). Quel que soit le type d'espace considéré, la structure mise en évidence au paragraphe 5.2.1.1 se répète : 75% des comtés, quel que soit leur profil, affichent une activité réseau ponctuelle alors qu'une minorité d'individus enregistrent l'activité la plus forte :

- le point fléché en rouge sur le graphique 1 correspond au comté urbain de Harris, sur lequel s'étend la ville de Houston (cf. carte de la figure 5.27) : le nombre de tweets de crise

géolocalisés émis sur la période enregistrée (4 332) est 110 fois supérieur à la médiane des comtés urbains (Q2 = 39,5) ;

- le point fléché en orange sur le graphique 3 correspond au comté périurbain de Ellis, au sud de l'aire métropolitaine de Dallas (cf. carte de la figure 5.27) : le nombre de tweets émis sur la période enregistrée (2 551) est 132 fois supérieur à la médiane des comtés périurbains (Q2 = 17) ;

- le point fléché en jaune sur le graphique 5 correspond au comté rural de Grimes, au nord-ouest du comté de Harris (cf. carte de la figure 5.27) : le nombre de tweets émis sur la période enregistrée (1 724) est 132 fois supérieur à la médiane des comtés ruraux (Q2 = 13). Après exploration du type d'activité enregistrée dans ce comté, il s'avère que 99,5% des tweets de crise émis sont associés à un compte automatique émettant des bulletins météorologiques en un unique point du territoire. La poignée de tweets restants (7 tweets) correspond à une activité sporadique localisée dans le sud du comté, émanant également d'un unique compte annonçant la survenue de crues éclair en des lieux précis.

- Le point fléché en vert sur le graphique 7 correspond au comté de Donley, à l'est de la ville d'Amarillo (aucune aire urbaine ou métropolitaine ne s'étend sur ce comté, cf. carte et image satellite de la figure 5.27). Le nombre de tweets émis sur la période enregistrée (95) est 10 fois supérieur à la médiane des comtés hyper-ruraux (Q2 = 9). L'activité tweeting est ici liée aux émissions de cinq comptes. En revanche, les quelques pôles d'habitations de ce comté ne regroupent ici que 10% des tweets émis : en fait, l'activité majeure se concentre à proximité des voies de communication.

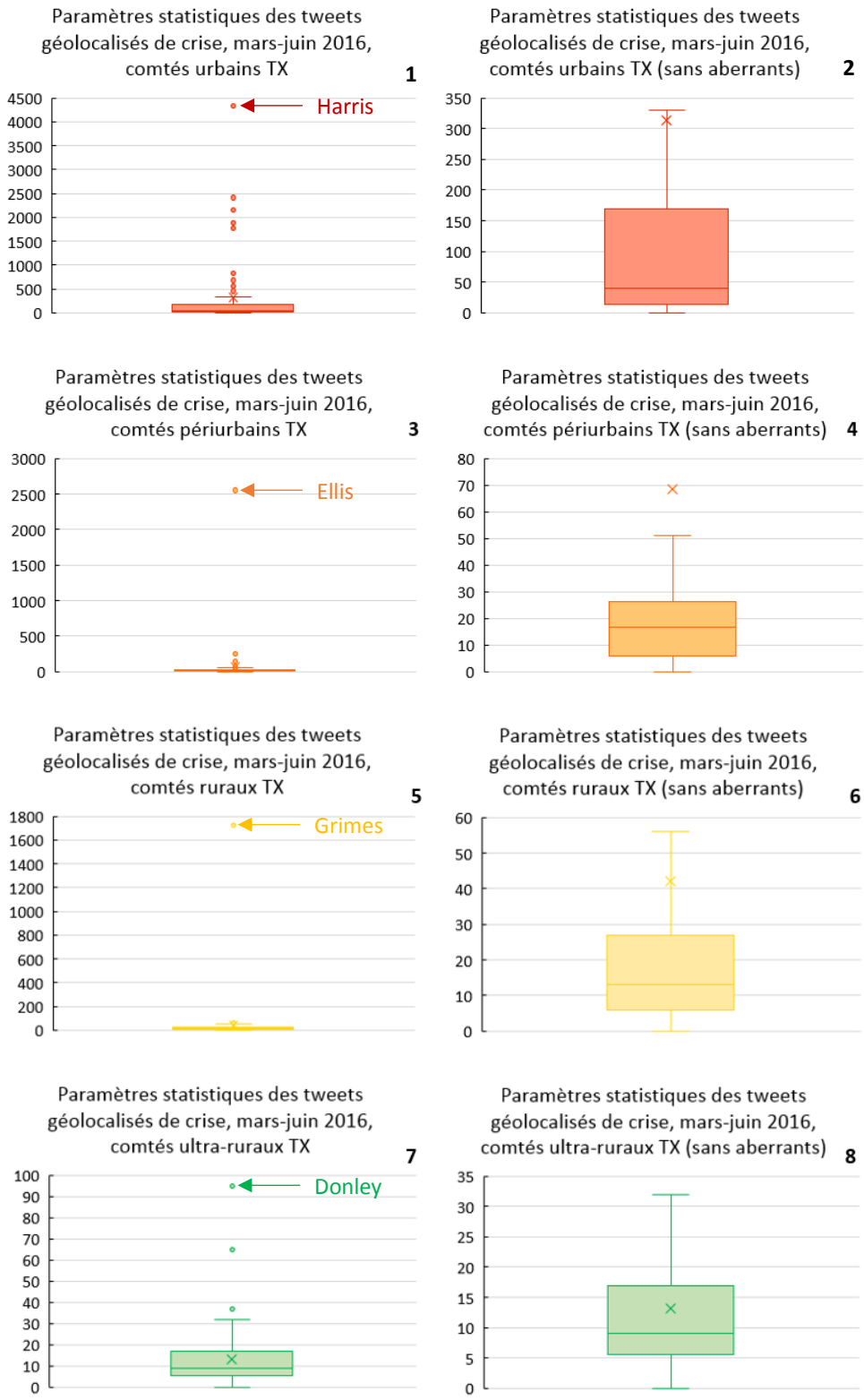


Figure 5.26 : Distribution statistique des tweets de crise géolocalisés en fonction du profil des populations dans les comtés du Texas

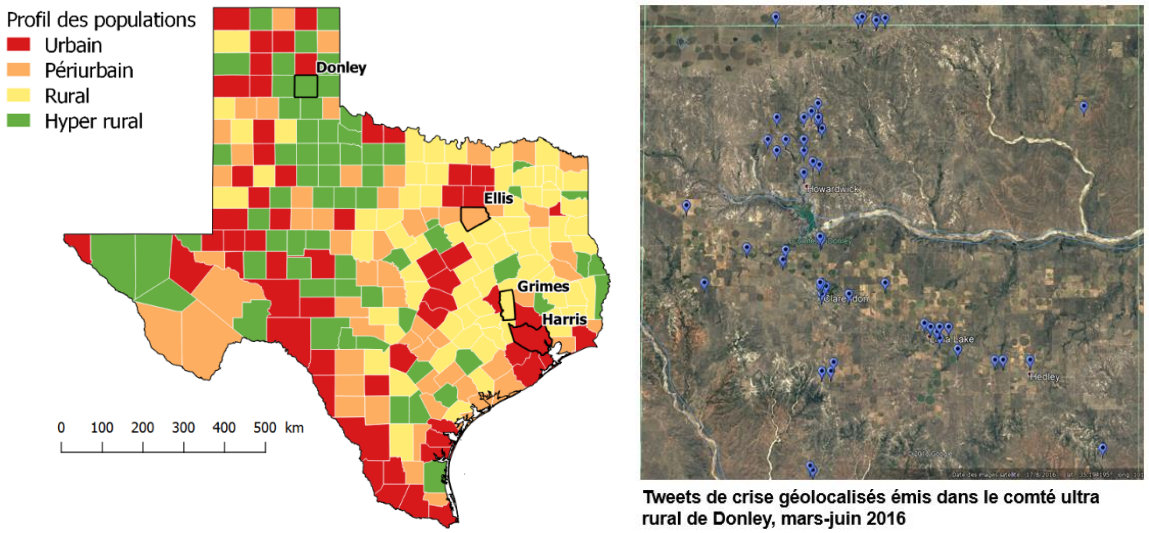


Figure 5.27 : Profil des comtés en fonction de leur population et distribution spatiale des tweets de crise géolocalisés dans le comté hyper-rural de Donley (Google Earth pour l'image satellite)

A petite échelle, tout type d'espace concentrant une importante activité de tweets de crise géolocalisés correspond alors, conformément aux pratiques statistiques traditionnelles en géographie, à un individu aberrant. A partir d'un tableau de contingence (type de comté / nombre de tweets répartis en cinq classes selon la méthode des quantiles), nous avons alors effectué un test de Khi² pour vérifier l'éventuelle existence d'une relation entre le profil des comtés et le nombre de tweets de crise émis sur l'ensemble de la période collectée : il tend à valider l'hypothèse d'une relation entre les deux variables (Khi² observé = 62,76 et Khi² théorique de 26,22 pour 12 degrés de liberté et un risque d'erreur de 1%) mais de faible intensité (V de Cramer = 0,14)[9]. Le graphique des écarts entre valeurs observées et valeurs du modèle d'indépendance (figure 5.28) souligne une nouvelle fois les tendances opposées entre activité *tweeting* forte/faible et les populations urbaines/rurales :

- parmi les populations urbaines, les plus fortes quantités de tweets sont sur-représentées (phénomène certainement imputé à l'activité des habitants des grandes métropoles) ; à l'inverse, les faibles quantités de tweets sont sous-représentées ;
- les milieux où la population est rurale présentent la situation opposée : attirance pour les faibles quantités de tweets, et répulsion pour les quantités les plus élevées.

[9] Des résultats analogues sont observés si l'on retire les individus aberrants de chaque type de comté : Khi² observé de 50,51 (sa valeur théorique est identique à celle annoncée dans ce paragraphe) et V de Cramer de 0,13.

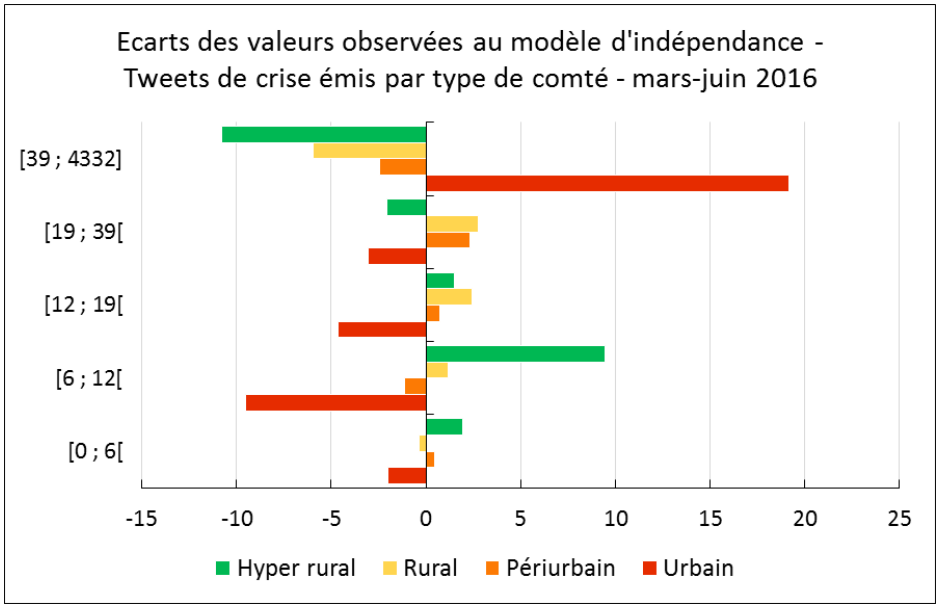


Figure 5.28 : Ecarts entre les effectifs de tweets observés et les effectifs en situation d'indépendance statistique

Les territoires de prédilection de l'activité réseau restent donc les territoires concentrant des populations urbaines ; l'activité réseau enregistrée dans les territoires périurbains et ruraux reste marginale : 75% des 136 comtés dont les populations sont classées comme rurales et ultra-rurales contiennent moins de 21 tweets de crise géolocalisés (troisième quartile des deux séries confondues). Pour autant, si les populations urbaines semblent plus enclines à tweeter en période de crise, il reste encore à nuancer ce résultat : seuls 50% des comtés aux populations urbaines contiennent plus de 39 tweets géolocalisés (médiane de la série des comtés aux populations urbaines, pour un maximum de 4 432 tweets de crise dans le comté de Harris). En outre, parmi cette catégorie de comtés regroupant les plus fortes quantités de tweets, 93% de ces tweets sont situés dans des aires métropolitaines. En conséquence, même si on considère l'existence de populations urbaines dans des milieux ruraux, c'est la population urbaine des métropoles qui est à l'origine des foyers d'activité réseau les plus conséquents. Il existe donc un risque de disposer de trop peu de tweets voire de manquer un phénomène grave qui frapperait un territoire non métropolitain.

5.2.1.3. Exploration des lieux de réactivité de l'événement virtuel du 16 au 21 avril 2016

Cette section explore le contenu des lieux de réactivité détectés pendant la perturbation enregistrée du 16 au 21 avril 2016, en fonction des différentes phases de la crise mais également en fonction des problématiques identifiées dans les paragraphes précédents. En confrontant données officielles, contenu textuel et localisation des tweets, nous cherchons à savoir si l'événement virtuel permet non seulement de détecter le type d'un phénomène réel mais encore à mettre en évidence l'éventuelle absence de phénomènes d'intensité

exceptionnelle sur le réseau. Chaque territoire/journée exploré est illustré comme suit : une carte affiche les tweets de crise géolocalisés étiquetés en fonction d'un thème principal identifié (consignes de sécurité, comportements individuels ou collectifs, conditions environnementales, *etc.*) et le nuage de mots (mots simples ou associations lexicales en fonction du nombre de tweets dans l'entité géographique considérée) représente le contenu lexical des tweets.

Exploration d'une poche d'activité métropolitaine permanente - aire métropolitaine de Houston.

L'activité virtuelle de crise enregistrée sur le réseau le 16 avril 2016 (figure 5.29) s'avère faible (onze tweets de crise géolocalisés) ; parmi eux, on peut en distinguer six faisant explicitement référence à l'anticipation d'un phénomène hydrométéorologique prévu. Ces tweets mentionnent des événements collectifs annulés ("*well sadly the second day of the bp ms 150*[10] *has been cancelled due to inevitable thunderstorms*") ainsi que l'adaptation d'un comportement individuel ("*last walk outside before a weekend of rain @ hermann park jogging trail*").

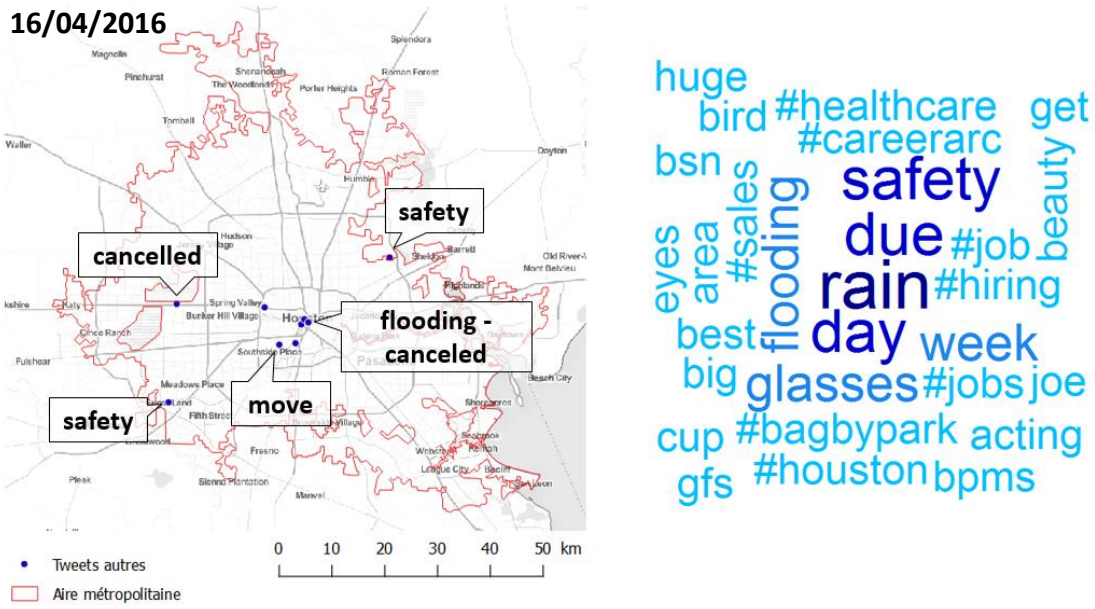


Figure 5.29 : Activité virtuelle de crise à Houston - 16/04/2016 (C.Cavalière)

En début de phénomène physique, l'événement virtuel reste ponctuel (24 tweets géolocalisés émis le 17 avril) mais sa qualité sémantique s'avère pauvre (figure 5.30) : seuls neufs tweets témoignent d'un phénomène en cours et d'intensité variable (de la faible pluie à l'orage). Les autres tweets contiennent le vocabulaire de crise mis en évidence mais dans des contextes dont le rapport avec le phénomène réel en cours n'est pas apparent (on peut

[10] *BP MS 150* fait référence à une course cycliste annuelle qui rallie Houston à Austin.

alors les considérer comme du bruit) : "*Even though y'all messed up my hat; I sat through a tornado*" ; "*Natural light floods this modernist home's living space. A minimalist haven on boulevard Oak's*".

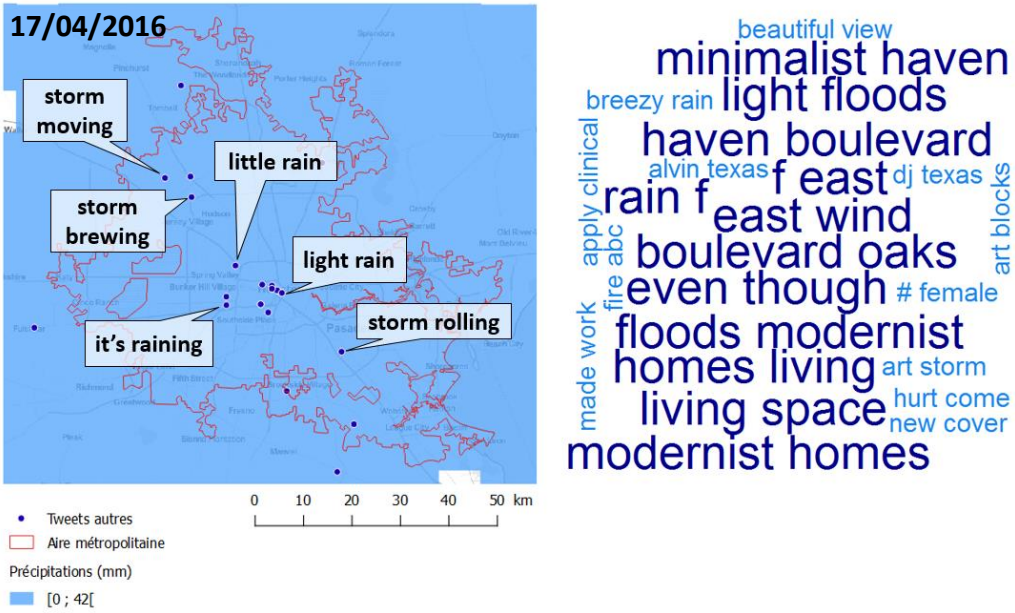


Figure 5.30 : Activité virtuelle de crise à Houston - 17/04/2016 (C.Cavalière)

Le 18 avril marque l'arrivée du phénomène physique intense par l'ouest de la métropole, qui se poursuit le lendemain. Notons d'emblée que l'agitation maximale enregistrée sur le réseau ne coïncide pas avec le début de la phase violente du phénomène réel mais apparaît dans les 24 heures suivantes : le 18 avril enregistre 53 tweets géolocalisés dont 6 tweets d'alerte alors que le 19 avril cumule un total de 418 tweets géolocalisés et 50 tweets d'alerte sur la métropole. L'analyse du contenu sémantique marque par ailleurs une évolution des thématiques mises en évidence dans les tweets (figure 5.31) :

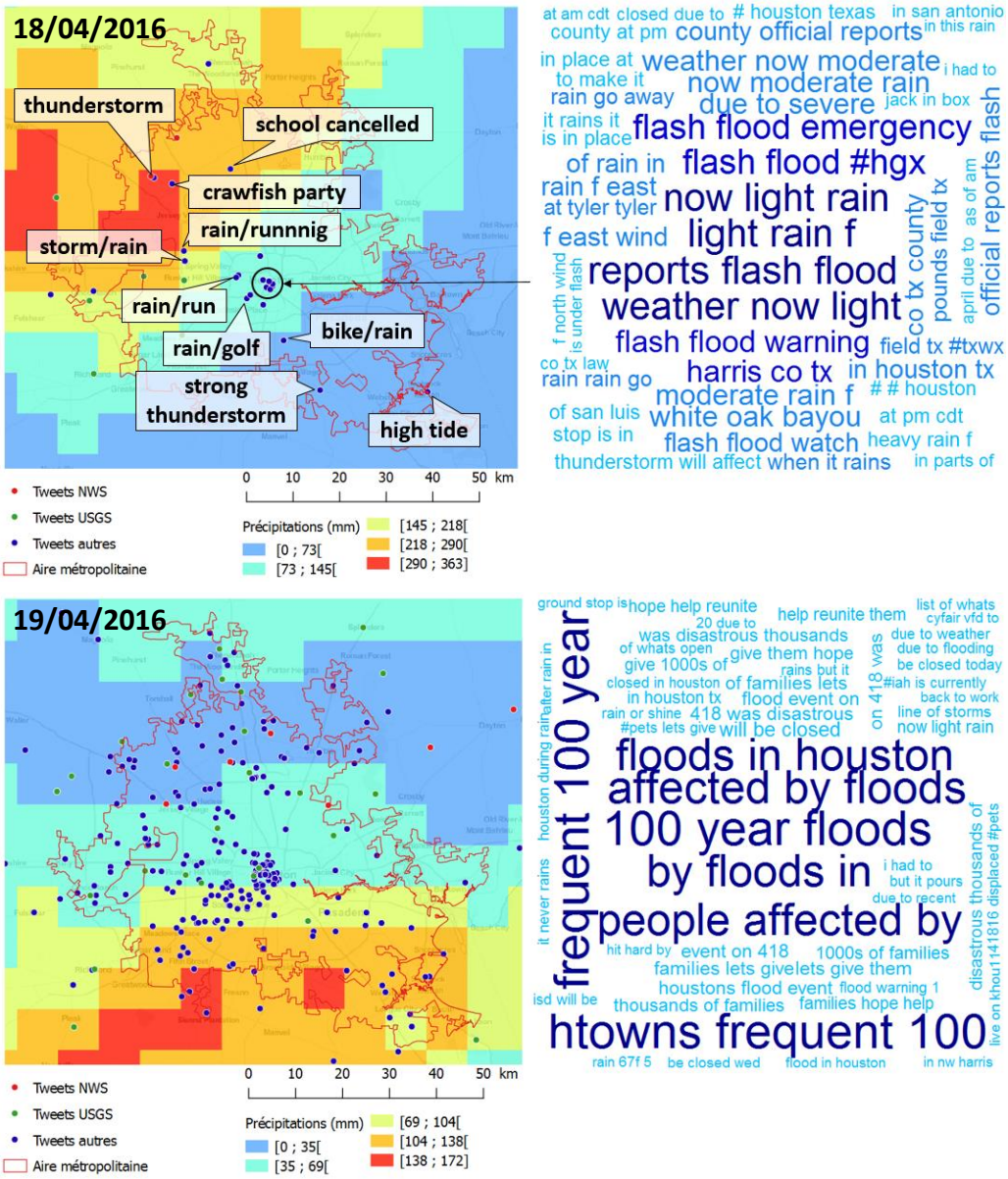


Figure 5.31 : Activité tweeting de crise à Houston - 18/04/2016 et 19/04/2016 (C.Cavalière)

- les tweets émis le 18 avril témoignent de pluies en cours à des intensités variables (*light rain*, *moderate rain*, *heavy rain*), d'inondations prévues et d'alerte aux crues éclair en cours (*reports flash flood*, *flash flood warning*). En outre, on retrouve des informations traduisant des comportements en situation perturbée (en l'occurrence, des activités individuelles ou collectives pratiquées à l'extérieur maintenues malgré les conditions météorologiques : "*rain or shine, there will be one big crawfish party in cypress, tx today*").

- Les tweets émis le 19 avril indiquent un changement de phase dans la crise virtuelle ; plusieurs faits peuvent alors être notés : en premier lieu, on observe un changement de tonalité dans le discours : si l'on focalise l'attention sur le nuage des associations lexicales, les termes les plus fréquents du 18 avril (soit "*weather now light rain*, *flash flood emergency/warning*, *reports flash flood*") sont remplacés, le lendemain, par des termes

faisant directement référence aux conséquences du phénomène naturel : les sinistrés ("*people affected by floods*") ainsi qu'une information révélatrice de l'intensité d'occurrence centennale du phénomène survenu la veille ("*100 years floods in Houston*"). En outre, on ne distingue plus de messages annonçant la poursuite d'une activité malgré les conditions environnementales perturbées ; le vocabulaire témoigne de perturbations et de mesures de sauvegarde face aux dangers, relatives aux sinistrés des inondations ("*people affected by floods*"), aux troupeaux mis en sécurité ("*displaced pets*"), aux axes routiers fermés à la circulation ("*will be closed due to flooding*"). De même, on retrouve les messages de prières/espoir, caractéristiques des situations de crise, quel que soit le type de phénomène survenu dans le monde réel. Malgré tout, les temps verbaux employés semblent indiquer la coexistence de deux événements virtuels : un événement évoquant le phénomène réel au passé et la reprise éventuelle des activités normales ("*was disastrous*", "*after rain*", "*back to work*") ainsi qu'un second événement toujours en cours : "*but it pours*", "*flood warning*". La carte témoigne en effet de la poursuite des pluies soutenues dans la partie sud de la métropole, qui regroupe moins de tweets de crise que le centre et le nord. Une exploration lexicale plus approfondie du réseau peut-elle alors indiquer la coexistence d'un territoire en phase de résilience et de sauvegarde alors qu'un autre territoire reste en crise au sud ?

Nous avons alors extrait et analysé la sémantique des tweets de crise inclus dans différents territoires de l'aire métropolitaine, en fonction des cumuls pluviométriques enregistrés le 19 avril 2016 et de la distance des tweets entre eux. La figure 5.32 présente alors la sémantique des tweets analysés sous forme de nuages de mots (simples si le nombre de tweets collectés est inférieur à dix, et sous forme d'associations lexicales si le nombre de tweets collectés est supérieur à dix).

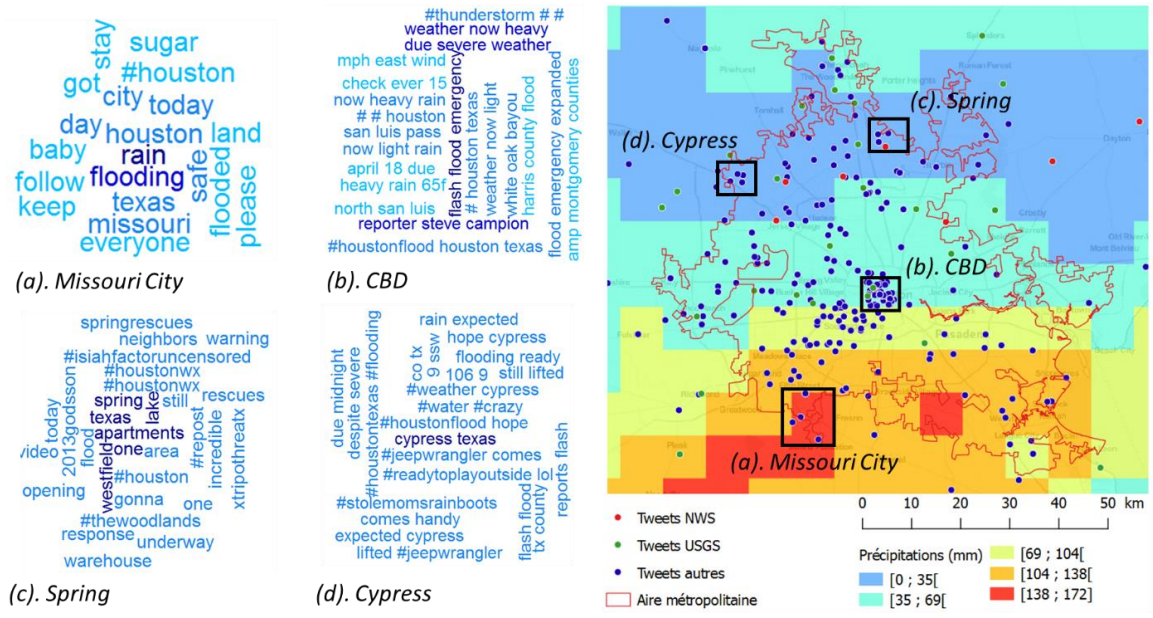


Figure 5.32 : Exploration sémantique de poches d'activités en fonction des cumuls pluviométriques sur Houston le 19 avril 2016 (C.Cavalière)

Les poches d'activité explorées dans les différents lieux de l'aire métropolitaine témoignent-elles de situations variées au 19 avril 2016 ? *A priori*, les tendances relevées dans le discours général représenté par la figure 5.31 ne sont ici pas perceptibles : à *Missouri City* (zone rouge), le lexique témoigne de pluies (*rain*) et de zones inondées (*flooded*, *flooding*) ; en revanche, les termes employés ne fournissent pas d'indications afin d'estimer l'intensité des précipitations et des inondations. En fait, lorsqu'on parcourt les nuages de mots et d'associations lexicales, les situations d'urgence et lieux affectés semblent davantage se concentrer dans le *CBD* (*flash flood emergency extended*, *severe weather*, *heavy rain*) ainsi que dans la ville de *Spring* (*rescues*, *warning*, *response* ainsi qu'un tweet indiquant qu'une entreprise met à disposition ses entrepôts pour les familles évacuées). L'éventuelle diversité des situations d'échelle locale ne transparaît pas dans les nuages des tendances globales. Ce constat pourrait s'expliquer par la forte concentration de tweets de crise géolocalisés dans le *CBD* : la lecture des tweets en question révèle que l'information diffusée depuis le centre des affaires arbore une tonalité générale ou médiatique, et concerne des territoires autres que ce centre, et de diverses échelles : "*happening now: people and pets being rescued from Houston flood in Waller County*" (le comté de *Waller* est situé dans la partie nord-ouest de la métropole) ; "*Jersey Village flooding: residents rescued by boat from homes. @HoustonNews*" (*Jersey Village* est un quartier résidentiel aisé situé à 25 km au nord-ouest du *CBD*).

De la même manière, il est difficile de savoir si les poches d'activité locale explorées reflètent de manière objective cette diversité des phénomènes et événements résultants, alors en cours dans les lieux en crise (cf. figure 5.32) :

- d'une manière générale, les quatre lieux explorés sont inondés au 19 avril ;
- les deux lieux de *Cypress* et de *Missouri City* ont subi des cumuls pluviométriques conséquents, respectivement le 18 avril puis le lendemain, mais seuls les tweets de *Missouri City* font état d'évacuations ; à *Cypress*, on pourra cependant identifier un vocabulaire d'intensité ressentie (*crazy*) qui est absent de *Missouri City* ("*#houstontexas #flooding #carstuck #water #crazy #weather @ cypress*") ;
- le centre des affaires est inondé mais les événements locaux sont bruités par la diffusion d'informations relatives à d'autres territoires.

Les 20 et 21 avril, le phénomène intense est terminé pour la métropole : en revanche, l'événement virtuel persiste à moindre mesure (les deux journées cumulent 344 tweets géolocalisés) et semble figé dans cette phase de gestion des perturbations et de sauvegarde (figure 5.33) : on retrouve le vocabulaire lié aux impacts du phénomène sur les populations et animaux ("*displaced pets*"), ainsi que sur les infrastructures et notamment l'aéroport de Houston (*George Bush Intercontinental Airport*, dont le code est IAH) : "*#Iah is currently experiencing departure delays*". De la même manière, on trouve un lexique fournissant des indications quant à l'intensité de cette crise : "*severe flooding*", "*flooding widespread*", ou encore "*Texas record rainfall*" (le lieu d'origine de ces informations reste le *CBD*).

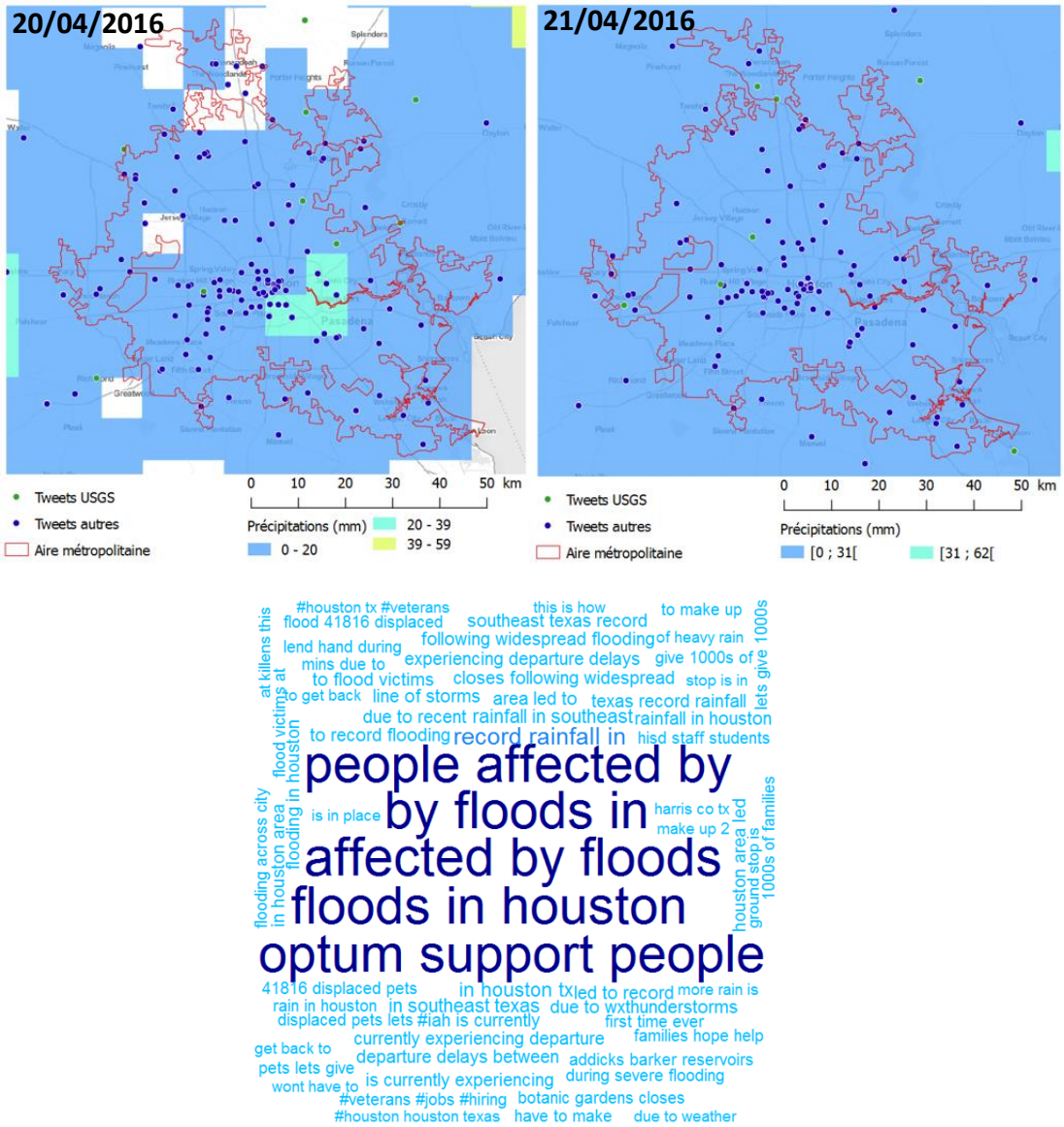


Figure 5.33 : Activité tweeting de crise à Houston - 20/04/2016 et 21/04/2016 (C.Cavalière)

Exploration d'une poche d'activité temporaire en milieu rural (16 et 17 avril 2016, nord de l'Etat du Texas).

La poche d'activité détectée dans le nord de l'Etat les 16 et 17 avril, sur la figure 5.22, correspond à une activité sporadique localisée sur des comtés situés en marge de la hiérarchie urbaine de l'Etat. Il s'agit des comtés de Dallam, Hartley, Sherman, Moore et Oldham (ce dernier est localisé au nord de la ville d'Amarillo et ne contient aucune aire urbaine). Seul le comté de Dallam présente ce profil paradoxal mis en évidence précédemment : les populations sont considérées comme urbaines car concentrées en un pôle urbain local alors que le territoire hors milieu urbain est constitué de champs circulaires. Dans ces territoires en marge des grandes métropoles, le comportement du réseau s'avère radicalement différent (figure 5.34) : en période de crise, on trouve des émissions en quantités quasi équivalentes

quels que soient les acteurs virtuels identifiés (43 tweets *autres* et 43 tweets *NWS/USGS* pour le 16 avril dans les cinq comtés, puis 19 tweets *autres* et 11 tweets *NWS/USGS* le 17 avril, alors qu'à Houston, pour les 18 et 19 avril, on enregistrait 8 fois plus de tweets *autres* que de tweets d'alerte émis depuis ces mêmes comptes officiels). De plus, le contenu lexical des tweets se focalise exclusivement sur la description du phénomène physique en cours. On peut alors connaître la variabilité de l'intensité de ce phénomène en des lieux précis : en l'occurrence, il s'agit d'averses de grêle dont l'intensité est exprimée dans les tweets selon les gradients de l'échelle officielle utilisée par le *NWS* (cf. annexe 3). L'activité non liée aux comptes *USGS* ou *NWS* reste pour autant concentrée dans le pôle urbain local de Dalhart (comté de Dallam) et, d'après la forme du contenu lexical des tweets, semble davantage apparentée à d'autres acteurs officiels locaux (médias, autorités) plutôt qu'à des individus lambda : le 16 avril, les 43 tweets *autres* proviennent ainsi de cinq comptes différents dont quatre d'entre eux cumulent 36 tweets qui arborent la forme suivante : "*at 5.56 PM, Dalhart [Dallam co, TX] trained spotter reports hail of tennis ball size*".

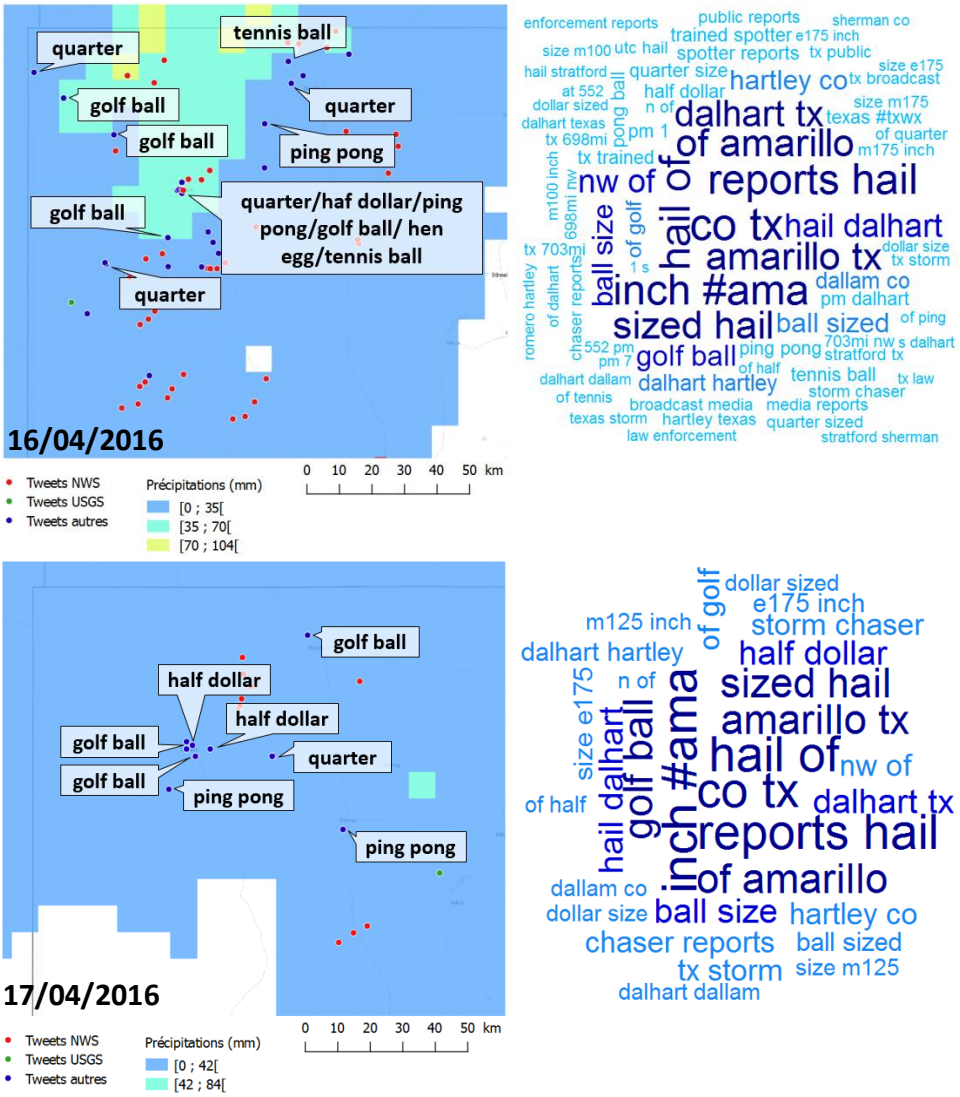


Figure 5.34 : Activité tweeting enregistrée dans le comté de Dallam - 16/04/2016 et 17/04/2016 (C.Cavalière)

A ce stade des analyses, le terrain métropolitain s'avère certes riche en termes de quantités de tweets de crise géolocalisés émis mais ces traces, considérées à une échelle spatiale fine, restent éparses et passent sans doute sous silence une partie des événements locaux consécutifs aux phénomènes naturels (que se passe-t-il dans les mailles enregistrant les cumuls pluviométriques les plus importants, représentés en rouge [cf. figure 5.32], et qui pourtant ne contiennent que quelques tweets épars qui ne font pas état de situations urgentes ?). L'information géolocalisée diffusée depuis le terrain rural offrirait, quant à elle, une perspective de cartographie de l'intensité du phénomène (en raison de sa cohérence lexicale) mais elle ne renseigne pas les événements locaux et les comportements des individus pendant la crise.

Pour dresser un bilan de ce premier volet d'exploration, nous pouvons souligner les propriétés suivantes du tweet géolocalisé en situation de crise naturelle : l'événement virtuel a la capacité de mesurer l'intensité du phénomène physique en des points précis du territoire, et quel que soit le milieu à l'étude (métropole comme pôle urbain en milieu rural). Cette mesure passe par l'emploi d'un vocabulaire précis et officiel (comme l'échelle des grêlons du *NWS*) ou de qualificatifs (*light/heavy* pour la pluie, *severe/widespread/100 year flood* pour les inondations). En revanche, les temporalités de l'activité *tweeting* et leur signification se révèlent différentes en fonction du milieu : dans la métropole qui reste active sur le réseau, on identifie des témoins d'une période de crise et d'une période de mise en œuvre de mesures de sauvegarde, qui cumule les plus fortes quantités de tweets alors qu'en milieu rural, l'activité virtuelle s'exerce exclusivement pendant la manifestation du phénomène naturel et reste peu diversifiée dans ses thématiques. D'après l'exploration sémantique des tendances virtuelles générales et locales, nous pouvons dégager quatre types de contenus de l'événement virtuel (les termes définis en italique seront conservés pour les analyses ultérieures, et le vocabulaire de référence cité provient des nuages de mots des figures 5.29 à 5.34 ou de la lecture des tweets en entier le cas échéant) :

- le contenu relatif à l'*alerte*, diffusé dans des tweets à consonance officielle : "*flash flood warning*", "*flash flood watch*" ;
- le contenu relatif aux *phénomènes physiques en cours*, diffusé dans des tweets à consonance officielle ou dans des tweets émis par des individus transcrivant des observations environnementales directes : "*weather now heavy rain*", "*chaser/spotter reports*", "*flash flood emergency*", "*flooding*" , "*crazy rain*";
- le contenu relatif à des *mesures de sauvegarde* des populations affectées : "*rescued*", "*closed due to flooding*", "*displaced pets*". Nous définissons comme une mesure de sauvegarde des populations (ou animaux) toute action destinée à éviter une mise en danger des individus dans un territoire perturbé (comme la fermeture d'une route inondée) ou toute action destinée à secourir des individus en danger immédiat ;
- le contenu relatif à des *mesures de soutien matériel immédiat* aux populations sinistrées : "*help*", "*support people affected*", "*in response to the floods, Godson is opening his warehouse*". Nous définissons les mesures de soutien aux populations affectées comme tout

acte mentionnant des actions immédiates d'entraide auprès des sinistrés. Ce vocabulaire, perçu le 19 avril 2016, se retrouve mêlé aux messages de prières "*hope*" ; il reste présent dans les tendances globales des 20 et 21 avril 2016 et peut alors être qualifié de *mesure de résilience*. La résilience qualifiera ici toute action postérieure à la survenue du phénomène physique et destinée à retourner, dans les plus brefs délais, à l'état d'équilibre initial entre environnement, territoires et populations sinistrées[11]. Dans ce premier cas d'étude, la thématique de la résilience reste peu diversifiée : elle concerne majoritairement la levée de fonds et la collecte de biens matériels "*organizations collecting goods*", "*need donated items*", "*raised funds for flood victims*" ; dans un deuxième temps, le trafic aérien de l'aéroport intercontinental qui reste perturbé "*#IAH is currently experiencing departure delays*" ; enfin, l'expression "*get back to*", visible sur la figure 5.33 nous a interpellés : il s'agit cette fois d'un tweet personnel "*happy to get back to work today; a nice break from being trapped indoors from the #houstonflood"*, émis dans le sud-ouest de la métropole, à 10 km au nord de Missouri City (cf. figure 5.32). L'auteur de ce tweet informe ainsi la possibilité de sortir et de circuler (et effectivement, le tweet est localisé sur une rocade). Un seul tweet a été émis, les jours précédents (en l'occurrence le 18 avril), dans un rayon de cinq cents mètres : il signale alors une inondation (mais sans qualificatif supplémentaire précisant son intensité ressentie par l'utilisateur). Peut-on alors considérer deux tweets comme des marqueurs d'un lieu moins affecté ou plus résilient face au phénomène naturel ?

- Enfin, le contenu relatif à tout autre type d'*informations* qui décrivent le phénomène *a posteriori* "*record rainfall*", "*100 year floods*".

Visibilités officielle et numérique des phénomènes et des territoires.

A l'aide du bulletin de la SEDB publié en avril 2016, nous cherchons maintenant à tester la visibilité numérique des phénomènes physiques explorés par les tweets de crise géolocalisés : le phénomène inventorié dans la SEDB est-il présent sur le réseau ? Le réseau a-t-il la capacité d'enrichir voire d'identifier un phénomène non inventorié dans la SEDB ? En ce qui concerne l'événement virtuel identifié du 16 au 21 avril 2016, la SEDB répertorie des phénomènes de types grêle, tornade, vents violents, crues éclair et inondations dans 80 comtés. Ces 80 comtés regroupent 74% de l'ensemble des tweets émis pendant la période indiquée, tous comptes confondus. Dans les comtés étudiés ci-avant et inventoriés dans la SEDB, le contenu sémantique des tweets correspond aux descriptions des phénomènes et événements réels :

- dans le comté de Oldham (nord de l'Etat), les types de phénomènes et dates de survenue concordent : la SEDB mentionne des averses de grêle, de taille *half dollar*. En revanche, le réseau s'avère plus précis quant à l'identification des lieux affectés : alors que la SEDB indique le seul lieu-dit de *Vega*, les tweets permettent d'identifier deux lieux supplémentaires : *Boys Ranch* et *Wildorado*.

[11] Soit la définition donnée dans l'introduction du manuscrit (Hecker, 2014).

- Dans l'aire métropolitaine de Houston, qui englobe le comté de Harris et s'étend sur les comtés limitrophes de Galveston, Montgomery, Waller, Fort Bend et Brazoria, on trouve une description plus détaillée de l'événement réel, copiée ci-dessous dans la figure 5.35 :

> Heavy rain caused extensive flooding especially over western half of the county where 10 to 15 inches of rain fell in less than a 12 hour period. An estimated 40,000 vehicles and 10,000 homes were flooded. Seven people in Harris County drowned in their vehicles when they drove into flooded roadways. There were numerous high water rescues. F49VE, M49VE, M56VE, F41VE, M66VE, F25VE, M61VE.
>
> A slow moving upper low over the Southwestern U.S. combined with near record level moisture aided in producing extremely heavy rainfall and devastating flooding over portions of Harris, Waller and Fort Bend Counties. Northwest to southeast orientated bands of precipitation commenced during the early evening hours of April 17th across extreme southwestern and western Harris County as well as north and west into Grimes, Waller, Fort Bend, Austin and Colorado Counties. Between 8:00 p.m. and 9:00 p.m. thunderstorms began to greatly intensify and slow their northward movement over Waller County and, by late evening, had stalled and began shifting eastward into western Harris County. Excessive rainfall spread across northwestern Harris County during the late evening hours of April 17th and into the early morning hours of April 18th. Slow thunderstorm movement and rain rates over 4 inches per hour resulted in a large portion of northwest Harris and Waller Counties receiving between 10 and 20 inches of rainfall over mainly a 12 hour period. A few CoCoRaHS gauges in Waller County measured over 20 inches.
>
> The flooding resulted in 9 direct fatalities over the region, all drownings in vehicles. Seven of these were in Harris County with 1 in Waller County and another in Austin County. An estimated 40000 cars and trucks were flooded. Several bayous and creeks were flooded. The Addicks Barker Reservoir was severely impacted. At least 10,000 homes were flooded. Damage was estimated from Damage Survey Reports to be near $60 million.

Figure 5.35 : Description du phénomène de pluies inondations des 18 et 19 avril 2016 sur les comtés de l'aire métropolitaine de Houston (NOAA, Storm Events DataBase)

Dans le cas de la métropole, la SEDB peut s'avérer plus précise en termes de chiffres : de "*thousands of families*" dans les tweets, on trouve dans la source officielle le chiffre *d'au moins 10 000 foyers et 40 000 véhicules inondés*. En revanche, les tweets restituent le nombre exact de sept victimes noyées dans leur véhicule sur des routes inondées pour le comté de Harris. Par ailleurs, le mot *victims* apparaît de manière ponctuelle mais sans faire partie du vocabulaire le plus fréquent (dans l'aire métropolitaine, le terme apparaît le 19 avril mais ne se retrouve que dans 22 tweets). Après vérification des tweets concernés, seuls trois d'entre eux évoquent les individus noyés dans leur véhicule ; les autres tweets emploient ce terme en évoquant l'aide aux sinistrés des inondations. En ce qui concerne les paramètres physiques du phénomène[12], les tweets nous apportent une information absente de la description de la SEDB : l'occurrence centennale du phénomène d'inondation survenu à partir du 18 avril.

Qu'en est-il des territoires *invisibles* sur le réseau ? En moyenne, 96% des mailles de dix kilomètres enregistrant des précipitations pendant chaque journée de la période considérée sont vides de tweets de crise géolocalisés : ces territoires affectés sont donc invisibles sur le réseau. De la même manière, parmi les 80 comtés ayant subi un phénomène inventorié dans la SEDB entre le 16 et le 21 avril 2016, cinq ne disposent d'aucun type de tweet géolocalisé. Pour autant, parmi les 75 comtés enregistrant un événement virtuel, 50% disposent de moins de douze tweets pour décrire le réel. Ces territoires invisibles (ou peu présents sur le réseau) se retrouvent à toutes les échelles : sur l'aire métropolitaine de Houston, la perturbation arrive

[12] Dans la SEDB, la description physique des différents stades du phénomène dans le temps est assez complète. Nous avons également vu que le réseau inventoriait différents adjectifs relatifs à l'intensité des précipitations.

le 18 avril au nord-ouest, par le comté de Waller. Comme l'indiquaient les figures 5.29 à 5.33 et la description de la SEDB, les précipitations les plus intenses ont été enregistrées sur les territoires ouest et sud de l'aire métropolitaine. Or, pour les journées du 18 au 21 avril 2016, ces espaces les plus violemment affectés ne regroupent qu'une poignée de tweets[13] (le *hotspot* de l'activité *tweeting* de crise de Houston semblant concentré dans le CBD de la métropole).

A contrario, certains espaces non présents dans les données officielles s'agitent sur le réseau : la ville frontalière d'El Paso, dans l'ouest du Texas, reste active pendant l'ensemble de la période alors qu'elle est affectée par des précipitations modérées pendant la seule journée du 17 avril 2016 et qu'aucun phénomène n'est répertorié dans la SEDB. Après exploration, la persistance de cette activité virtuelle est liée à un unique compte qui annonce un possible épisode pluvieux le 17 avril (confirmé par les données officielles) mais qui évoque principalement, en téléprésence, l'événement réel survenu à Houston (un habitant d'El Paso étant compté parmi les sept victimes des inondations dans le comté de Harris).

5.2.1.4. Existe-t-il une répétivité des structures identifiées ?

Conformément à la démarche de recherche expliquée dans le chapitre 4, nous cherchons à savoir si les structures de l'activité *tweeting* mises en évidence dans le paragraphe précédent s'identifient à chaque nouvel événement. En d'autres termes, nous considérons les résultats de ces premières analyses comme des hypothèses qui vont être testées dans le présent paragraphe, à partir d'un autre phénomène réel :

- Hypothèse 1 : l'activité *tweeting* majeure enregistrée sur le réseau est le fruit de populations urbaines habitant les métropoles ; dans ces milieux particuliers, le réseau a la capacité de détecter des comportements, de restituer les différentes phases de la crise ainsi que des informations quantitatives vérifiées dans la SEDB. En outre, il semble que la phase associant mesures de sauvegarde et d'amorce d'actions de résilience, draine les flux de tweets les plus conséquents.
- Hypothèse 2 : les lieux de concentration de l'activité *tweeting* de crise sont des individus hors-normes : une faible surface concentre la majorité des tweets de crise émis en réponse au phénomène. Ainsi, quels que soient l'échelle géographique et le milieu considérés, on trouvera des territoires affectés dans le réel mais invisibles sur le réseau.
- Hypothèse 3 : certains espaces restent des poches d'émissions peu nombreuses mais actives en permanence, que le territoire concerné soit affecté ou non par un phénomène naturel : ces émissions apparaissent liées à la téléprésence au phénomène.

Délimitation de l'événement virtuel.

[13] Sans pour autant qu'ils ne fassent partie des marges urbaines de la métropole exclues ou en retard dans l'intégration aux réseaux numériques performants.

La période du deuxième événement virtuel destiné à tester les hypothèses est également identifiée à partir des dates enregistrant des émissions de tweets supérieures à la valeur médiane mensuelle. Si les graphiques de la figure 5.20 indiquent une multitude de pics qui annoncent autant d'événements virtuels, nous ne sélectionnons pas pour autant un pic au hasard : étant donné que nous avons focalisé la précédente étude d'un milieu métropolitain sur Houston, nous nous orientons sur un événement virtuel enregistré à partir d'un phénomène ayant frappé cette même métropole. Par conséquent, nous cherchons d'abord, dans les bulletins de la SEDB du printemps 2016, un phénomène ayant affecté le comté de Harris, puis nous déterminons statistiquement les dates de début et de fin d'événement virtuel :

- d'après le bulletin SEDB du mois de mars 2016, le comté de Harris enregistre, le 18 mars 2016, des averses de grêle ainsi que de forts cumuls de précipitations engendrant des crues éclair dans la ville de Houston. La description complète est copiée ci-dessous, dans la figure 5.36 :

> High water rescues were conducted near the intersection of Interstate 45 and South Lockwood Drive and near the intersection of Almeda Road and Holcombe Boulevard.
>
> A lone thunderstorm formed when a gravity wave interacted with the sea breeze and a slowing cold frontal boundary as it neared the city of Houston. This storm produced large hail and high rainfall rates that lead to flash flooding within the city.

Figure 5.36 : Description du phénomène physique ayant frappé le comté de Harris le 18 mars 2016, (NOAA, Storm Events DataBase)

- d'après ce même bulletin, 41 comtés ont été frappés, entre le 17 et le 19 mars, par des averses de grêle ou des vents violents. Parmi eux, le comté d'Orange, situé à l'est de Houston, à la limite de la Louisiane, a été frappé par des inondations conséquentes à un épisode pluvieux intense survenu la semaine précédente (figure 5.37) :

> Flood water from heavy rain on the 9th and 10th gradually flowed down the Sabine River to Orange County. The river at Orange crested on the 17th at 7.62 feet. This was 2.24 feet below the record set during Hurricane Ike in 2008 and was the second highest crest recorded. Areas north of Interstate 10 had much higher levels since the freeway and railroads acted as a damn holding back water. Most homes along the river flooded with some taking several feet of water, especially north of I-10.
>
> Roughly 1,500 structures were affected during the event with 190 structures flooded which includes 177 residences. This caused an estimated $2 million in damage to public infrastructure and roughly $4.4 million to private structures.
> Multiple days of heavy rain fell across the Sabine River Valley causing massive flooding in the basin. Across the Toledo Bend Reservoir rainfall amounts averaged 15 to 20 inches. This pushed the lake level to a record of 174.36 which is several inches higher than the previous record set in 1989. All operational flood gates were fully opened to stabilize and gradually lower the lake level. Two gates were kept shut since maintenance were being performed. An estimated 205,000 cfs was being released at the peak of the event and this produced record flooding at most sites downstream and north of Interstate 10.

Figure 5.37 : Description des phénomènes ayant affecté le comté d'Orange, le 17 mars 2016 (NOAA, Storm Events DataBase)

- pour le mois de mars 2016, l'étude des paramètres de position des tweets géolocalisés indiquent une médiane située à 106 tweets quotidiens émis pour les tweets de la catégorie *autres* ; la valeur médiane des tweets émis par les comptes du *NWS* est de 3. Nous retenons alors, comme période d'événement virtuel, les journées comptabilisant des émissions supérieures à ces valeurs, soit un événement réseau enregistré entre le 17 et le 20 mars 2016 inclus.

Répartition de l'activité tweeting à l'échelle globale.

La tendance de concentration spatiale mise en évidence dans la section précédente s'avère de nouveau palpable dans l'événement virtuel détecté du 17 au 20 mars 2016 (tableau 5.3). Les comtés urbains dans lesquels au moins un phénomène est inventorié dans la SEDB regroupent plus de 64% des émissions de tweets de crise non liées à des comptes officiels (alors que seules 10,67% des émissions sont localisées en milieu rural et hyper-rural). La localisation des tweets du *NWS* et les informations récoltées dans la SEDB indiquent un phénomène physique survenu dans les comtés attenant à l'aire métropolitaine de Dallas : ce phénomène étant susceptible d'introduire un biais dans les quantités de tweets émis en milieu urbain, nous avons calculé le rapport, pour chaque profil de comté, entre les émissions de tweets catégorisés comme *autres* et les émissions liées au *NWS* (tableau 5.3)

Tableau 5.3 : Emissions de tweets de crise géolocalisés par profil de comté, du 17 au 20 mars 2016

Profil Comté	Nombre de tweets NWS	Nombre de tweets autres	Fréquence tweets autres (%)	Rapport tweets autres /tweets NWS
Urbain	88	404	64,43	4,59
Périurbain	65	156	24,88	2,4
Rural	35	51	8,13	1,46
Hyper-rural	13	16	2,55	1,23

Indéniablement, sensibilité et réactivité des utilisateurs de Twitter s'inscrivent davantage dans les milieux urbains et particulièrement en milieu métropolitain : pour l'émission d'un seul tweet d'alerte en milieu urbain, la réactivité (et donc l'implication des individus à tweeter) est presque quatre fois plus forte qu'en milieu hyper-rural. Pour autant, sur les 41 comtés texans répertoriés dans la SEDB entre le 17 et le 20 mars, sept restent invisibles sur le réseau dont cinq contiennent des populations urbaines situées en dehors des aires métropolitaines de l'Etat (figure 5.38).

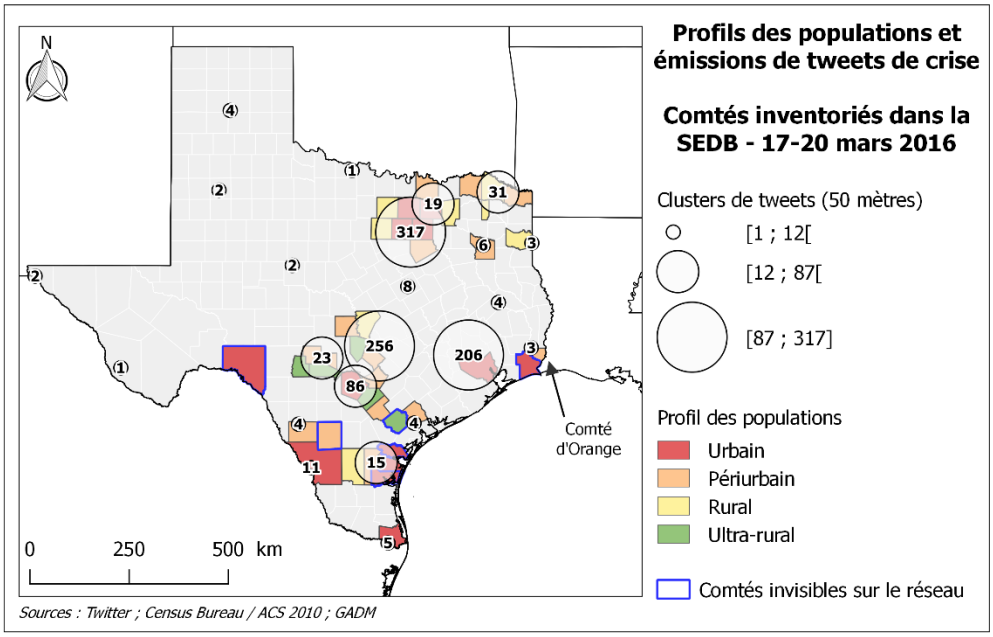


Figure 5.38 : Foyers d'émissions de tweets géolocalisés clustérisés et comtés inventoriés dans la SEDB du 17 au 20 mars 2016 (C.Cavalière)

Exploration de l'activité tweeting liée à l'aire métropolitaine de Houston.

L'événement virtuel enregistré sur la métropole du 17 au 20 mars s'avère de moindre ampleur que précédemment, malgré la survenue de précipitations intenses accompagnées de crues éclair : sur cette période de quatre jours, la métropole de Houston n'enregistre en effet qu'un total de 106 tweets géolocalisés. D'après l'exploration de leur contenu sémantique, on peut de nouveau dégager les différentes phases de l'événement virtuel (figure 5.39) :

Figure 5.39 : Nuages de mots des tweets de crise émis sur l'aire métropolitaine de Houston, du 17 au 20 mars 2016

- une phase de *prévision* (mais sans alerte, le mot *warning* étant quasiment absent) : le 17 mars, le réseau annonce la survenue de "*scattered thunderstorms*" prévue le lendemain. Les tweets mentionnant ce type de phénomène se présentent cependant sous la forme de bulletins météorologiques : "*7:30PM: Sunset Tomorrow's forecast for Pasadena: scattered thunderstorms, 82/63°F*". Contrairement à l'événement virtuel enregistré à la mi-avril, on ne détecte pas de tweets relatifs à des problématiques comportementales individuelles ou collectives d'anticipation. Les tweets mentionnant le terme de *safety* font ici référence à des offres d'emploi.

- Une phase associant *prévision* et *description* du phénomène physique en cours : le 18 mars, certains tweets contiennent de nouveau le vocabulaire des bulletins météorologiques "*current weather*", "*todays forecast*" ; d'autres tweets de crise font directement référence à l'épisode orageux en cours ("*strong thunderstorm*", "*thunderstorm over sugarland*") ainsi qu'aux averses de grêle ("*reports hail*"), en ayant toujours recours à l'échelle définie par le *NWS*. En revanche, le phénomène de crue éclair reste peu visible sur le réseau. Seuls deux tweets sur les 54 émis dans l'aire métropolitaine le 18 mars en font mention : un tweet évoque des routes inondées à proximité d'un centre médical (mais le nom de l'établissement reste inconnu dans le texte du tweet) ; le second tweet fait état de la survenue d'une crue éclair (sans mentionner le nom du cours d'eau) et se trouve géolocalisé à 878 mètres de l'intersection évoquée dans la SEDB (*Almeda Road*/*Holcombe Boulevard*). De même, le tweet

le plus proche de l'intersection entre l'*Interstate 45* et *South Lockwood Drive* se situe à plus de 3 km et ne mentionne aucune inondation ou crue éclair.

- Une phase rapide associant *phénomènes en cours* d'une part, et *mesures de sauvegarde* d'autre part s'observe uniquement le 19 mars : les tweets évoquent des routes fermées en raison des inondations ("*closed due to flooding*" et cette fois, il est question de l'*Interstate 45*). Cette même journée du 19 mars révèle une information supplémentaire quant aux perturbations consécutives au phénomène naturel : l'explosion d'un réservoir d'hydrocarbures frappé par la foudre "*tank explodes after lightning strike*".

Le 20 mars, l'événement virtuel est terminé : seuls sept tweets géolocalisés sont émis dans l'aire métropolitaine et le terme de "*safety*" se trouve de nouveau associé à des offres d'emploi (de même, l'expression "*first aid*", visible dans le nuage de mots du 19 mars faisait référence à une annonce d'emploi).

Visibilité des territoires et des phénomènes sur le réseau en dehors des aires métropolitaines.

Afin de vérifier les dernières hypothèses formulées, nous nous appuyons de nouveau sur la SEDB pour explorer d'une part, le contenu des tweets émis dans des territoires non métropolitains affectés et d'autre part, les événements détectés sur le réseau sans qu'un phénomène naturel ne soit officiellement enregistré. Notre attention se focalise en premier lieu sur le comté d'Orange, frappé par des inondations survenues les 9 et 10 mars, qui se poursuivent jusqu'au 17 mars. Ce comté est classé comme périurbain : il jouxte l'aire urbaine de Beaumont, située dans le comté voisin de Jefferson. Du 17 au 20 mars, le comté d'Orange n'enregistre que trois tweets de crise mentionnant une averse de grêle (ils sont émis le 18 mars). Aucun événement virtuel mentionnant les inondations ou leurs conséquences n'est détecté bien que les informations contenues dans la SEDB décrivent un phénomène majeur lourd de conséquences. En fait, l'événement virtuel lié au phénomène d'inondation de ce comté est quasiment imperceptible : si l'on filtre les tweets de crise émis dans le comté d'Orange du 7 au 12 mars, on ne retourne que deux tweets. Le premier, émis le 9 mars, évoque la survenue d'une crue éclair ; le second, émis le 10 mars, évoque les précipitations mais sans aucune description de leur intensité.

En ce qui concerne la recherche d'un foyer d'émission de tweets de crise en l'absence de phénomène inventorié par les sources officielles, nous avons identifié l'aire métropolitaine d'Austin (la capitale de l'Etat). Du 17 au 20 mars, 168 tweets de crise géolocalisés et catégorisés comme *autres* sont émis sans qu'aucun tweet d'alerte *NWS* ne soit détecté (l'activité virtuelle est par ailleurs plus forte qu'à Houston pour la même période). S'agit-il alors d'une agitation faisant écho à des phénomènes distants, d'un événement réel manqué par la SEDB ou d'une sensibilité à une pluie sans conséquence majeure ? L'analyse lexicale met en évidence des pluies, vents et orages d'intensités variables (figure 5.40) : "*light rain*", "*it's raining*", "*now thunderstorm*" avec sans doute une phase de précipitations intenses : "*severe*

storm", "*showers storms*". Pour autant, aucune expression relative à une inondation ou crue éclair n'est détectée (figure 5.40).

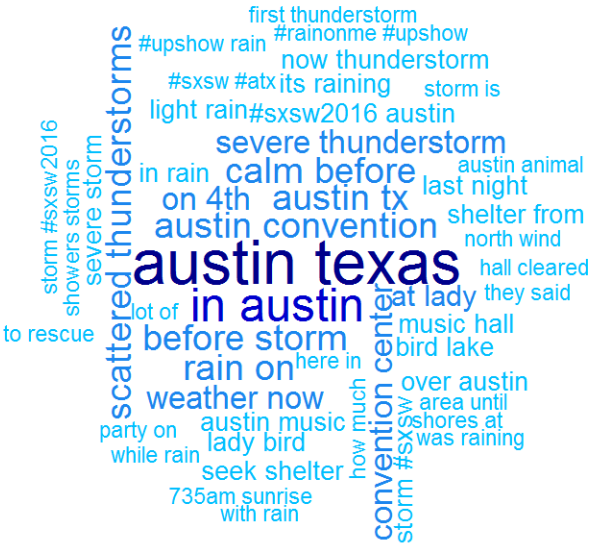


Figure 5.40 : Cooccurrences lexicales des tweets géolocalisés émis sur l'aire métropolitaine d'Austin - 17/03/2016-20/03/2016

Néanmoins, une série de termes ou de cooccurrences lexicales visibles sur le nuage de mots précédent nous ont interpellés :

- le terme de *shelter* (visible au bas et également à droite de la figure 5.40) apparaît spontanément le 18 mars entre 18h et 19h; il est susceptible de rapporter un événement réel consécutif aux pluies et orages. Nous avons alors vérifié le contexte associé à ce mot : les six tweets contenant ce terme ne font pas référence aux refuges mis à disposition des habitants en cas de catastrophe naturelle mais correspondent à des invitations à se mettre à l'abri dans l'immédiat alors que les orages débutent ; ils sont donc associés à une thématique comportementale de protection face à un danger éventuel.

- L'association *they said* (partie droite de la figure 5.40) peut faire référence à un utilisateur relayant un discours provenant d'un acteur officiel. Le tweet en question "*It's just drizzle" they said. "Don't bother with rain gear" they said* ne mentionne nommément aucune source mais semble afficher un décalage net entre le discours entendu par l'utilisateur et la réalité *perçue* de l'intensité du phénomène physique au moment de tweeter.

- Quatre tweets emploient le terme *rescue* mais seul un est lié au thème des secours pendant une période de crise : le contenu sémantique est précis en termes quantitatifs mais n'apporte pas l'information souhaitée quant à l'échelle géographique concernée par les chiffres annoncés : "*Spending today w Texas National Guard, which has run 58 missions incl 36 recues, 121 evacs in past 24 hrs".*

- L'expression *Lot of* est employée à deux reprises : un tweet, émis le 19 mars, fait état de la survenue de nombreux coups de tonnerre la *nuit précédente* ; le second tweet, émis le 20 mars, fait état d'inondations ayant causé de nombreux dégâts à Austin, *l'année précédente*.

- Une dernière association lexicale a suscité notre intérêt : *Austin convention*. S'agissait-il d'un événement culturel, social ou politique annulé/reporté ? Dans les faits, il s'agit de l'*Austin Convention Center* (centre des congrès) qui accueille une partie du festival annuel *South By Southwest* ayant eu lieu du 11 au 20 mars 2016 à Austin (on retrouve par ailleurs mention de ce festival dans les tweets géolocalisés par le hashtag *#SXSW*). Quatre tweets de crise géolocalisés contenant ce hashtag sont inventoriés sur l'aire métropolitaine de la capitale texane dans notre jeu de crise. Ils sont émis le soir du 18 mars, entre 17h39 et 19h07 et concernent tous une problématique comportementale, collective puis individuelle : deux tweets mentionnent que le concert programmé est maintenu malgré les intempéries ; les deux autres tweets informent que leur unique auteur n'a pas l'intention de se rendre à ce concert le soir en question, ni même de sortir, en raison des éclairs.

Nous avons alors recherché l'existence éventuelle d'autres tweets géolocalisés mentionnant des comportements adoptés vis-à-vis de cet événement culturel dans un contexte physique perturbé. A partir du jeu de tweets bruts collectés de mars à juin 2016 sur l'ensemble du Texas, nous avons filtré les tweets géolocalisés contenant le hashtag *#SXSW* émis sur l'aire métropolitaine d'Austin entre le 17 et le 20 mars. Cette requête retourne un total de 8 740 tweets géolocalisés. Le premier constat est formel : le phénomène pluvio-orageux (pour rappel, 168 tweets géolocalisés extraits) est annexe par rapport à l'événement culturel annuel de la capitale de l'Etat. Le tableau ci-dessous (figure 5.41) fait alors état de la recherche de cooccurrences lexicales relatives à d'éventuelles perturbations des événements programmés, consécutives aux phénomènes pluvio-orageux survenus pendant cette période.

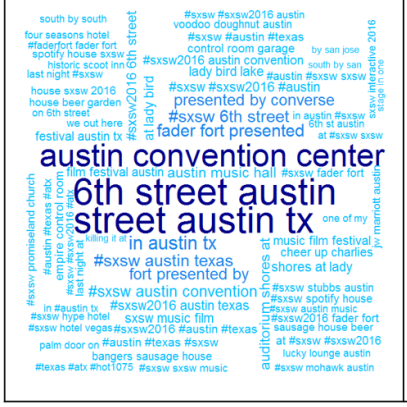
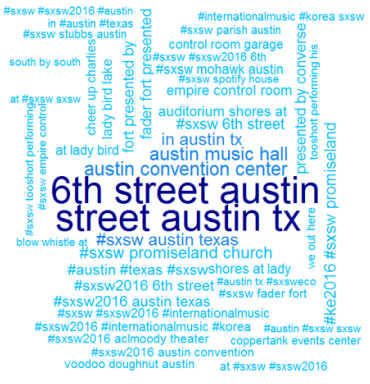
Tous les tweets #SXSW émis entre le 17 et le 20 mars 2016, Austin	Tweets #SXSW émis le 18 mars 2016, entre 17h00 et minuit, Austin	Tweets #Wolfmother émis le 18 mars 2016, entre 17h00 et minuit, Austin
		

Figure 5.41 : Cooccurrences lexicales identifiées dans les tweets géolocalisés liés à l'événement South By Southwest à Austin

Sur l'ensemble du jeu de 8 740 tweets géolocalisés, le vocabulaire météorologique est totalement masqué par l'événement culturel. On peut néanmoins mettre en évidence l'existence d'une minorité de tweets à problématique comportementale à condition de filtrer les cooccurrences lexicales de fréquences les plus basses (1 ou 2 occurrences) et contenant des mots précis comme *rain* ou encore *lightning*. Un tweet émis le soir du 18 mars à 20:10 indique ainsi que son auteur n'est pas effrayé par les intempéries et maintient son intention d'assister à un concert malgré la pluie ("*Braving the rain for @IntoItOverIt here at #SXSW Lightning = not skurred*[14] *#musicglue*"). En filtrant les tweets émis pendant les soirées, c'est-à-dire aux périodes pendant lesquelles les concerts sont programmés, même constat : les tweets à problématique comportementale sont imperceptibles. En revanche, en filtrant par le hashtag correspondant au nom du groupe qui se produit sur scène, on peut enfin souligner l'existence de perturbations de l'événement programmé par les intempéries. Dans la figure 5.41, l'exemple affiché correspond au concert de *Wolfmother* (tweets filtrés à partir du hashtag *#Wolfmother*), le soir du 18 mars, qu'on avait identifié précédemment car associé avec le hashtag *#SXSW*. On peut alors apprendre :

- à 17h39 : en dépit de la pluie, le concert programmé est effectivement maintenu ;
- à 20h05 : le concert est finalement suspendu en raison des conditions météorologiques ;
- au niveau du ressenti des utilisateurs face au phénomène météorologique prévu et aux perturbations envisageables de l'événement programmé : on retrouve les tweets précédemment mentionnés de l'utilisateur qui renonce à sortir, auquel s'ajoute le tweet d'un utilisateur assistant au concert et pensant ne pas être affecté par les pluies, émis à 17h50 ("*Back here for Wolfmother and Twinkies... I'm 90% sure we are staying dry over here at the park #wind #SXSW*").

[14] En argot américain, *skurred* signifie *scared*.

Bilan

L'activité *tweeting* émise en période de crise est avant tout le reflet d'un comportement numérique statistiquement hors-normes, caractéristique des grandes métropoles. Dans ces milieux, l'analyse sémantique des émissions de tweets de crise restitue la succession des différentes phases du phénomène naturel et des événements sociaux en réponse : prévisions, alerte, dispositions individuelles ou collectives de protection, phénomène en cours, mesures de sauvegarde, de soutien et de résilience. En revanche, l'ensemble des thèmes identifiés ne sont pas présents dans la totalité des événements virtuels explorés, en fonction des milieux (métropolitain, rural) mais aussi des échelles (ensemble de la métropole ou quartier). Nous émettons ainsi quatre réserves :

- dans les territoires actifs sur le réseau, la quantité et la richesse lexicale des contenus géolocalisés de crise semblent étroitement liées à la gravité ressentie du phénomène physique : le phénomène de pluies intenses et d'inondations, survenu en avril à Houston, se révèle plus documenté sur le réseau que le phénomène détecté en mars. A ce stade, on ne sait pas quels paramètres entrent en compte, pour un individu, dans sa représentation de la gravité d'un phénomène.

- Nous nous interrogeons sur le sens de la composante spatiale du tweet géolocalisé : à toutes les échelles, nous avons mis en exergue des territoires non affectés dans le réel mais pourtant actifs sur le réseau et, au contraire, des territoires affectés par des phénomènes précis mais invisibles sur le réseau, même en milieu métropolitain.

- De la même manière que les lieux qui concentrent les plus fortes émissions de tweets géolocalisés sont peu nombreux, les émissions de tweets de crise géolocalisés apparaissent comme un événement virtuel mineur aux côtés d'événements réels d'origine non naturelle (à Austin, l'événement virtuel pluvio-orageux enregistré sur le réseau du 17 au 20 mars 2016 ne représentait que 1,9% de l'événement virtuel lié au festival annuel). Par ailleurs, comme indiqué dans le chapitre 2, nous ne pouvons pas entrevoir si les pensées et pratiques de cette minorité d'utilisateurs sont représentatives de l'ensemble des utilisateurs de Twitter ou encore de l'ensemble de la population d'un territoire.

- L'information sémantique affichée dans les nuages de mots révèle les tendances générales, qui arborent un profil d'information diffusée par des acteurs officiels : si ces tendances permettent d'identifier l'existence des thèmes mentionnés ci-avant qui se relaient au cours du temps, elles ne mettent pas en exergue l'information centrée sur les individus au cœur de leur environnement perturbé. Cette information-là, qui peut traduire un comportement ou encore la perception des éléments du territoire dans un contexte particulier, existe sur le réseau mais reste malheureusement éparse.

Bilan (suite)

En conséquence, nous remettons en cause le caractère exhaustif et prédictif des tweets géolocalisés : le tweet de crise géolocalisé n'est pas un matériau exhaustif car, à toutes les échelles considérées, des territoires affectés ne sont pas représentés dans l'événement virtuel. De même, dans les territoires virtuellement visibles, certains phénomènes semblent passés sous silence, comme l'épisode de crues éclair à Houston et les inondations dans le comté d'Orange en mars 2016. Le tweet de crise géolocalisé n'est pas non plus un matériau prédictif : à ce stade, nous ne pouvons pas confirmer que la simple survenue d'un phénomène physique dans le réel activera le réseau et provoquera la création de contenus captant l'ensemble des facettes du réel. Ce dernier paramètre semblerait davantage lié au type de milieu considéré ainsi qu'à la gravité des effets du phénomène pour les populations qui le vivent : dans la métropole de Houston, l'inondation majeure (occurrence centennale) survenue les 18-19 avril 2016 suscite un événement virtuel huit fois plus conséquent que l'événement pluvio-orageux et les crues éclair résultantes au mois de mars. De la même manière, un phénomène d'inondation n'a pas le même retentissement selon qu'il frappe la métropole de Houston ou le comté périurbain d'Orange, limitrophe de la Louisiane (alors que la SEDB indique des dégâts importants et coûteux). Enfin, à Austin, l'événement virtuel consécutif au phénomène naturel se trouve masqué par l'événement virtuel faisant écho à un événement culturel.

5.2.2. Spatialisation et temporalités d'un événement virtuel en réponse à un phénomène physique extrême rare

5.2.2.1. Logique de distribution spatiale des tweets de crise géolocalisés à l'échelle globale du phénomène

Ce premier paragraphe présente un nouveau test articulé autour de la composante spatiale du tweet de crise géolocalisé à l'échelle globale de la survenue du phénomène : précédemment, nous avons émis des réserves quant au sens de cette composante, dans la mesure où elle ne représentait pas l'ensemble des territoires affectés et, considérée à l'échelle locale, ne semblait pas quantitativement proportionnelle à l'intensité des phénomènes. La localisation des tweets de crise en réponse à l'ouragan Harvey suit-elle alors cette même logique de distribution ou peut-on identifier des spécificités aux phénomènes rares ?

De nouveau, la logique de distribution spatiale des tweets de crise géolocalisés n'apparaît pas liée à la variabilité spatiale de l'intensité des précipitations, comme l'indique la figure 5.42. Les foyers d'émission les plus conséquents restent localisés dans les deux grandes

métropoles de l'Etat, Houston et Austin, dont les aires métropolitaines cumulent 61,5% des émissions de tweets de crise géolocalisés émis entre le 23 et le 31 août 2017.

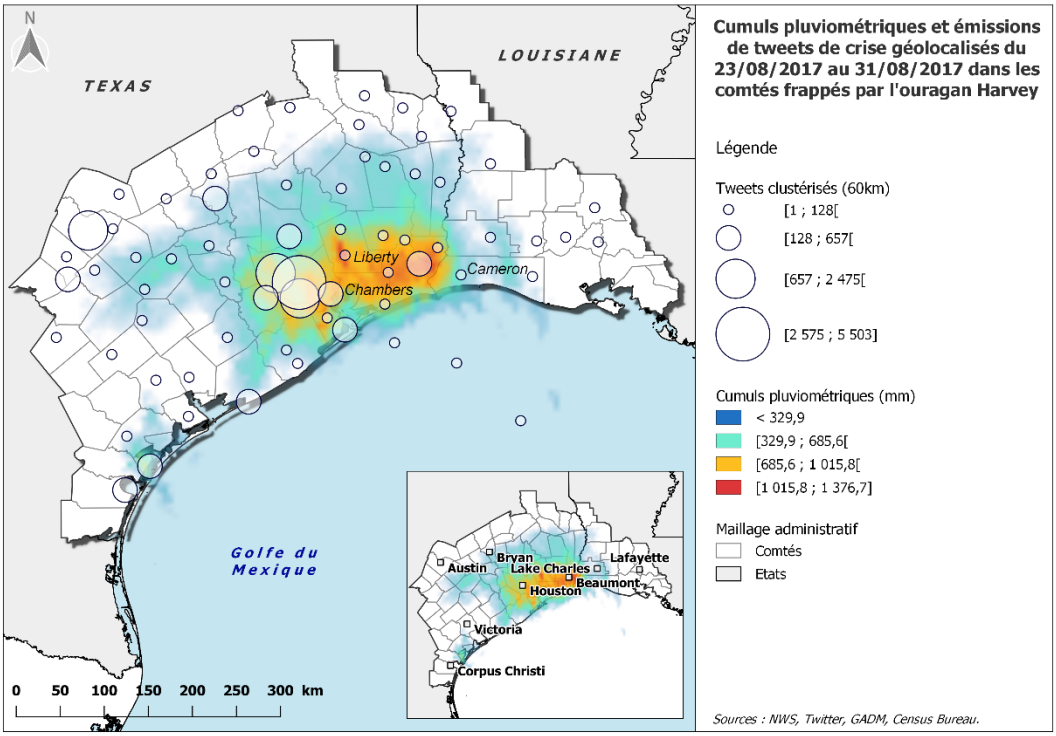


Figure 5.42 : Cumuls pluviométriques (mm) et foyers de tweets émis entre le 23 et le 31 août 2017 – Texas et Louisiane (C.Cavalière)

Que dire alors de l'équité de la représentation spatiale des territoires affectés dans l'événement virtuel ? Si l'aire urbaine de Beaumont constitue un foyer d'émission de tweets de crise ayant subi des cumuls pluviométriques parmi les plus forts, certains espaces enregistrant des cumuls analogues restent en marge de l'événement virtuel, comme le comté de Liberty (situé au nord-est de l'aire urbaine de Houston) : conformément à la typologie des comtés, établie dans le paragraphe 5.2.1.2, il est classé comme comté rural avec trois pôles urbains (Liberty, Dayton et Cleveland, qui forment une conurbation avec la périphérie de l'aire métropolitaine de Houston). En dépit de sa situation géographique (proximité directe à l'aire métropolitaine de Houston) et des cumuls pluviométriques, ce comté n'enregistre qu'un total de 35 tweets de crise géolocalisés, soit 0,18% des émissions totales capturées entre le 23 et le 31 août 2017. Il en est de même pour le comté de Chambers (au sud du comté de Liberty) qui n'enregistre que 40 tweets de crise émis sur la même période ainsi que le comté de Cameron en Louisiane (limitrophe du Texas et sur la côte du Golfe du Mexique) : bien que les territoires situés à proximité du fleuve Sabine à l'ouest figurent parmi les plus forts cumuls pluviométriques, le comté de Cameron n'enregistre que 11 tweets géolocalisés.

En fait, en observant le tableau 5.4 ci-après, on pourra préciser les tendances de localisation des foyers d'émission de tweets de crise en fonction des territoires.

Tableau 5.4 : Proportions quotidiennes de tweets de crise localisés en milieu urbain et métropolitain

Jour	Nombre total de tweets	% de tweets localisés dans les aires urbaines	% de tweets des aires urbaines inclus dans les métropoles (Houston et Austin)
23/08/2017	769	78,7	53,4
24/08/2017	1 102	80,1	67,2
25/08/2017	2 149	75,5	68,7
26/08/2017	2 677	68,5	73,6
27/08/2017	3 726	67,7	84,6
28/08/2017	3 382	78,8	84,9
29/08/2017	3 127	82,1	86,4
30/08/2017	2 501	79,6	82,7
31/08/2017	1 815	79,5	67,9

Malgré la tendance des tweets de crise géolocalisés à se concentrer dans les aires urbaines pendant l'ensemble de la période représentée dans le tableau ci-dessus (qui oscille entre 68,5% et 82,1%), la proportion des tweets géolocalisés augmente dans les deux métropoles de Houston et d'Austin dès qu'elles sont physiquement frappées par l'ouragan (cf. données de précipitations des cartes de la figure 5.43, présentée ci-après) : en effet, si elles ne cumulent que 53,4% des tweets de crise émis en milieu urbain le 23 août, cette proportion augmente progressivement dès le 24 août pour atteindre 86,4% de tweets concentrés dans les deux métropoles le 29 août. En revanche, si l'on considère l'ensemble des aires urbaines affectées, on pourra identifier deux propriétés du comportement de l'événement virtuel généré par l'ouragan :

- si l'on compare les chiffres indiquant la concentration des tweets de crise géolocalisés dans les aires urbaines et dans les deux aires métropolitaines, on constatera rapidement que, en dehors du 23 août, la majorité de l'activité urbaine s'avère en fait métropolitaine (puisque Houston et Austin inventorient à elles seules entre 67,2% et 84,9% de l'activité enregistrée en milieu urbain). La tendance à l'hyperconcentration virtuelle dans une minorité d'espaces persiste à l'échelle globale de l'événement généré en réponse à l'ouragan.

- contrairement aux émissions métropolitaines qui s'accroissent entre le 23 et le 29 août par rapport à l'ensemble des émissions urbaines, le pourcentage de tweets inclus dans l'ensemble des aires urbaines baisse les 26 et 27 août (chiffres représentés en rouge dans le tableau 5.4). En conséquence, on pourra constater l'existence vraisemblable de foyers virtuels ponctuels d'émissions de tweets de crise, localisés dans des espaces physiquement

périphériques. L'activité *tweeting* de ces marges virtuelles est-elle alors uniquement décelable le jour où elles sont frappées par le phénomène physique[15] ?

En réponse au tableau 5.4, si l'on observe maintenant la distribution spatiale quotidienne des tweets de crise géolocalisés par rapport aux pluies (figure 5.43), on constate de nouveau cette tendance à la concentration permanente des émissions en milieu métropolitain : le 25 août, la métropole de Houston est déjà active sur le réseau alors qu'elle n'est physiquement pas affectée (les premières pluies l'atteignent par le sud le 26 août) ; les deux métropoles restent d'ailleurs actives sur l'ensemble de la période.

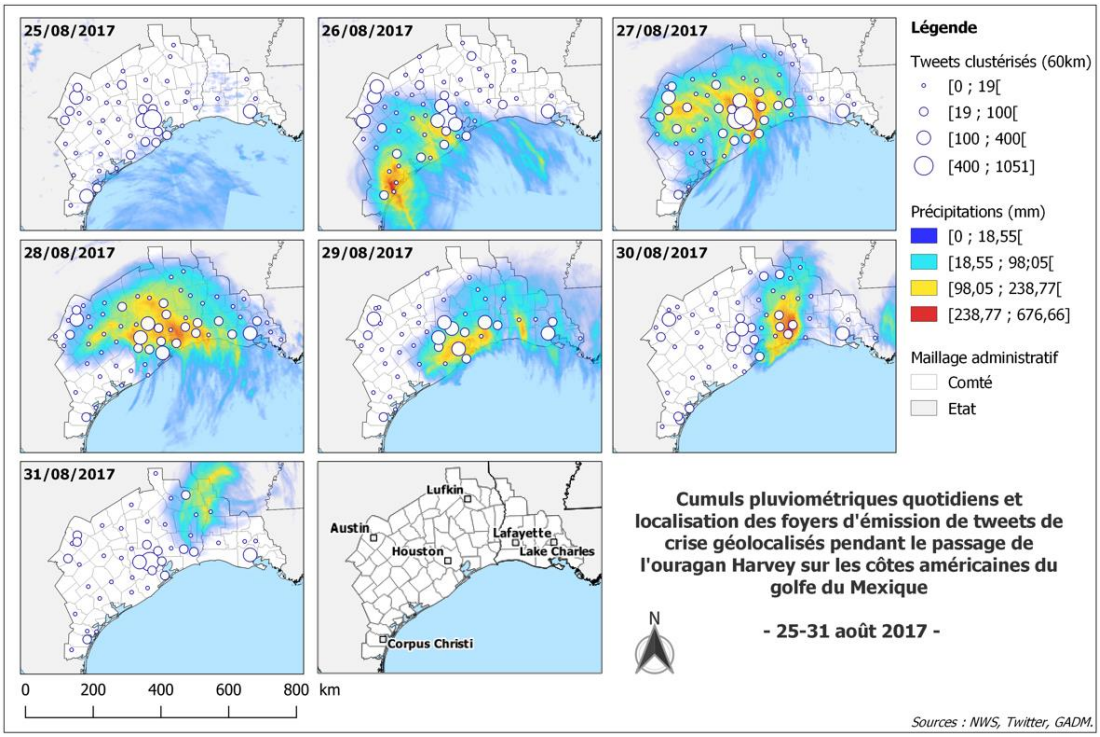


Figure 5.43 : Foyers d'émission de tweets géolocalisés et cumuls pluviométriques quotidiens du 25 au 31 août 2017 (C.Cavalière)

Néanmoins, si l'on appréhende l'activité tweeting de crise indépendamment des deux métropoles, on pourra alors visualiser l'existence d'une dynamique spatiale du réseau virtuel en réponse à la dynamique spatiale de l'ouragan. Malgré tout, la dynamique de l'événement virtuel se repère davantage en termes de *pics d'émissions* de tweets de crise (cf. prochain paragraphe 5.2.2.c) que de *variabilité spatiale* de ces mêmes émissions : pour constituer le graphique 5.44 ci-dessous, nous avons fractionné l'espace réel de crise en mailles de dix kilomètres de côté, qui sont ensuite annotées en fonction de leur activité ou inactivité quotidienne. Les valeurs indiquées dans le tableau correspondent au pourcentage de nouvelles mailles activées au jour indiqué, entre le 23 et le 31 août 2017 (autrement dit il s'agit

[15] L'activité virtuelle concentrée en milieu urbain augmente en effet de nouveau à partir du 28 août et se stabilise entre 79% et 82% de tweets émis depuis ces territoires.

des mailles dans lesquelles on comptabilise au moins un tweet au *jour j* alors qu'elles étaient inactives la veille ou les jours précédents).

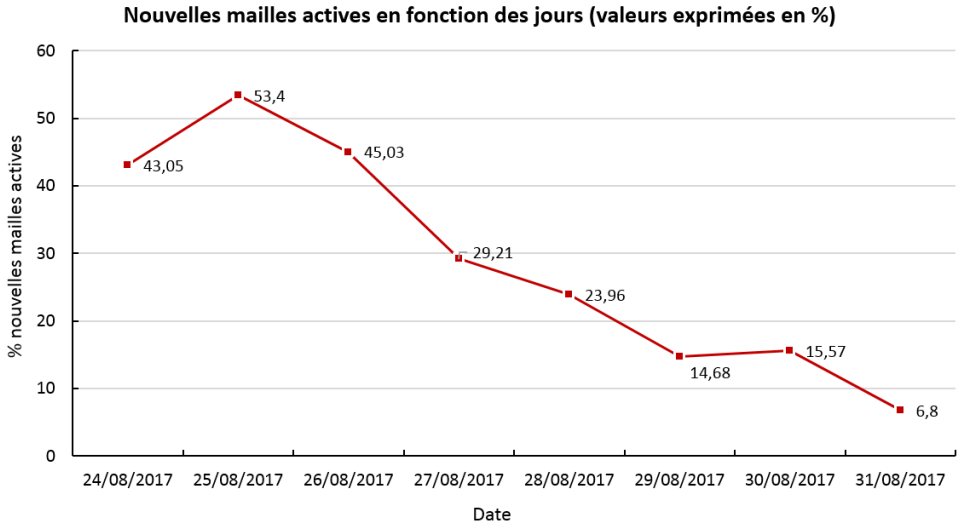


Figure 5.44 : Proportions de mailles nouvellement actives en fonction des jours

On observe une première phase courte d'anticipation du phénomène, pendant laquelle de nouvelles mailles s'activent entre le 24 et le 25 août, suivie par une seconde phase de décroissance, à partir du 26 août, pendant laquelle la proportion de nouvelles mailles actives diminue régulièrement (hormis pour la journée du 30 août, lorsque l'ouragan gagne l'est du Texas et la Louisiane, déclenchant vraisemblablement un nouvel événement virtuel dans des territoires jusqu'alors peu actifs sur le réseau). La plupart des mailles qui s'agitent avant l'arrivée du phénomène restent alors actives pendant et après son passage. Pour autant, les comportements virtuels spatiaux ne peuvent être standardisés (cf. figure 5.43) :

- même si les deux métropoles restent actives en permanence, on observe des variabilités quantitatives en fonction de la dynamique de l'ouragan. A Houston et Austin, les utilisateurs de Twitter communiquent avant l'arrivée du phénomène. Les dynamiques d'émission augmentent dès qu'elles sont frappées par les pluies : les 23 et 24 août, les deux métropoles cumulent 1 052 tweets géolocalisés (soit 56,2% des émissions totales des deux journées) ; les 27 et 28 août, elles cumulent 4 991 tweets de crise (soit 70,2% des émissions totales enregistrées).

- *A contrario*, des territoires s'avérant actifs avant l'arrivée de l'ouragan ont tendance à s'effacer du réseau pendant et après son passage. Ce comportement est nettement perceptible sur le littoral du golfe du Mexique, de Corpus Christi à la baie de Galveston (lagune située au sud de la métropole de Houston) : des foyers d'activité sont visibles le 25 août alors que l'ouragan gagne la côte sud ; le 26, ces foyers persistent mais l'activité s'avère plus ténue que la veille alors que les comtés de Nueces (comté ayant pour siège Corpus Christi) et de San Patricio (comté limitrophe de Nueces, au nord) sur la côte sud sont frappés de plein fouet :

183 tweets de crise sont émis le 25 et seulement 85 le lendemain. Du 27 au 31 août, l'activité virtuelle reste, de la même manière, ponctuelle mais témoigne d'une légère intensification le 30 août (97 tweets enregistrés sur ces deux mêmes comtés).

- enfin, certains territoires, situés à l'intérieur des terres (entre Austin et Houston, et entre Houston et le fleuve Sabine qui fixe la limite entre le Texas et la Louisiane), restent en veille (on distingue ainsi de faibles quantités de tweets de crise émis quotidiennement) mais s'activent dès lors qu'ils sont affectés par le phénomène : les comtés de Liberty, Jefferson, Chambers et Orange (localisés entre Houston et le fleuve Sabine) témoignent de ce comportement. Une activité mineure est palpable dès le 25 août. Contrairement au cas précédent, cette activité s'accentue et reste stable entre le 27 et le 30 août (soit l'ensemble de la période pendant laquelle ces territoires sont physiquement affectés par les pluies les plus intenses). Dans le comté de Jefferson, on passe ainsi de 8 tweets de crise émis le 25 août à une moyenne de 72 tweets de crise émis par jour entre le 27 et le 31 août 2017.

- en Louisiane, on ne distingue qu'un seul foyer d'émission de tweets géolocalisés, au sud de Lake Charles (après vérification, il s'agit d'un compte automatique associé à une station météorologique). Bien que l'activité virtuelle reste très réduite (entre le 25 et le 31 août, elle ne représente que 6,9% des émissions totales et 1,8% si on retire les émissions de tweets liées à la station météorologique), un état de veille est perceptible. Pour autant, les pics d'émissions ne se distinguent que dans les comtés du sud, limitrophes avec le Texas et sur une seule journée, le 28 août, alors que l'ouragan se déplace vers l'est les 30 et 31 août.

5.2.2.2. Représentation et visibilité des territoires sur le réseau

Dans le paragraphe dédié à la visibilité numérique des territoires affectés par des phénomènes naturels récurrents, nous avons mis en évidence l'inéquitable visibilité spatiale des territoires dans l'événement virtuel. Ce constat est identique pour le phénomène extrême peu fréquent : en effet, bien qu'on parvienne à identifier des foyers d'émission extérieurs aux métropoles, plus conséquents en termes d'effectifs de tweets par rapport aux quantités émises en réponse aux phénomènes récurrents et habituels[16], les plus forts flux de tweets restent concentrés dans les aires métropolitaines. Ainsi, en reprenant les mailles de dix kilomètres de côté et en agrégeant le nombre de tweets inclus dans chaque maille du 23 au 31 août 2017, on observe de nouveau une minorité de mailles au comportement hors-normes, concentrant la majorité des émissions de tweets de crise géolocalisés (figure 5.45). Parmi ces huit mailles identifiées, cinq sont incluses dans l'aire métropolitaine de Houston (et 32% des émissions totales s'avèrent en fait concentrées dans cinq mailles de cette métropole) ; une maille se trouve dans la métropole d'Austin ; une maille est localisée entre les aires métropolitaines de San Antonio et d'Austin ; la dernière correspond à l'activité de la station météorologique détectée dans le comté de Vermilion en Louisiane.

[16] Dans l'aire urbaine de Beaumont (comté de Jefferson), on n'enregistrait aucun tweet de crise géolocalisé pendant le mois de mars 2016. Pendant l'ouragan, on enregistré 145 tweets de crise.

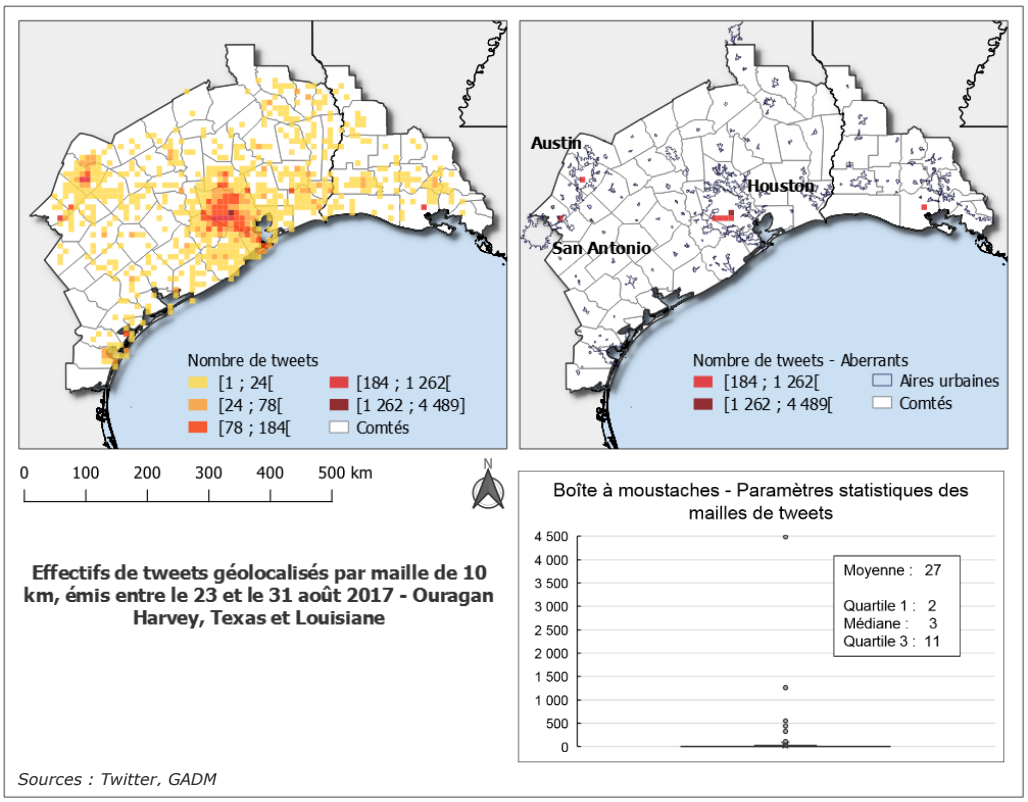
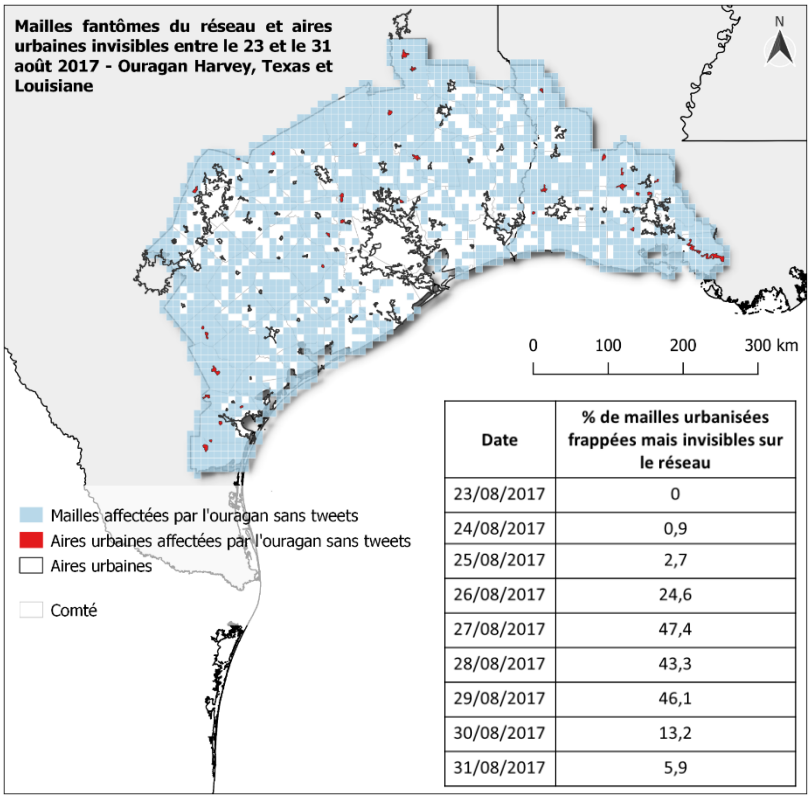


Figure 5.45 : Distribution spatiale et paramètres statistiques des tweets par maille - du 23 au 31 août 2017 (C.Cavalière)

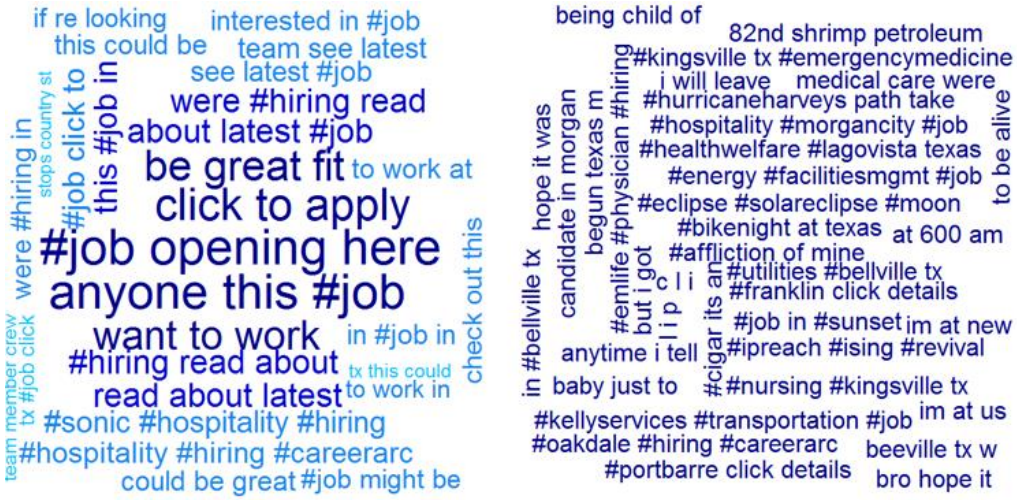
En écho à la carte précédente, la figure 5.46 ci-aprèss met en évidence les *mailles fantômes* dans lesquelles aucun tweet de crise géolocalisé n'a été capturé du 23 au 31 août 2017 ; elles représentent 70,4% de l'ensemble des mailles de la carte. Elles peuvent bien entendu concerner des espaces à faibles enjeux humains mais s'identifient également dans des territoires urbanisés. On les retrouve ainsi dans les marges d'aires urbaines comme San Antonio, Austin ou encore Beaumont. Par ailleurs, certaines villes de Louisiane, ainsi que du nord et du sud-ouest du Texas restent quasiment invisibles sur le réseau (représentées en rouge sur la carte). Le tableau qui accompagne la carte indique alors le pourcentage quotidien de mailles affectées par la pluie et croisant une aire urbaine, dans lesquelles aucun tweet de crise n'est détecté. Du 27 au 30 août, plus de 45% des mailles englobant ou recoupant des territoires urbains sont invisibles du réseau.



Date	% de mailles urbanisées frappées mais invisibles sur le réseau
23/08/2017	0
24/08/2017	0,9
25/08/2017	2,7
26/08/2017	24,6
27/08/2017	47,4
28/08/2017	43,3
29/08/2017	46,1
30/08/2017	13,2
31/08/2017	5,9

Figure 5.46 : Mailles affectées par les pluies mais absentes de l'événement virtuel (C.Cavalière)

Nous avons alors exploré les facteurs qui pourraient expliciter cette absence de l'événement virtuel : ces pôles sont-ils déjà inactifs en situation normale ? Avons-nous masqué un vocabulaire précis mais minoritaire car non employé par les utilisateurs plus nombreux et plus prolixes des métropoles ? Parmi les 33 aires urbaines totalement invisibles du réseau en période de crise, seules cinq n'affichent aucun tweet géolocalisé émis sur l'ensemble de la période collectée dans le jeu de tweets bruts destiné à extraire le corpus de l'ouragan : il s'agit des aires urbaines de Bishop, Eagle Lake, Hempstead, Deerwood et de Rusk. Les 28 autres aires urbaines cumulent un total de 601 tweets géolocalisés bruts émis entre le 23 août et le 1er septembre 2017. Le nuage de mots (a) de la figure 5.47 représente les cooccurrences lexicales dont la fréquence d'apparition est supérieure à 10, dans le jeu de 601 tweets bruts. Il est accompagné d'un second nuage (b) qui représente les associations lexicales n'apparaissant qu'une seule fois (contrairement aux métropoles ou sièges de comtés, il semble en effet que la mise en évidence d'un vocabulaire de crise dans ces petits pôles urbains passe plutôt par la recherche du vocabulaire marginal, étant donné qu'une majorité de tweets géolocalisés sont liés à des offres d'emploi).



(a). Associations lexicales d'occurrence minimale 10

(b). Associations lexicales d'occurrence 1

Figure 5.47 : Nuages de mots – Vocaculaire des tweets géolocalisés dans les mailles invisibles du réseau

En fait, on peut identifier une minuscule part de tweets relatifs à l'ouragan Harvey (24 tweets, soit 4% des 601 tweets bruts extraits) localisés sur trois villes du Texas seulement : Beeville et Kingsville dans le sud de l'Etat, à l'ouest de Corpus Christi, cumulent 23 tweets rattachés au thème de l'ouragan ; le dernier est localisé à Navasota, au nord-ouest de l'aire métropolitaine de Houston (aucun tweet de crise n'est décelable sur les aires urbaines initialement invisibles de Louisiane). Une lecture de ces tweets souligne rapidement les raisons pour lesquelles ils nous ont échappé :

- ils contiennent un vocabulaire minoritaire différent des associations qui émergent lorsqu'on inclut les tweets des métropoles dans la recherche de lexique. Ici, on pouvait ainsi mettre en évidence le terme de *sandbag*, présent sur le réseau virtuel de Beeville dès le 23 août (la ville est atteinte par l'ouragan le 26 août), de même que le hashtag *#ConvoyOfHope*, témoin de l'intervention de l'association humanitaire éponyme d'aide aux victimes[17].

- certains tweets ne contiennent aucun mot de vocabulaire explicitement rattaché au phénomène mais d'après leur sens, on peut supposer leur lien à l'ouragan : "*Heading to Port Aransas to help my mermaid sista!*" (soit un utilisateur allant assister sa soeur, vraisemblebalement sinistrée des inondations) ; "*Reminder: Church, be sure to check in on anyone you haven't seen or heard from in the past week*" (soit une consigne demandant aux utilisateurs de penser à prendre des nouvelles de leurs proches).

[17]Source :https://eu.news-leader.com/story/news/2017/09/07/max-simultaneous-disasters-stretch-springfield-based-convoy-hope/633260001/ (Consulté pour la dernière fois le 30/04/2019)

Si au final, les villes de Beeville et de Kingsville, dans le contexte particulier de la survenue de cet ouragan, sont présentes numériquement dans les pages Web des médias traditionnels et autres plateformes 2.0[18] (en particulier les vidéos amateures postées sur *YouTube*), qu'en est-il de leur visibilité dans les données officielles et des villes dans lesquelles aucun tweet de crise géolocalisé n'a finalement été détecté ?

- Beeville (et le comté de Bee) est totalement absente de la SEDB ; les tweets font pourtant état de réunions de gestion de crise, de dégâts et attestent du passage de l'association humanitaire précédemment mentionnée. A Kingsville, les informations contenues dans la SEDB soulignent principalement les vents violents survenus le 25 août (mais sans faire de dégât majeur ou de victime) : un tweet émis le même jour témoigne de la même tonalité "*Getting pretty comfortable here in Kingsville, nothing gnarly is happening*".

- Dans les cinq autres aires urbaines n'enregistrant aucun tweet de crise géolocalisé, aucun des cinq noms n'apparaît dans le bulletin d'août 2017 de la SEDB. En consultant la presse en ligne, on ne trouve que mention de Hempstead, dans le comté de Waller, à l'ouest de la métropole de Houston : sans rapport direct avec des événements locaux, une vidéo postée sur *YouTube* informe de la mobilisation d'une force militaire opérationnelle basée à Hempstead vers Houston. En revanche, un tweet inclus dans un autre article en ligne (proposant une revue de tweets relatifs aux impacts de l'ouragan sur les troupeaux[19]) suggère la survenue d'inondations à proximité de Hempstead (figure 5.48).

Figure 5.48 : Emission d'un tweet suggérant une inondation non inventoriée (Source : Successfull Farming)

[18] Quelques titres sont visibles via le moteur de recherche Google, mais l'accès au site reste bloqué en France.
[19] Source : https://www.agriculture.com/news/livestock/12-striking-farmland-livestock-scenes-from-hurricane-harvey (Consulté pour la dernière fois le 11/05/2019)

Que penser alors de ces résultats ? Certains territoires périphériques présentent une activité *tweeting* géolocalisée mince mais existante en temps normal, alors comment expliquer cette faible empreinte, voire cette absence de tweets de crise quand des précipitations ont été enregistrées sur ces mêmes territoires ? Médias numériques consultés et SEDB laisseraient à penser qu'en dépit des pluies, aucun phénomène exceptionnel ne serait survenu (par conséquent, la sensibilité des utilisateurs du réseau ne s'active pas). On peut également penser que la sensibilité à la moindre pluie qui caractérise les habitants des grandes métropoles est caduque dans les pôles des espaces périphériques. Troisième hypothèse possible et plus probable : les tweets de crise existent dans ces milieux mais, contrairement aux habitants des territoires métropolitains, les utilisateurs ne sont ici pas adeptes des fonctionnalités de géolocalisation, à la manière du tweet présenté dans la figure 5.48 qui n'inclut pas de métadonnées spatiales : ici, la mention du nom de lieu se présente sous forme sémantique, incluse dans la chaîne de caractères, et non dans l'activation des fonctionnalités de géolocalisation proposées par Twitter.

Finalement, pour sortir de l'impasse de la question de l'(in)visibilité de territoires affectés par l'ouragan Harvey, nous avons eu recours au jeu de données diffusées par la FEMA, inventoriant les dégâts subis par les bâtiments des propriétés privées au 29 août 2017. A cette date, un total de 233 624 bâtiments endommagés sont enregistrés par l'organisation fédérale. Si aucun bâtiment n'est inventorié sur nos aires urbaines invisibles de l'événement virtuel, ils sont en revanche 5,6% localisés dans des mailles où aucun tweet géolocalisé de crise n'a été capturé. Comme nous l'avons fait pour les aires urbaines précédemment, nous avons alors souhaité vérifier l'existence éventuelle de tweets de crise non répertoriés comme tels dans notre base de données, en nous focalisant sur des espaces au profil différent mais affichant au moins 12 bâtiments ayant subi des dégâts majeurs ou détruits (valeur médiane à 11). Nous avons alors inventorié deux types d'espaces présentant ce profil : des espaces situés dans les marges des aires urbaines ou métropolitaines ainsi que des espaces ruraux ne croisant aucune aire urbaine. Pour l'exploration de ces mailles particulières, nous avons retenu deux mailles de marges urbaines (Houston et Beaumont) ainsi que deux mailles d'espaces périphériques aux territoires urbains et métropolitains. La carte présentée dans la figure 5.49 indique la localisation des espaces tests et le tableau 5.5 expose les résultats de la recherche d'éventuels nouveaux tweets de crise.

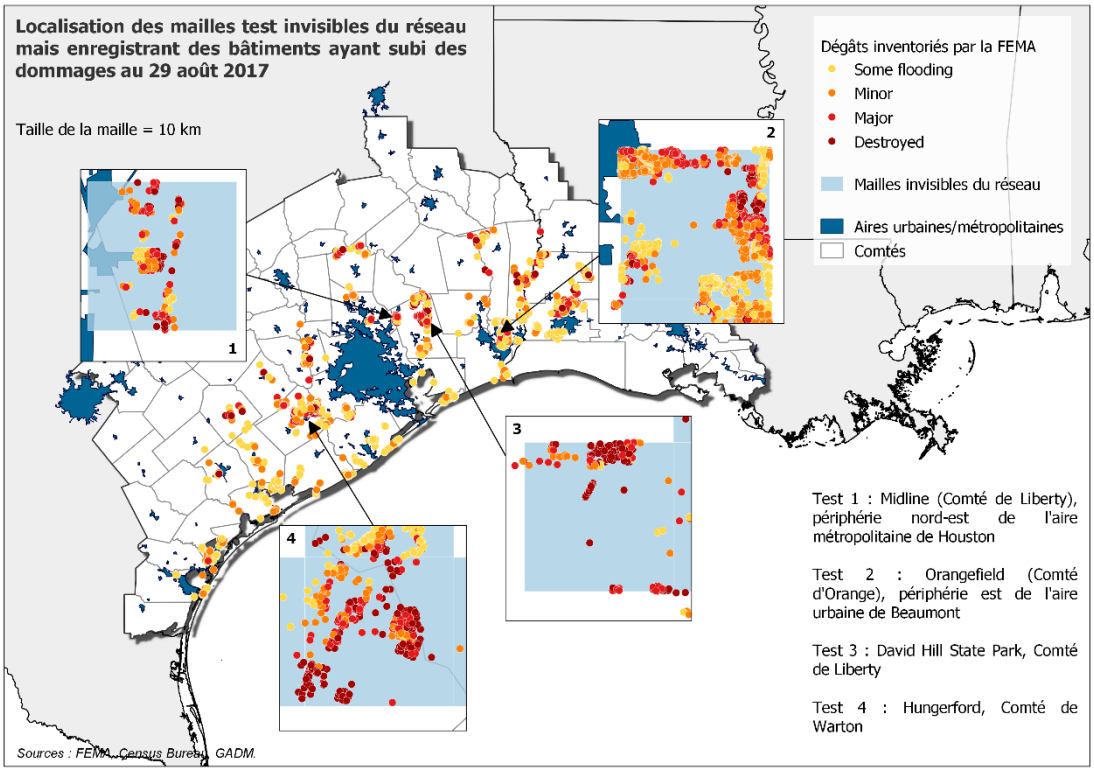


Figure 5.49 : Localisation des mailles test pour la recherche d'éventuels nouveaux tweets de crise géolocalisés (C.Cavalière)

Tableau 5.5 : Exploration des tweets géolocalisés inclus ou voisins des mailles invisibles du réseau

Maille test	Dommages	Nombre de tweets non triés inclus dans la maille	Distance du tweet le plus proche d'une propriété affectée (m)	Thèmes des tweets non triés inclus dans la maille	Thèmes des tweets non triés situés à moins de 5 km de la maille
Test 1	Some flooding – 39 Minor – 71 Major – 81 Destoyed – 50	13	1 820	- Offres d'emplois - Routes inondées	- Bulletins météorologiques - Alertes aux inondations (tweets rédigés en espagnol)
Test 2	Some flooding – 379 Minor – 612 Major – 261 Destoyed – 61	0	199 (hors maille test 2)		- Offres d'emplois - Maisons inondées/endommagées - Individus sains et saufs
Test 3	Some flooding – 11 Minor – 36 Major – 48 Destoyed – 109	0	3 419 (hors maille test 3)		- Activité liée à un média : aide aux populations affectées ; formulaire en ligne de déclaration des dégâts - Alerte tornade - Prières
Test 4	Some flooding – 18 Minor – 111 Major – 106 Destoyed – 150	0	3 705 (hors maille test 4)		- Offres d'emplois - Alerte tornade - Alerte inondation et crues éclair

L'existence de mailles et, par là même, de territoires d'échelle locale ayant subi des dégâts et dont la visibilité numérique par la géolocalisation se présente comme un événement marginal voire nul, reste avérée dans le cas d'un phénomène rare. Doit-on alors considérer ces territoires comme inexistants sur les réseaux numériques ? Une nouvelle fois, il se peut

que les habitants concernés soient actifs sur le réseau sans pour autant avoir adopté une pratique systématique de géolocalisation par les fonctionnalités Twitter. Cette invisibilité numérique pourrait également s'expliquer par l'évacuation précoce de certains territoires (bien que les tweets ne le mentionnent pas et qu'une recherche effectuée sur le Web indique que les territoires explorés n'avaient pas fait l'objet d'obligations d'évacuations préventives aux inondations[20]). Cela étant, l'exemple de la géolocalisation des tweets de crise en réponse à un phénomène rare laisse en suspens les interrogations énoncées précédemment (cf. bilan de la section 5.2.1) quant à la signification géographique de la composante spatiale des tweets à coordonnées GPS. Face aux faits constatés dans les cartes, la première loi de a-t-elle un sens dans les tweets de crise géolocalisés ? En effet, si l'on reprend l'exemple des quatre mailles explorées ci-avant, on constate des décalages entre la réalité du terrain (bâtiments des propriétés ayant subi des dégâts d'intensité variée) et le contenu des tweets inclus ou situés à moins de cinq kilomètres de la maille : d'une part, le test 1 (Midline) indique des tweets mentionnant des routes inondées mais ne laisse transparaître aucune information relative aux dégâts des propriétés. D'autre part, 57% des tweets géolocalisés à moins de cinq kilomètres de la maille d'Orangefield (test 2) représentent des offres d'emplois diffusées en continu.

5.2.2.3. Les temporalités virtuelles du phénomène extrême peu fréquent

Si le sens de la composante spatiale des tweets de crise géolocalisés reste pour l'instant hasardeux, que peut-on dire de leur composante temporelle dans la survenue d'un phénomène extrême rare ? Dans le cas précédent des phénomènes récurrents, nous avons défini les périodes d'événement virtuel à partir de paramètres statistiques calculés pour chaque mois. Ici, la réponse virtuelle à un phénomène tel qu'un ouragan peut indubitablement être qualifiée d'événement virtuel hors-normes : en effet, au printemps 2016, la journée du 18 avril qui enregistrait le plus fort pic d'émission de tweets de crise géolocalisés culminait à un total de 1 175 tweets (toutes catégories confondues) ; pendant le passage de l'ouragan, la journée qui a drainé les flux les plus conséquents s'avère être le 27 août 2017, avec un total de 3 574 tweets de crise géolocalisés (soit un flux multiplié par trois pour la journée la plus active lors du phénomène extrême rare, par rapport à la journée la plus active des phénomènes récurrents).

Pics d'émission de tweets de crise géolocalisés.

Comme les temporalités de ce phénomène sont précisément documentées sur le Web, notre objectif ne consistait pas à délimiter temporellement l'événement virtuel consécutif à l'ouragan mais à repérer, par les pics d'émission de chaque journée, les périodes pendant lesquelles les flux de tweets sont anormalement élevés, en comparant tweets d'acteurs officiels et tweets catégorisés comme *autres*. Nous pouvons alors mettre en évidence les périodes d'activité *tweeting* de crise significative sur le réseau, localiser les espaces concernés

[20] Source : https://en.wikipedia.org/wiki/Hurricane_Harvey#Texas (Consulté pour la dernière fois le 02/05/2019)

et identifier des territoires à explorer lexicalement tout en vérifiant la disponibilité d'informations dans le bulletin SEDB adéquat. La figure 5.50 ci-dessous affiche les quantités de tweets émis par heure entre le 23 et le 31 août 2017.

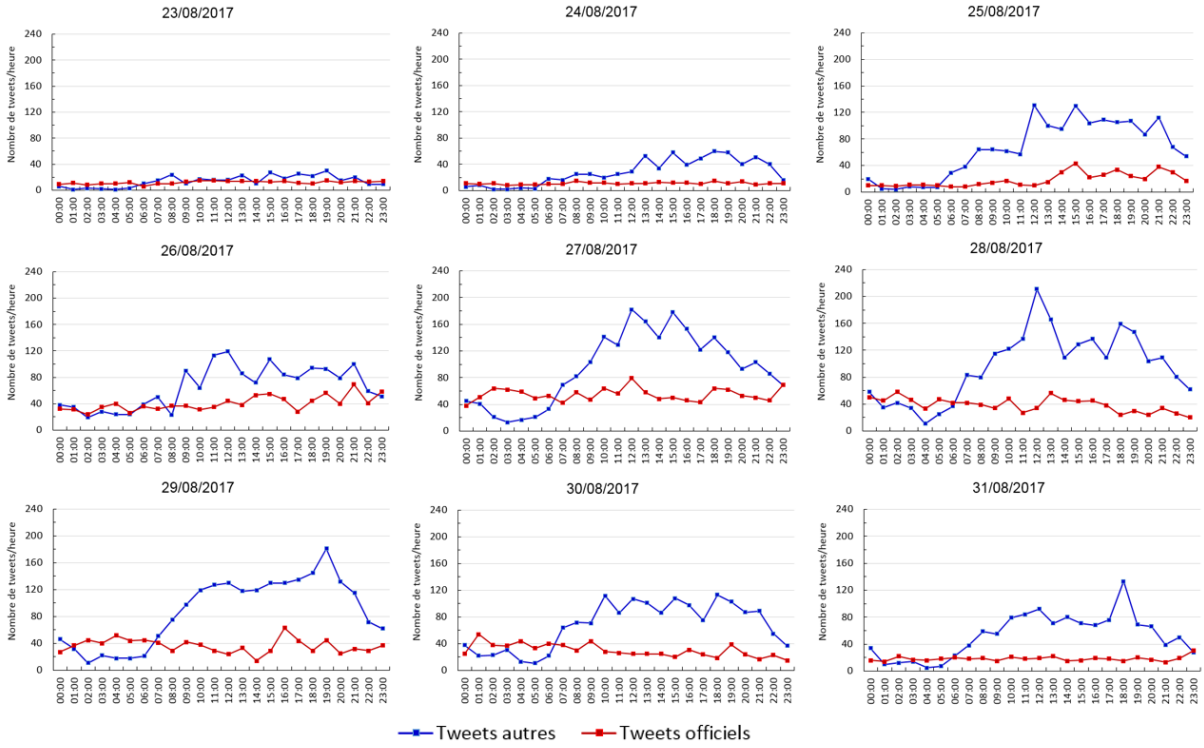


Figure 5.50 : Flux horaires de tweets de crise géolocalisés émis entre le 23 et le 31 août 2017

Sans surprise, on identifie des pics d'émission quotidiens (en revanche, si c'est le 27 août qui draine la plus forte quantité de tweets sur la journée, le plus fort pic horaire est marqué le 28 août entre midi et 13h). Malgré tout, un fait nous a interpellés dans ces graphiques : la nette régularité des créneaux horaires enregistrant les périodes d'activité *tweeting* poussée, qui nous fait davantage penser à la distribution temporelle de l'activité *tweeting* routinière. En effet, si l'émission de tweets par les acteurs officiels du réseau paraît régulière dans la journée, avec des quantités de tweets émis plus fortes pendant que l'ouragan frappe les terres, les flux de tweets catégorisés comme *autres* suivent la logique d'activité quotidienne de l'individu : amorce de l'activité le matin à partir de 7h, premiers pics à la mi-journée, nouveaux pics pendant la soirée et déclin de l'activité après 22h. La figure 5.51 ci-après indique alors les périodes pendant lesquelles on va considérer qu'est survenu un événement virtuel, pour les deux catégories de tweets de crise : afin d'atténuer l'effet pic de tweets, ces créneaux correspondent aux heures pendant lesquelles le flux de tweets enregistrés est supérieur au flux médian calculé pour chaque jour.

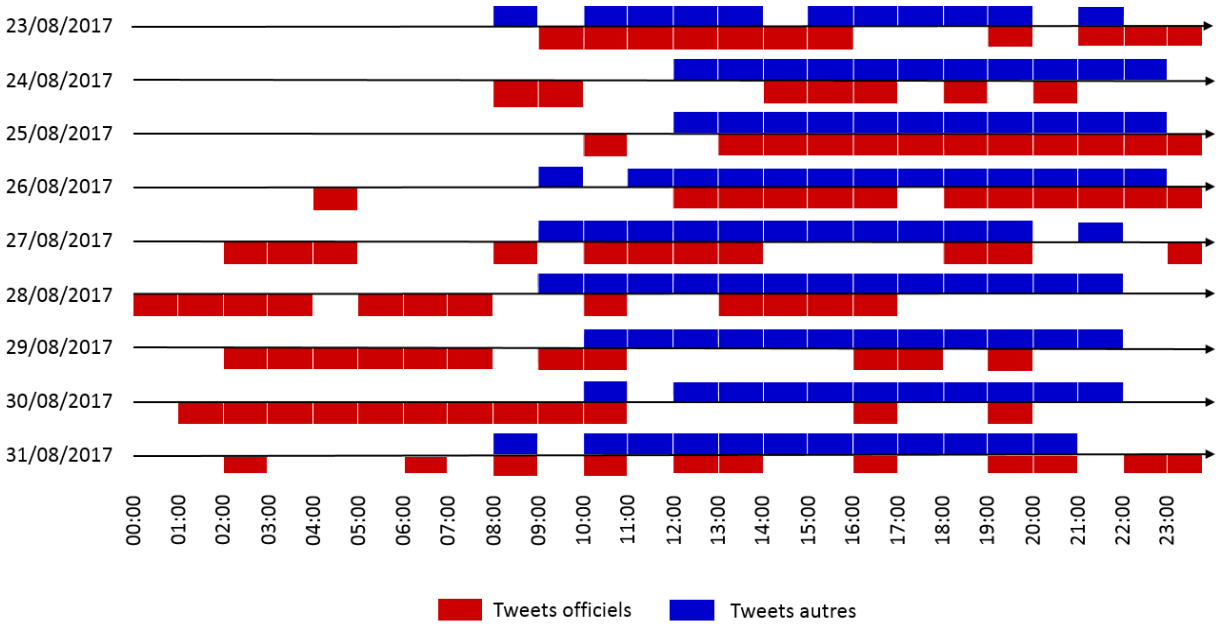


Figure 5.51 : Périodes enregistrant un événement virtuel du 23 au 31 août 2017

Une logique temporelle propre à la crise est-elle alors perceptible à l'échelle globale de l'ouragan ? Par cet exemple, il s'avère périlleux de trouver un sens particulier aux dynamiques et pics d'émission constatés. En effet, si les événements virtuels émanant des émissions de tweets officiels apparaissent par périodes distinctes, les temporalités de l'événement virtuel généré par les autres tweets ne constituent pas une réponse symétrique : cet événement virtuel *autre* se révèle régulier et quasi-continu, généralement entre 10h et 21h. De la même manière, si ces périodes de forte activité coïncident avec des événements virtuels officiels enregistrés pendant des périodes identiques, force est de constater que les événements virtuels officiels survenus dans les premières heures de la matinée (et notamment du 27 au 30 août) n'engrangent aucune réponse majeure dans les tweets catégorisés *autres*.

Doit-on alors focaliser l'attention sur les quelques événements virtuels *autres* isolés en journée afin d'identifier un événement particulier dans l'événement virtuel global ? Ou faut-il considérer les paramètres statistiques comme non significatifs et appréhender l'événement virtuel dans son ensemble sans chercher à souligner des particularités ? Pour mettre en évidence l'éventuelle existence de micro-événements dans l'événement virtuel global, nous marquons ces périodes isolées ; leur exploration lexicale permettra *a priori* de déterminer si elles contiennent un sens spécifique :

- le 24 août, de 8h à 9h (absence d'événement virtuel officiel), puis de 21h à 22h ;
- le 26 août, de 9h à 10h (absence d'événement virtuel officiel) ;
- le 27 août, de 21h à 22h (absence d'événement virtuel officiel) ;
- le 30 août, de 10h à 11h ;
- le 31 août, de 8h à 9h ;

Persistance de l'activité tweeting au cours du temps.

Dans la même perspective de recherche d'espaces propices à l'exploration lexicale des tweets de crise géolocalisés, nous avons représenté les mailles de 10 km évoquées dans le paragraphe précédent (cf. 5.2.2.2), non pas en fonction des quantités de tweets émis par maille (en raison des biais statistiques introduits par les métropoles) mais en fonction du nombre de jours pendant lesquels chaque maille a été active (figure 5.52).

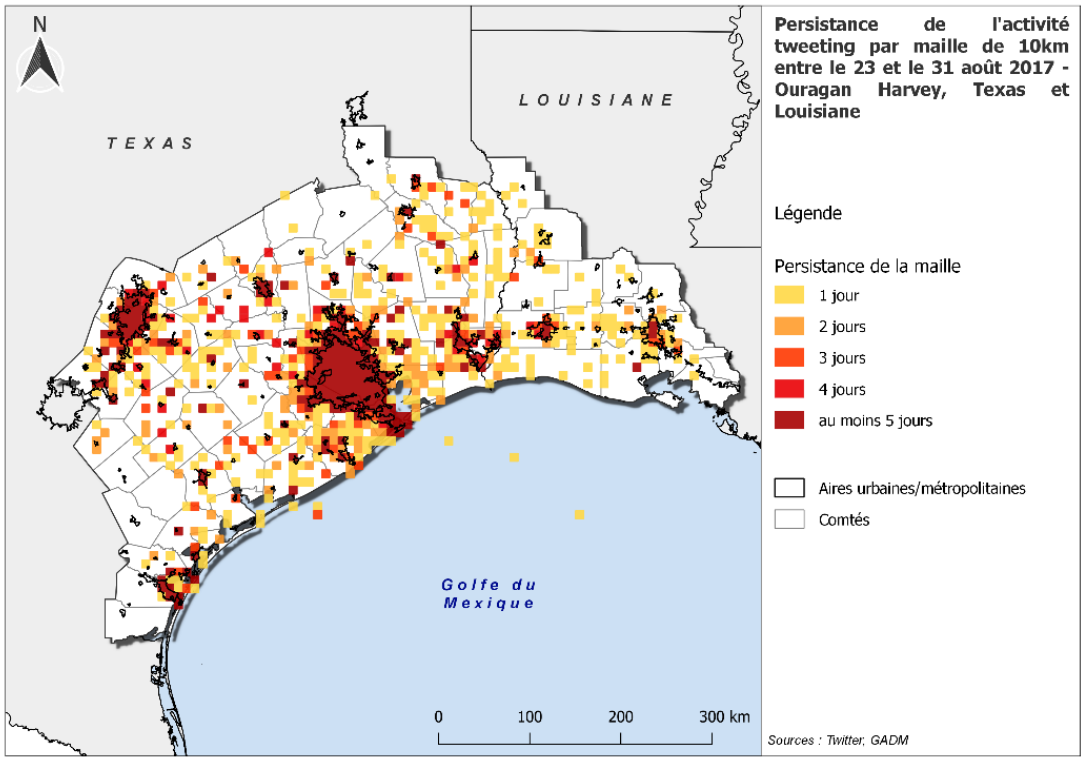


Figure 5.52 : Persistance de l'activité tweeting de crise par maille entre le 23 et le 31 août 2017 (C.Cavalière)

On retrouve sans surprise le comportement distinct aux aires urbaines et métropolitaines qui cumulent 90% des mailles actives pendant au moins cinq jours. Néanmoins, des espaces actifs persistants apparaissent ponctuellement en dehors de toute aire urbaine : le tableau 5.6 ci-après permet alors de dégager des tendances qui nuancent les constats habituels. Une maille active pendant plus de cinq jours ne cumule pas nécessairement les quantités de tweets les plus importantes : l'une des mailles actives pendant neuf jours ne contient que 24 tweets de crise géolocalisés alors qu'une autre maille perçue pendant seulement deux jours en cumule 64 (les autres situations inattendues figurent en rouge dans le tableau et concernent des mailles à l'activité limitée dans le temps mais cumulant un nombre conséquent de tweets de crise géolocalisés).

Tableau 5.6 : Paramètres statistiques des mailles en fonction de la persistance de leur activité

Nombre de jours d'activité de la maille	Nombre de mailles	Fréquence en %	Nombre minimal de tweets	Nombre maximal de tweets	Nombre de tweets médian	Localisation de la maille improbable
9	67	2,7	24	4489	94	Centre de l'aire métropolitaine d'Austin
8	34	1,4	12	168	39	
7	27	1,1	7	123	33	
6	24	0,9	6	55	18,5	
5	36	1,5	5	112	13	Comté de Liberty, marge nord-ouest de l'aire métropolitaine de Houston
4	40	1,6	4	36	7,5	
3	64	2,6	3	96	5	Au large des côtes du Golfe du Mexique
2	122	4,9	2	64	3	Aire urbaine de Palacios, sud-ouest de Houston
1	324	13,1	1	11	1	Crosby (Est de Houston)
0	1746	70,4				

En dépit de ces valeurs inhabituelles, le tableau 5.6 permet tout de même d'observer une baisse constante du nombre médian de tweets de crise géolocalisés en fonction du nombre de jours pendant lesquels chaque maille a été active. Par ailleurs, une minorité d'individus témoigne de nouveau de l'empreinte virtuelle la plus conséquente : seules 6,1% des mailles restent actives plus de cinq jours. Comme dans le paragraphe 5.2.1.2, nous souhaitons savoir si les comportements temporels et la (non) persistance de l'activité tweeting de crise enregistrée dans les mailles, sont liés au profil du territoire. Les mailles sont ainsi regroupées en trois catégories : mailles des aires métropolitaines de Houston, Austin et San Antonio, mailles des aires urbaines et enfin, mailles localisées en dehors de tout milieu urbain. Dans chaque groupe, les mailles hors-normes sont retirées (en effet, en ce qui concerne les mailles situées en dehors de tout milieu urbain, nous ne savons pas encore s'il s'agit d'une activité liée à des individus ou à des comptes automates non détectés). La figure 5.53 représente ainsi, pour chaque journée et chaque catégorie de mailles, la proportion de tweets de crise géolocalisés émis entre le 23 et le 31 août 2017 (les flux de tweets variant d'un milieu à un autre, ils sont exprimés en pourcentage).

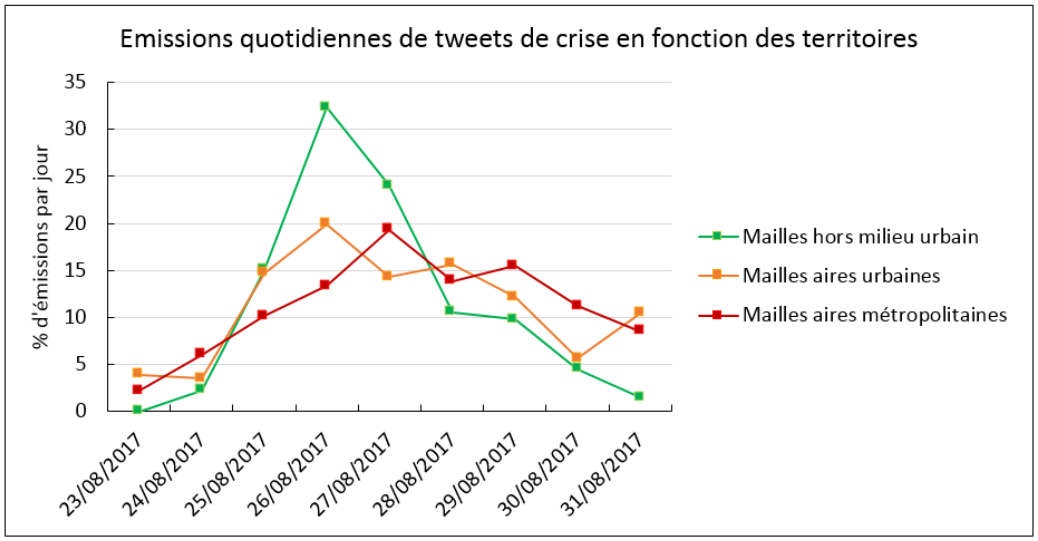


Figure 5.53 : Proportions quotidiennes de tweets de crise émis chaque jour par maille et par milieu

Le graphique met alors en évidence l'alternance entre des périodes de croissance du flux, de pics et de décroissance, mais la persistance de l'activité tweeting de crise apparaît comme un comportement urbain : si l'allure des courbes représentant l'activité temporelle des mailles situées dans les aires urbaines et métropolitaines témoigne de ces différents cycles, elle indique également une activité plus homogène et plus diffuse que la tendance observée en dehors de tout milieu urbain. La courbe verte, qui correspond à ce dernier profil, témoigne en effet d'une concentration des émissions en un jour précis, le 26 août 2017. Par ailleurs, si la période de décroissance consécutive au pic d'émissions s'avère moins brutale que la période de croissance qui la précède, elle reste plus rapide que dans les milieux urbains et métropolitains.

5.2.2.4. Exploration des lieux de réactivité virtuelle consécutive au passage de l'ouragan Harvey

Ce paragraphe présente le résultat des analyses lexicales focalisées sur les lieux d'intérêt identifiés d'une part, via l'exploration temporelle des tweets de crise géolocalisés, et d'autre part, via la consultation de la presse locale en ligne (qui permet notamment de cibler rapidement les territoires ayant subi les dégâts les plus importants). L'objectif s'avère double :

- dans un premier temps, il s'agit de rechercher des facteurs explicatifs des faits constatés dans le paragraphe 5.2.2.3 (relatifs aux temporalités des événements virtuels et à la persistance de l'activité des mailles) en identifiant et explorant les lieux à l'origine de cette activité virtuelle.

- Dans un second temps, il s'agit de mesurer la capacité qu'ont (ou pas) les tweets de crise liés au phénomène à retracer son historique à l'échelle locale, les comportements des

individus et les effets du phénomène sur les populations, en dehors de tout milieu métropolitain (Houston faisant l'objet d'une étude à part entière dans le chapitre suivant).

Exploration des lieux d'intérêt ciblés par la dynamique temporelle de l'activité virtuelle.

Dans un premier temps, nous focalisons notre attention sur les mailles remarquables mises en évidence dans le tableau 5.6 : pour rappel, il s'agit des mailles qui s'activent de manière ponctuelle (de 1 à 5 jours) et qui cumulent un nombre important de tweets[21], mais également d'une maille qui reste active pendant neuf jours et qui pourtant ne contient qu'un nombre très restreint de tweets de crise géolocalisés (24 tweets). Le tableau 5.7 ci-dessous récapitule la localisation des mailles en question et présente le profil du lieu concerné ainsi que les thèmes identifiés dans l'exploration lexicale des tweets contenus dans chaque maille :

Tableau 5.7 : Exploration des lieux et du lexique associés aux mailles au profil atypique

Profil de la maille	Profil du lieu	Informations contenues dans la sémantique des tweets
Austin – 9 jours d'activité pour 24 tweets	Quartier résidentiel aisé de West Lake Hills	- Absence d'informations relatives à des situations urgentes ou des perturbations graves - Organisation d'une *hurricane party* - Collecte de dons pour les sinistrés
Comté de Liberty, Huffman – 5 jours d'activité pour 112 tweets	Quartier résidentiel puis espace forestier et agricole à l'est de la maille	Un seul compte automatique émetteur de bulletins météorologiques
Eaux territoriales du Golfe du Mexique – 3 jours d'activité pour 96 tweets	Océan	*Idem*
Palacios – 2 jours d'activité pour 64 tweets	Littoral de la baie de Matagorda	*Idem*
Crosby (Est de Houston) – 1 jour d'activité pour 11 tweets	Résidentiel peu dense et ranchs	Alerte ponctuelle faisant état de l'activation du plan d'urgence chimique le 31 août

La maille incluse dans le milieu métropolitain et qui enregistre un effectif de tweets assez faible contient des informations concernant les individus et leurs activités dans leur environnement, qui s'avèrent assez hétéroclites, mais dans lesquelles on peut identifier deux thèmes de tweeting (notons cependant qu'il n'y a aucun signalement, pendant l'ensemble de la période, d'une perturbation physique ou d'une situation d'urgence locale) : d'une part, le partage d'informations et de liens destinés à organiser la collecte de dons pour les sinistrés de l'ouragan (à l'échelle globale du phénomène et non locale), et d'autre part, l'organisation

[21] Par rapport aux valeurs des paramètres statistiques généraux des mailles actives le même nombre de jours.

d'une *hurricane party*[22] locale. En dehors du milieu métropolitain, on peut détecter le signal d'une perturbation s'exerçant dans des temporalités précises mais celui-ci provient d'automates ou d'acteurs officiels et n'engendre pas de réponse humaine (en ce qui concerne les tweets géolocalisés émis par smartphone). Par exemple, le plan d'urgence chimique activé au 31 août 2017 à Crosby fait référence à une explosion suivie d'un échappement de fumées toxiques depuis une usine (suite à une panne d'un système de réfrigération, consécutive aux inondations), ayant entraîné l'évacuation d'environ deux cents individus pendant plus d'une semaine[23] ; en dépit de sa gravité, cet événement local reste invisible dans notre jeu de tweets de crise, du point de vue des populations affectées[24].

Dans un second temps, nous avions noté, à partir de la figure 5.51, de courtes périodes pendant lesquelles le réseau s'agitait en dehors des événements virtuels officiels. Nous avons exploré trois de ces périodes : le 24/08 entre 8h et 9h (soit avant l'arrivée de l'ouragan), le 27/08 entre 21h et 22h (soit pendant que l'ouragan frappe le sud-est du Texas) et enfin, le 31/08 entre 8h et 9h (soit après le passage de l'ouragan sur le Texas et le sud de la Louisiane, mais pendant son passage sur le nord de ce même Etat). Pour ce nouveau test, nous nous focalisons sur les dynamiques d'apparition et de disparition de l'activité tweeting dans une entité donnée (en l'occurrence, il s'agit des mailles de dix kilomètres de côté) : ces dynamiques sont considérées comme suit : (1) chaque créneau horaire donné ci-avant constitue la période de référence pendant laquelle on enregistre une agitation virtuelle ; (2) pour l'analyse, on identifie les mailles qui s'avèrent actives pendant cette période de référence mais qui étaient invisibles dans le créneau horaire précédent et qui disparaissent dans le créneau horaire suivant. En conséquence, l'activité tweeting de crise n'est pas considérée par un critère quantitatif mais simplement par la présence d'un tweet susceptible d'envoyer un signal à un moment précis. Les cartes suivantes (figures 5.54 à 5.56) présentent la localisation des mailles ponctuelles correspondant aux critères énoncés. Les tweets inclus dans ces mailles ont été parcourus un à un et classés selon des thèmes (définis au fur et à mesure de la lecture des tweets). Dans les cartes suivantes, la sémantique des tweets n'est ainsi pas représentée par le vocabulaire sous forme de nuages de mots, mais pas des icônes représentant chaque thème. Plus la taille de l'icône est grande, plus le thème est présent dans les tweets contenus dans la maille en question.

La période analysée le 24 août (figure 5.54) indique des témoignages ponctuels d'anticipation de l'arrivée du phénomène ; il s'agit principalement d'utilisateurs manifestant leur intention de faire leurs provisions avant d'être confinés et vraisemblablement d'une association qui recherche déjà des bénévoles. Ces cellules d'activité semblent alors davantage

[22] Dans le sud-est des Etats-Unis, le terme de *hurricane party* désigne un événement rassemblant en un lieu précis et pendant le passage de l'ouragan, des individus n'ayant pas évacué et qui apportent des vivres, boissons (alcool compris) et matériels de premiers soins. Source : https://en.wikipedia.org/wiki/Hurricane_party

[23] Source : https://www.nytimes.com/2018/08/03/business/arkema-chemical-plant-explosion-texas.html (Consulté pour la dernière fois le 29/10/2019).

[24] Le constat est identique si l'on se base sur le jeu de tweets bruts : l'ensemble des tweets émis depuis Crosby le 31 août 2017 ne proviennent pas d'individus ayant vécu l'accident chimique.

liées au créneau horaire (8h-9h) qu'à la survenue d'événements locaux rapportés par les utilisateurs actifs.

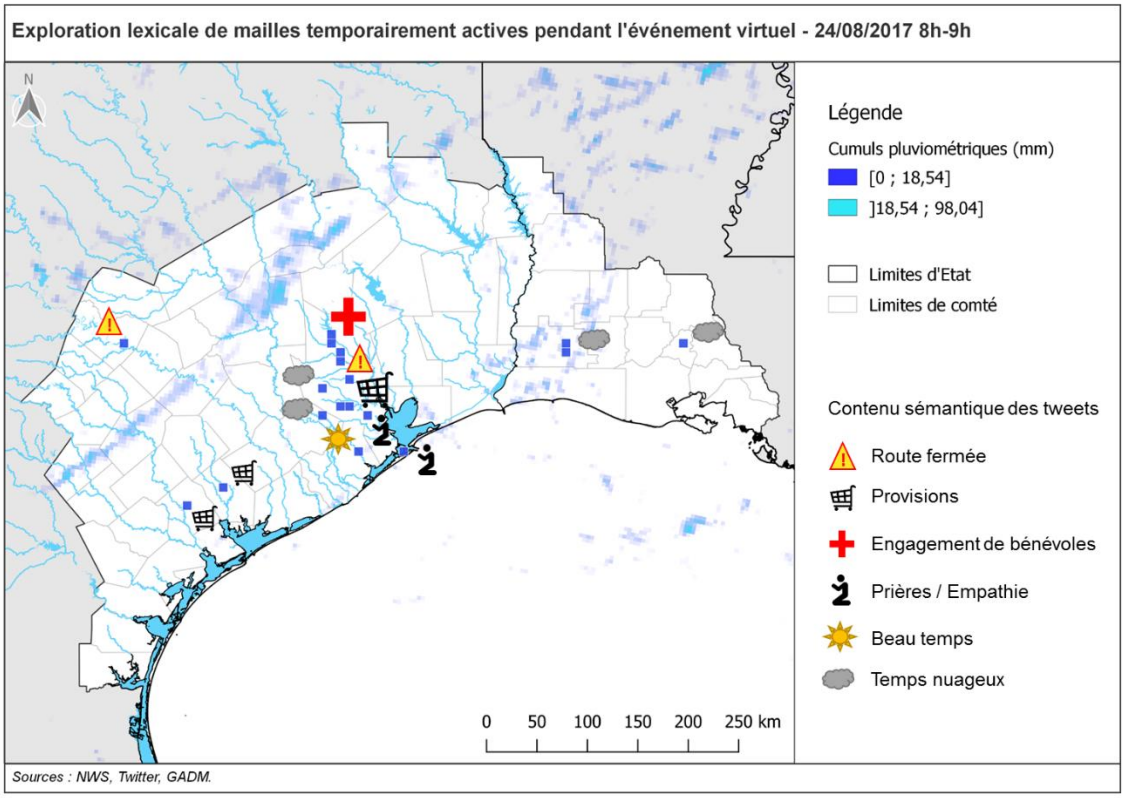


Figure 5.54 : Exploration lexicale des mailles ponctuelles détectées le 24 août 2017 entre 8h et 9h (C.Cavalière)

Dans la soirée du 27 août (figure 5.55), on peut détecter le signal de phénomènes et événements consécutifs d'échelle locale : le thème inondation apparaît spontanément dans une maille d'une marge de l'aire métropolitaine de Houston (le texte du tweet traduit d'ailleurs l'intensité de l'inondation "*water up to knee*") ; dans deux mailles, la photographie apparaît comme témoin d'un événement d'échelle fine dans des lieux où l'activité tweeting reste marginale par rapport aux quantités de tweets émis dans le centre des métropoles. Dans le sud, on détecte sans doute le signal d'un événement local plus grave : dans le créneau horaire étudié, quatre tweets mentionnent l'évacuation d'un centre de soins, les refuges ouverts ainsi qu'un décès.

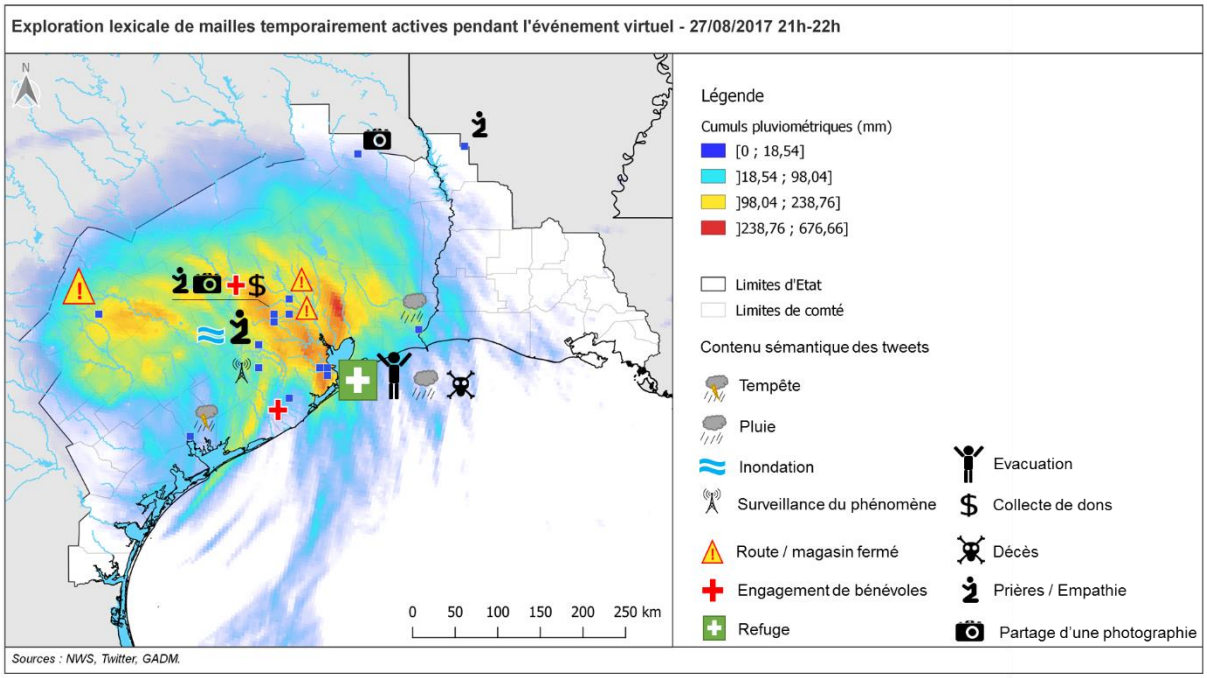


Figure 5.56 : Exploration lexicale des mailles ponctuelles détectées le 27 août 2017 entre 21h et 22h (C.Cavalière)

Le 31 août, on identifie de nouveau des mailles témoignant d'une activité spontanée liée à un événement local (figure 5.56) : dans un premier temps, on retrouve l'accident de l'usine chimique de Crosby et des partages de photographies des inondations sur le quartier de Kingwood, dans la partie nord-est de la métropole de Houston.

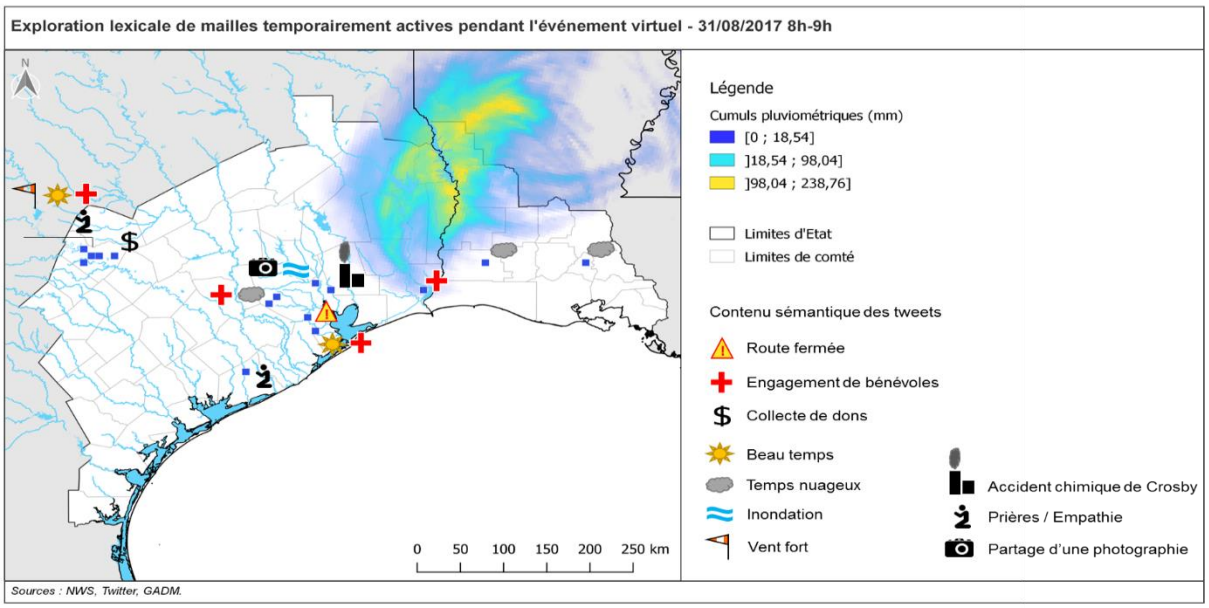


Figure 5.55 : Exploration lexicale des mailles ponctuelles détectées le 31 août 2017 entre 8h et 9h (C.Cavalière)

Ce même jour, l'entraide est le thème le plus partagé sur le réseau (engagement bénévole et collecte de dons) mais nous ne considérons pas les mailles contenant les tweets en question comme signaux d'événements locaux (les messages qui appellent à la mobilisation des personnes et à la collecte de dons sont généraux et ne contiennent aucune mention de lieu précis).

Au final, les dynamiques d'apparition/disparition de l'activité tweeting de crise permettent de détecter un signal propre à un événement ou à un phénomène particulier d'échelle locale (inondations, évacuations, ou encore l'accident chimique de Crosby). Ces mailles à l'activité sporadique ont la propriété suivante : elles ne se situent pas dans les hauts-lieux de l'activité tweeting de crise jusqu'alors identifiés (comme le centre de la métropole de Houston), mais dans les marges de cette métropole ainsi que dans les pôles urbains secondaires. Cette activité temporaire est-elle alors significative ou s'agit-il de comportements isolés ?

Nous avons déjà vu que l'accident chimique de Crosby n'avait obtenu aucune réponse autre que les tweets signalant l'activation du plan d'urgence chimique. Nous répétons alors ce test de visibilité des événements et phénomènes ponctuels avec deux nouveaux exemples : les évacuations mentionnées sur le littoral de la baie de Galveston le 27 août entre 21 et 22 heures (cf. figure 5.55) ainsi que les photographies des inondations dans le quartier de Kingwood le 31 août entre 8 et 9 heures (cf. figure 5.56). La recherche est étendue à l'ensemble des tweets bruts émis dans la (ou les) maille(s) en question pendant le créneau horaire considéré (au cas où le filtrage des tweets ait écarté de l'information utile). Si l'on se focalise, dans un premier temps, sur le quartier de Kingwood, l'activité de la maille s'avère en fait déclenchée par le retour des habitants dans leur quartier et la constatation de l'ampleur des inondations et des dégâts, comme l'indiquent ces deux tweets : "*This is amazing that water is literally across the street from us*", "*Kingwood is still a mess… people waiting to get in to see their homes by boat!*". Dans les villes de League City et de Dickinson, on trouve d'autres tweets témoins, initialement inclus dans le jeu de tweets de crise, de la survenue de pluies intenses et d'inondations : "*In case anybody in the homestown needs it. Parent's house flooded*", "*It's official the water is in, I know it's all materialistic stuff…*" (ces deux tweets cités sont en revanche émis dans l'heure qui suit l'annonce des évacuations).

Dans les deux lieux ciblés, qui correspondent à des territoires dont l'activité virtuelle géolocalisée de crise reste quantitativement faible, la dynamique d'apparition/disparition d'une maille se trouve donc liée à la survenue de phénomènes et d'événements locaux (inondations et retour des habitants). Ceux-ci sont par ailleurs documentés par des utilisateurs humains (et non associés à des comptes automatiques). De même, on pouvait constater que le contenu lexical des tweets émis dans une même cellule d'activité (et dans des temporalités proches) se recoupait. Dans ces deux cas, la première loi de Tobler se vérifierait, cette fois, en ce qui concerne l'événement virtuel ; qu'en est-il alors par rapport au réel ? La carte ci-après

(figure 5.57) affiche la localisation des mailles spontanées, par rapport aux données de la FEMA relatives aux propriétés endommagées.

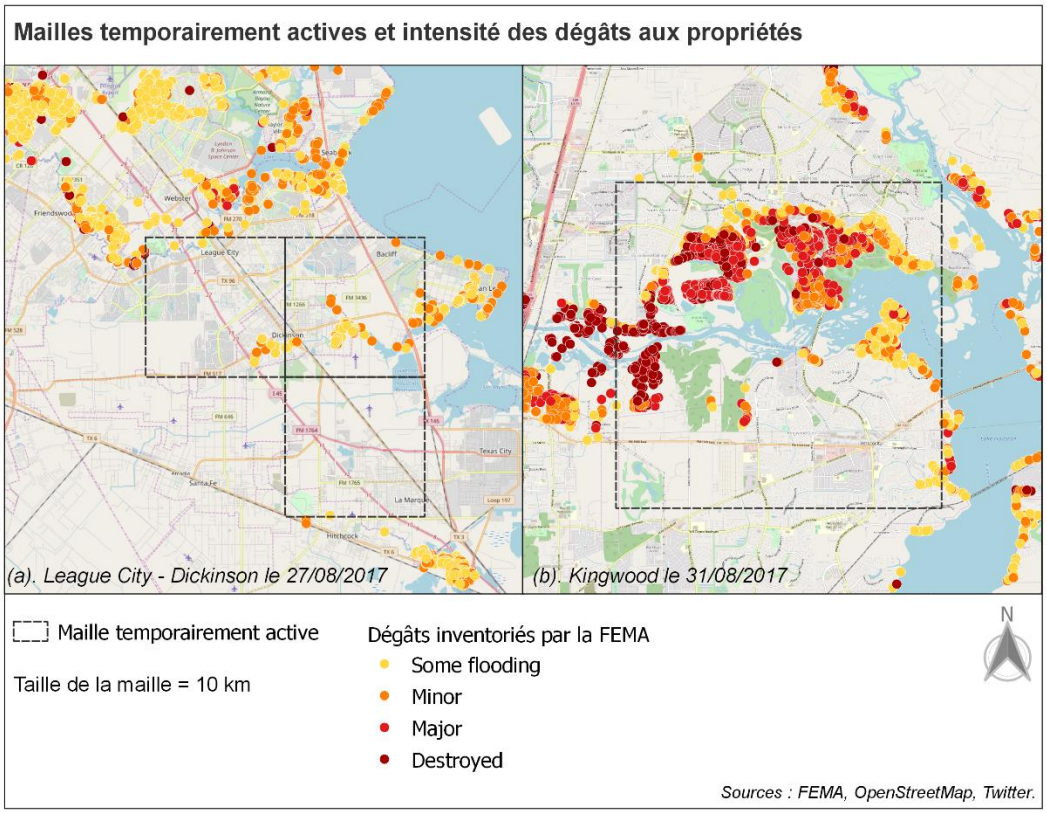


Figure 5.57 : Localisation des mailles temporaires et intensité des dégâts (FEMA), C.Cavalière

Dans le cas de Kingwood (b), le discours des tweets concorde avec l'intensité des dégâts relevés (pour rappel, il était question de rues inondées et d'habitants circulant en bateau pour atteindre leur domicile). En revanche, dans le sud-ouest de la baie de Galveston (a), les cellules ne se sont pas activées dans les territoires les plus affectés par les inondations (visibles plus au nord sur la carte [a]). Pour terminer, si l'on doit comparer les deux dynamiques de ces mailles, on se retrouve donc face à :

- l'absence d'un événement virtuel à Kingwood pendant la survenue des intempéries puisque celui-ci se déclenche en fait au retour des habitants.
- Dans le sud, l'activité virtuelle est enregistrée pendant la survenue du phénomène et des événements consécutifs mais se tarit après.

Dans les territoires situés en dehors des pôles de l'activité virtuelle, peut-on alors lier l'intensité réelle d'un phénomène à ces périodes d'activité et de silence (ce qui corcorderait avec l'approche adoptée par (Samuels *et al.*, 2018) ? Et en effet, le vocabulaire des tweets des deux lieux explorés adopte une tonalité différente (notons qu'on se trouve dans deux milieux humides) : dans la baie de Galveston, nous avons mis en évidence l'évacuation d'un centre de soins, ainsi que l'existence de maisons inondées dans l'heure qui suit cette évacuation. Les

tweets n'offrent guère de précision supplémentaire quant à l'ampleur de ces inondations, ce qui n'est pas le cas de Kingwood : dans ce quartier, on sait que des rues sont inondées et implicitement, que la circulation s'effectue par bateau pour atteindre les maisons inondées (il n'y a pas de tels témoins au sud-ouest de la baie de Galveston).

Consultation de la presse en ligne : territoires ayant subi le plus de dégâts matériels

Les lieux d'intérêt analysés ci-avant ont été définis en fonction de critères de visibilité et de dynamique de l'événement virtuel ; les tendances dégagées de l'activité virtuelle ont ensuite été recoupées avec les données de réalité-terrain. Pour ce nouveau test, nous adoptons la démarche en sens inverse : les lieux d'intérêt sont définis par les informations de réalité-terrain (en l'occurrence, les articles de presse en ligne publiés au lendemain de l'ouragan) et l'événement virtuel est exploré à partir des lieux identifiés au préalable.

Le 25 août, l'ouragan de catégorie 4 gagne les côtes texanes par la ville de Rockport (figure 5.58), située dans le comté d'Aransas (au nord de Corpus Christi), qui figure par conséquent parmi les territoires côtiers les plus affectés[25]. Ce comté avait fait l'objet d'évacuations préventives le jour même[26]. Concernant la ville de Rockport, le récit de la SEDB mentionne l'heure précise d'arrivée de l'ouragan sur les côtes, à 22h ; il indique par ailleurs des dégâts majeurs aux infrastructures, consécutifs à l'onde de tempête et le chiffre de deux victimes. Pour obtenir plus de précisions quant à ces dégâts, il faut se référer aux informations globales relatives au comté d'Aransas : la SEDB inventorie alors 1 500 maisons totalement détruites, 3 800 maisons ayant subi des dégâts majeurs, 5 350 maisons ayant subi des dégâts mineurs. Dans la description associée aux types de dégâts, on trouve des maisons effondrées, des étages arrachés aux maisons en bois, des arbres déracinés ou décapités, une panne d'électricité ayant duré deux à trois semaines ainsi que des troupeaux décimés.

[25] Source : https://www.nytimes.com/interactive/2017/09/01/us/hurricane-harvey-damage-texas-cities-towns.html (Consulté pour la dernière fois le 13/05/2019)

[26] Les évacuations préventives ne sont cependant pas obligatoires et concernent principalement l'habitat précaire de type *mobile home* ou maisons préfabriquées.

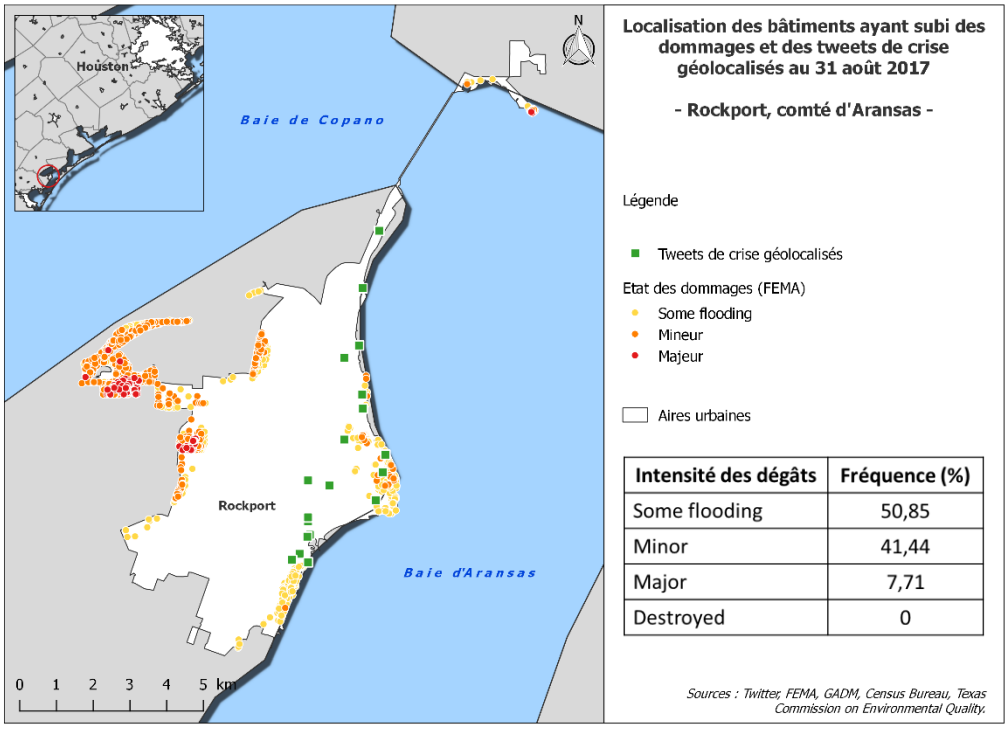


Figure 5.58 : Localisation des tweets de crise et intensité des dommages aux bâtiments des propriétés privées – Rockport (C.Cavalière)

Si l'on n'inventorie que 79 tweets de crise géolocalisés émis pendant la période du 24 au 31 août 2017, l'activité tweeting de crise est plus conséquente qu'en période normale, même si la zone, puisqu'évacuée, a sans doute perdu temporairement certains utilisateurs. Pendant l'ouragan, l'aire urbaine de Rockport représente 0,35% de l'ensemble des tweets géolocalisés émis dans la zone extraite alors qu'entre mars et juin 2016, l'activité virtuelle géolocalisés globale de ce même territoire représente un poids encore plus marginal, soit 0,0002% de l'activité. Néanmoins, le ratio entre le nombre total de tweets et le nombre de bâtiments endommagés reste très faible (0,05 tweet pour un bâtiment endommagé, cf. tableau 5.8).

Tableau 5.8 : Comparaison entre l'événement virtuel et les dommages provoqués par le phénomène physique (Rockport)

Nombre de tweets de crise (Rockport)	79
Bâtiments ayant subi des dommages (FEMA)	1 583
Ratio Nombre de tweets/Nombre de bâtiments endommagés	0,05
Distance minimale Tweet/Bâtiment ayant été endommagé	36 mètres
Distance moyenne Tweet/Bâtiment ayant été endommagé	5,87 kilomètres

Dans l'exploration de cette aire urbaine, un premier fait nous a interpellés sur la carte de la figure 5.58 : le décalage apparent en termes d'intensité des dégâts entre les données relevées par la *FEMA* d'une part et, d'autre part, le contenu de la *SEDB* et des tweets de crise. En effet, comme l'indique la *SEDB*, on constate dans les tweets de Rockport l'emploi récurrent d'un vocabulaire fort qui provient de divers utilisateurs (et non d'un seul individu submergé par l'émotion) : "*The initial reports are devastating and the storm isn't over yet*", "*absolute chaos* […]", "[…] *hundreds of home that have been destroyed*", "*High damage* […]", *etc*. Ainsi, si l'on consulte les chiffres de la *FEMA* indiqués dans le tableau accompagnant la carte de la figure 5.58 (7,71% des bâtiments marqués en dégâts majeurs, aucun marqué comme détruit), les dégâts semblent d'une intensité bien moindre par rapport au ressenti et au vécu des populations. Données témoins officielles et traces de l'événement virtuel ne concordant pas, le problème suivant se pose : pourquoi un tel décalage, à quelle source se fier ? Cette situation se retrouve-t-elle dans d'autres territoires ?

Un second fait s'avère marquant pour cette aire urbaine : l'absence du mot *flood*, auquel se substituent ici les qualificatifs de *swamped* (submerger, inonder) et de *waterlogged* (détrempé, gorgé d'eau). Toutefois, malgré l'évacuation de la zone, on peut distinguer différentes phases dans l'événement virtuel (tableau 5.9), qui soulignent de nouvelles incohérences avec les sources de données officielles : l'annonce de l'évacuation s'avère déjà présente sur le réseau au 24 août et le premier tweet émis relatif aux intempéries indique l'arrivée de l'ouragan à 22h30 sur les terres.

Tableau 5.9 : Thèmes des tweets de crise géolocalisés de Rockport et phases identifiées

Date	Fréquence Tweet (%)	Thèmes	Phase identifiée dans les tweets
24/08	7,59	Annonce des évacuations préventives à Rockport ; Prières	Anticipation
25/08	20,25	Arrivée de l'ouragan ; Manifestations physiques (vents, onde de tempête) ; Dégâts (arbres déracinés, toitures arrachées)	Phénomènes de la crise en cours
26/08	11,39	Poursuite des intempéries ; Dégâts majeurs ; Prières ; Demande de nouvelles des proches ; Signalement de pilleurs	Phénomène et événements de la crise en cours
27/08	11,39	Photo des dégâts ; Prières ; Recherche de survivants	Evénement - Sauvegarde
28/08	3,79	Dégâts (maisons, arbres)	Evénement
29/08	8,86	Photo des dégâts ; Appel aux dons	Evénement - Soutien matériel immédiat
30/08	18,98	Reconstruction ; Entraide	Soutien matériel - Résilience
31/08	12,66	Reconstruction/Déblaiement ; Entraide, Appel aux dons ; Prières	

Au niveau des types de dégâts, on observe une cohérence entre le discours des tweets et les informations fournies par la *SEDB* (toits arrachés, arbres déracinés), mais surtout, l'exemple de Rockport témoigne de la très forte présence d'un vocabulaire lié aux émotions et à la solidarité : les messages de prières sont quasi quotidiens et la période de soutien aux sinistrés, amorcée le 29 août, indique de multiples appels aux dons et de témoignages d'entraide pour la reconstruction des habitations. L'événement virtuel de Rockport ne contient finalement qu'un unique témoignage d'une mesure de sauvegarde ; en fait, le réseau passe directement de la manifestation des phénomènes de la crise en cours à la constatation des dégâts puis à l'engagement rapide d'une dynamique de reconstruction.

Les nuages de mots suivants (figure 5.59) sont classés en fonction de la nature grammaticale identifiée pour chaque mot, en ayant recours à la méthode de l'étiquetage morpho-syntaxique : ils présentent ainsi les noms, adjectifs et verbes employés dans les tweets de crise de Rockport.

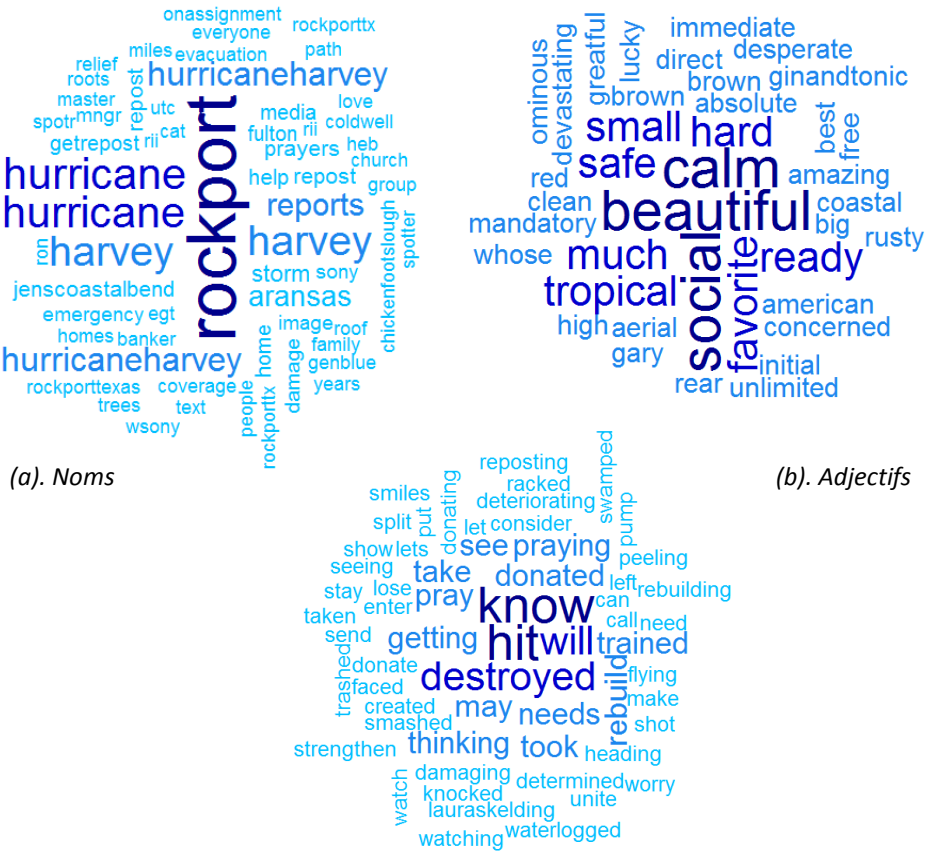


Figure 5.59 : Nuages de mots par nature grammaticale - Rockport, 24-31 août 2017

Les noms identifiés dans les tweets de Rockport se rapportent à l'ensemble des phases de la crise (*evacuation* pendant l'anticipation ; *hurricane*, *damage*, *reports*, *emergency* pendant la crise ; *relief* pendant la période de reconstruction) mais n'en traduisent pas les nuances. Ce sont en effet les adjectifs et verbes qui marquent ce vocabulaire fort identifié à la lecture des tweets ; s'ils qualifient également l'ensemble des phases de l'événement virtuel, ils traduisent en plus le *vécu* des individus face au territoire en crise :

- on retrouve le vocabulaire de l'émotion : des sentiments comme l'inquiétude (*ominous*, *concerned*, *desperate*) transparaissent alors que le *beautiful* est associé à un tweet de prières "*Thinking and praying for my beautiful Rockport I know we all love so much*" émis pendant la période post-crise (phase de soutien matériel dans le tableau 5.9).

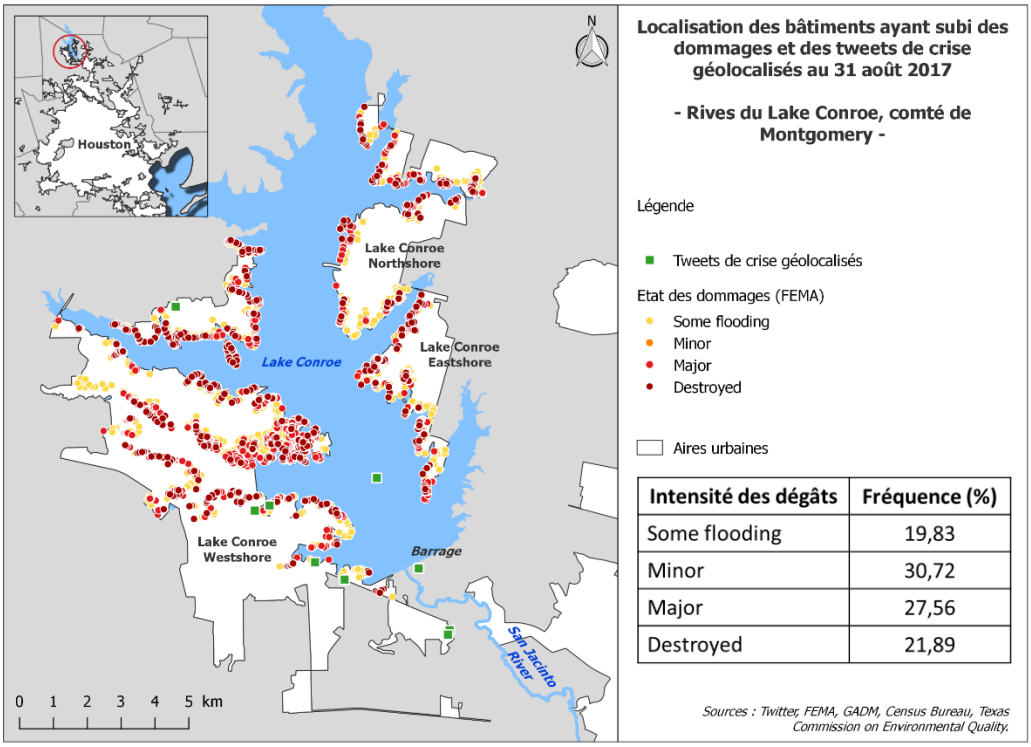
- Le vocabulaire lié à l'intensité des dégâts se trouve également dans ces catégories grammaticales : *destroyed*, *devastating*, *smashed*, *hit*, *etc*.

- Le vocabulaire témoignant de l'organisation de la communauté dans une dynamique de reconstruction est essentiellement perçu à travers les verbes : *faced*, *unite*, *need*, *rebuilding*, *donate*, *determined*, *etc*.

Dans le cas de Rockport, l'événement virtuel indiquerait ainsi l'amorce d'une phase de résilience de la crise à partir du 30 août (date à laquelle le mot *rebuild* apparaît pour la première fois, après les mesures de sauvegarde et les mesures de soutien immédiat via les appels aux dons et à l'entraide locale). Cependant, en raison du décalage entre la localisation des tweets de crise de Rockport et des lieux ayant subi le plus de dégâts (cf. carte de la figure 5.58), ainsi que de la distance moyenne des tweets aux bâtiments endommagés (5,87 km, cf. tableau 5.8), nous préférons qualifier cette période de *résilience virtuelle* : en un lieu précis, nous avons des tweets dont la sémantique décrit des actions caractéristiques de mesures de résilience (travaux de déblaiement et premières reconstructions) mais, les tweets de crise géolocalisés étant peu nombreux, nous ne savons pas si ces témoins sont représentatifs de l'éventuelle diversité des situations locales.

Pour capturer l'éventuelle diversité des événements virtuels enregistrés dans les territoires signalés comme les plus affectés par la presse, nous sélectionnons, pour comparaison, un second exemple d'aire urbaine qui ne constitue pas un territoire côtier. Parmi ces territoires[27], on trouve les trois aires urbaines entourant les rives du lac Conroe (nord-ouest de l'agglomération de Houston, figure 5.60) : *Lake Conroe Eastshore*, *Lake Conroe Northshore* et *Lake Conroe Westshore.*

[27] Source : https://www.nytimes.com/interactive/2017/09/01/us/hurricane-harvey-damage-texas-cities-towns.html (Consulté pour la dernière fois le 13/05/2019)



Intensité des dégâts	Fréquence (%)
Some flooding	19,83
Minor	30,72
Major	27,56
Destroyed	21,89

Figure 5.60 : Localisation des tweets et intensité des dommages aux bâtiments des propriétés privées - Lake Conroe (C.Cavalière)

Dans cet exemple, il s'agit essentiellement de zones résidentielles plutôt aisées : en 2017, le revenu médian par foyer pour ces trois aires urbaines était de 72 658 $ (en comparaison, cette même valeur était alors de 57 051 $ pour l'Etat du Texas, et en ce qui concerne l'exemple précédent de Rockport, le revenu médian par foyer était de 46 914 $). Ici, les dégâts ne sont pas consécutifs aux vents violents et à l'onde de tempête mais aux précipitations ayant entraîné la hausse du niveau des eaux du lac. Notons d'emblée que, d'après les données témoins relevées par la *FEMA*, les dégâts aux habitations semblent plus importants qu'à Rockport (21,89% des habitations notées comme détruites alors qu'aucune n'était inventoriée dans cette catégorie sur l'aire urbaine du littoral). La *SEDB* fournit alors les informations suivantes : les pluies intenses entraînent des crues éclair et inondations le long de la rivière San Jacinto qui traverse le lac du nord-ouest au sud-est. A partir du 26 août, l'élévation record du niveau des eaux du lac Conroe inonde des centaines de foyers sur ses rives et impose finalement, au niveau du barrage, des lâchers d'eau entraînant l'inondation de nombreux foyers et véhicules situés en aval.

En situation non perturbée, les territoires correspondants sont peu visibles sur le réseau : si l'on reprend les données brutes collectées entre mars et juin 2016, l'activité enregistrée sur ces trois aires urbaines était quasi négligeable (0,003% des tweets bruts géolocalisés étaient alors émis depuis les rives du lac Conroe). Pendant la période du 23 au 31 août 2017, cette situation marginale reste tangible puisque les émissions de tweets de crise géolocalisés ne représentent que 0,008% du corpus de crise extrait, et sont localisées dans la

seule aire urbaine de Lake Conroe Westshore. De la même manière, le ratio entre le nombre de tweets de crise géolocalisés et le nombre de bâtiments endommagés reste très faible (0,004 tweet pour une habitation endommagée, cf. tableau 5.10).

Tableau 5.10 : Comparaison entre l'événement virtuel et les dommages provoqués par le phénomène physique (Lake Conroe)

Nombre de tweets de crise (Lake Conroe)	**19**
Bâtiments ayant subi des dommages (FEMA)	**4 811**
Ratio Nombre de tweets/Nombre de bâtiments endommagés	0,004
Distance minimale Tweet/Bâtiment ayant été endommagé	33 mètres
Distance moyenne Tweet/Bâtiment ayant été endommagé	7,58 kilomètres

Au niveau de la sémantique (tableau 5.11), on peut parvenir à marquer le début des différentes phases de l'événement virtuel articulé au réel mais dans le cas des rives du *Lake Conroe*, l'information textuelle contenue dans les tweets reste trop pauvre pour documenter et délimiter précisément la durée de ces différentes phases :

- nous avons marqué certains thèmes comme *ambigus* : la tonalité ou le contenu de certains tweets marquent un décalage inattendu par rapport à la situation physique réelle ("*Hurricane celebration with my lovely bride and my mother in law at our club"* émis le 25 août). Le contenu relève alors parfois de l'interprétation du lecteur du tweet : "*Does my house really have to be a #HurricaneHarvey waypoint?"* émis le 27 août ; à cette date, on sait par la *SEDB* et les données du NWS que ce territoire a déjà été affecté. En outre, la géolocalisation référence le tweet au milieu du lac Conroe. Doit-on en déduire que la maison de cet utilisateur a été endommagée et qu'il a été évacué ?

- On peut dater la survenue des manifestations physiques de l'ouragan le 26 août (ce qui concorde avec la *SEDB*, notamment à propos de l'exceptionnelle montée des eaux du lac) mais les premiers tweets témoins de la phase de sauvegarde n'apparaissent que le surlendemain.

- Les jours suivants, on peut dater précisément la fin des précipitations à partir du 30 août mais aucune information relative aux dégâts des inondations ou encore à la poursuite de cette phase de sauvegarde n'apparaît (alors qu'il est logique d'envisager sa poursuite dans le monde réel). Un unique tweet pourrait évoquer un effet indirect des inondations : l'épanchement d'une nappe d'hydrocarbures sur deux routes. On ne distingue aucune trace de mesures de soutien aux populations affectées, ni d'éventuel engagement dans une dynamique de résilience virtuelle.

Tableau 5.11 : Thèmes des tweets de crise géolocalisés de Lake Conroe et phases identifiées

Date	Fréquence Tweet (%)	Thèmes	Phase présente dans les tweets
25/08	10,5	Photo ; *ambigu*	?
26/08	10,5	Vent violent ; inquiétude face à la montée des eaux du lac	Phénomènes de la crise en cours
27/08	10,5	Ambigu	?
28/08	26,3	Inondations majeures (constat; prières ; conseils si besoin de secours ou refuge ; secours d'un faon.	Phénomène de la crise en cours - Sauvegarde
29/08	10,5	Photo ; arc-en-ciel	Fin du phénomène pluvieux ?
30/08	10,5	Ambigu ; fin du phénomène	Fin du phénomène pluvieux (sûr)
31/08	15,8	Fin du phénomène ; Lac Conroe après l'ouragan	
01/09	5,3	Routes fermées en raison d'un épanchement d'essence	Effet indirect des inondations

Ce deuxième exemple témoigne d'un nouveau décalage entre données de terrain relevées par la *FEMA* et contenu des tweets : en effet, si les données témoins affichent un lieu dont les habitations ont manifestement subi plus de dégâts que sur le littoral de Rockport, le nombre de tweets de crise géolocalisés reste négligeable et leur vocabulaire se montre très différent de Rockport. Autour du lac, l'événement virtuel se révèle davantage focalisé sur la description physique du phénomène ("*Hurricane Harvey still blowing from the East*", "*Massive flooding*") mais pauvre en termes d'informations de vécu, de gestion de crise et d'organisation des populations affectées. Seul un tweet peut traduire un ressenti en rapport à l'observation d'un phénomène physique en mentionnant l'inquiétude de son auteur : "*Dee and I are watching the lake and the effect the storm is having on the lake!*"

Pour conclure sur ces deux premiers territoires d'exploration, force est de constater que les comportements lexicaux des poches d'activité ne sont pas normés : d'une part, l'aire urbaine de Rockport, bien que située dans un comté classé comme périurbain, comptabilise quatre fois plus de tweets de crise géolocalisés que les trois aires urbaines situées autour du lac Conroe, dans un comté classé comme urbain, qui s'avèrent en outre plus peuplées que Rockport et dont les populations n'ont pas été évacuées de manière préventive[28]. Dans un second temps, le récit local du phénomène s'avère hétérogène en fonction des lieux, témoignant ainsi d'une variabilité spatiale de la sensibilité des utilisateurs en fonction des différentes phases : sur les rives du lac, c'est l'ensemble des phénomènes physiques

[28] Notons également que les populations résidant autour du Lac Conroe sont plutôt jeunes : en 2017, l'âge médian était de 36 ans et 82% de la population avait moins de 60 ans.

consécutifs à l'ouragan qui constituent le thème central du corpus (63% des émissions de tweets de crise géolocalisés sont focalisées sur les conditions physiques en temps réel). A l'inverse, 67% des tweets de Rockport sont articulés autour des perturbations consécutives aux phénomènes physiques (et notamment les dégâts), ainsi qu'à la phase de gestion de crise (mesures de sauvegarde et de soutien) et à l'amorce de la période de résilience.

Pour autant, dans les deux cas, le récit de l'événement virtuel n'est constitué que de *bribes* témoignant d'une *séquence* de l'événement social du monde réel (la constatation de dégâts, un appel aux dons, un appel à l'aide d'urgence, le début du nettoyage d'une maison, *etc*.). Si l'on peut envisager la reconstitution de la trame d'un récit de crise avec moins d'une centaine de tweets de crise géolocalisés, cetains points restent en suspens :

- cette trame s'avère générale et temporelle (dans la mesure où l'on peut identifier des thèmes qui varient en fonction du temps) mais qu'en est-il de sa significativité spatiale ? A Rockport, nous avons supposé l'entrée dans une phase de résilience mais la partie occidentale de l'aire urbaine, qui concentre les dégâts les plus importants, reste vide de tweets de crise géolocalisés.
- Les territoires urbains localisés autour du lac Conroe sont des territoires peu actifs sur le réseau virtuel géolocalisé en temps normal : faut-il alors traduire leur silence comme un comportement normal ou comme révélateur d'effets plus graves des phénomènes sur les populations (ce que les données de la FEMA relatives à l'intensité des dégâts laissent supposer) ?

Rockport et les rives du lac Conroe témoignent cependant d'un point commun : cette dissonance entre l'intensité des dégâts indiqués par la *FEMA* (qui sera alors testée dans d'autres territoires pour observer une éventuelle répétition) et la localisation des tweets de crise. Ici, nous nous sommes davantage intéressés à l'exploration temporelle seule du contenu lexical des tweets et à la mise en évidence d'un phasage des différentes périodes de l'événement social généré par le phénomène physique perturbateur. En revanche, si l'on regarde les chiffres mentionnant les distances des tweets aux maisons endommagées, on trouve une distance minimale d'une trentaine de mètres (soit l'équivalent de l'incertitude liée à la mesure GPS) mais une distance moyenne des tweets à l'ensemble des maisons qui se compte en kilomètres. Effectivement, dans les deux cartes des figures 5.58 et 5.60, on trouve une poignée de tweets localisés à proximité des habitations affectées mais, une nouvelle fois, les lieux concentrant une majorité de ces habitations ne sont virtuellement pas renseignés (comme l'ouest de Rockport et les deux aires urbaines de Lake Conroe Northshore et de Lake Conroe Eastshore). Nous avons alors cherché à savoir si les tweets localisés à moins de cinquante mètres des maisons endommagées avaient un rapport direct avec les impacts des phénomènes physiques sur les bâtiments et leurs occupants (tableau 5.12).

Tableau 5.12 : Distance et contenu des tweets de crise géolocalisés par rapport aux habitations endommagées

Lieu	Nombre de tweets géolocalisés situés à moins de 50m d'une habitation	Nombre de tweets en rapport direct avec les effets du phénomène sur l'habitation ou les infrastructures voisines
Rockport (79 tweets)	3	0
Lake Conroe Westshore (19 tweets)	5	3

Cette fois, les tweets de crise géolocalisés semblent de nouveau échapper à la première loi de Tobler : à Rockport, même si l'on comptait davantage de tweets de crise mentionnant des dégâts, seuls trois d'entre eux sont géolocalisés à moins de cinquante mètres d'une habitation endommagée (soit 3,8% des tweets) et leur contenu ne porte pas sur les effets directs des intempéries dans le quartier ou sur la reconstruction. Dans l'aire urbaine de Lake Conroe Westshore, la proportion de tweets localisés dans le voisinage direct des maisons est plus forte (26,3%) mais seuls trois tweets évoquent l'inondation de maisons. Pour la ville de Rockport, les tweets les plus proches des maisons, qui évoquent leur nettoyage et leur reconstruction, sont localisés à une distance moyenne de 934 mètres aux bâtiments affectés.

Face à ces résultats et pour la suite des analyses lexicales, nous orientons nos pistes de cette manière : nous sélectionnons toujours des espaces non métropolitains mais souhaitons comparer les résultats présents avec d'autres territoires aux caractéristiques physiques analogues : un territoire de l'intérieur des terres et un territoire littoral. Observe-t-on les mêmes décalages entre données FEMA et contenu des tweets ? Les sensibilités sont-elles identiques en fonction des territoires ? Observe-t-on toujours une distance d'ordre kilométrique entre les territoires concentrant l'activité tweeting et les territoires concentrant les maisons endommagées ?

5.2.2.5. Vers une répétitivité des premières observations ?

Le territoire littoral de Port Arthur.

L'aire urbaine de Port Arthur, située au sud-est de Beaumont et bordant la rive texane du Lac Sabine, par lequel passe le fleuve du même nom avant de se jeter dans le golfe du Mexique, constitue le second territoire test du littoral. Du 27 au 30 août, la *SEDB* évoque des pluies intenses et continues avec un cumul pluviométrique maximal de 60,58 pouces (soit 1 538,73 mm) dans le comté de Jefferson (dans lequel se trouve Port Arthur) et leurs impacts : cinq victimes, des crues éclair et l'inondation de 64 000 foyers, la fermeture de plusieurs raffineries inondées ainsi que des dégâts sur les infrastructures de distribution et d'assainissement de l'eau.

Malgré l'enregistrement de 128 tweets de crise géolocalisés émis depuis l'aire urbaine entre le 23 août et le 1er septembre (tableau 5.13), le ratio entre ces tweets et le nombre de bâtiments endommagés reste toujours aussi faible (0,012 tweet par dommage inventorié). En ce qui concerne la distance de ces tweets aux habitations, les paramètres évoluent par rapport aux territoires précédents : même si, globalement, les tweets de crise s'avèrent plus distants des lieux de concentration des dégâts, le tweet le plus proche se situe quasiment dans une propriété mais surtout, 45,31% des tweets de l'aire urbaine sont localisés à moins de cinquante mètres d'une habitation endommagée.

Tableau 5.13 : Comparaison entre l'événement virtuel et les dommages provoqués par le phénomène physique (Port Arthur)

Nombre de tweets de crise (Port Arthur)	128
Bâtiments ayant subi des dommages (FEMA)	10 706
Ratio Nombre de tweets/Nombre de bâtiments endommagés	0,012
Distance minimale Tweet/Bâtiment ayant été endommagé	1,32 mètre
Distance moyenne Tweet/Bâtiment ayant été endommagé	14,18 kilomètres
Pourcentage de tweets situés à moins de 50 mètres d'une habitation endommagée	45,31

Dans un second temps, l'aire urbaine de Port Arthur s'avère en fait affectée par deux crises successives (tableau 5.14) : à partir du 27 août, elle est frappée par un premier épisode pluvieux intense car située à la marge de la perturbation qui touche alors Houston. Le 30 août, l'ouragan alors rétrogradé en tempête tropicale, touche terre une seconde fois à Port Arthur (mais à la différence de Rockport, Port Arthur n'a pas fait l'objet d'évacuations préventives). Quelle que soit la crise, une phase d'anticipation et de préparation reste marquée dans l'événement virtuel, par la diffusion de consignes collectives ("*Port Arthur: please be advised, tropical depression #Harvey is in the gulf of Mexico*" ; "*Port Arthur: we are asking residents to stay home and off the roads*") et de comportements individuels ("*We plan to worship this Sunday but we are keeping a close eye on hurricane Harvey*").

Tableau 5.14 : Thèmes des tweets de crise géolocalisés de Port Arthur et phases identifiées

Date	Fréquence Tweet (%)	Thèmes	Phase identifiée dans les tweets
23/08	1,56	Mise à disposition de sacs de sable ; Appel à la vigilance	Anticipation
24/08	0,78	Maintien d'une activité mais vigilance	Anticipation
25/08	2,34	Attente du phénomène ; Inquiétude	Anticipation
26/08	1,56	Soirée jeux	?
27/08	8,59	Premières pluies liées à l'arrivée de l'ouragan sur Houston ; Crues éclair ; Consignes de prudence	Anticipation / Phénomènes de la crise en cours
28/08	14,84	Crue éclair ; Prières ; Inquiétude ; Préparation (ouverture des refuges, fermeture des commerces)	Phénomènes de la crise en cours / Anticipation de la crise (2)
29/08	9,37	Routes inondées ; Invitation à se mettre en sécurité ; Fermeture des commerces ; Inquiétude	Evénements consécutifs à la crise (1) / Anticipation de la crise 2
30/08	36,72	Crues éclair et inondations ; Evacuations ; Appels aux secours ; Prières	Phénomènes de la crise (2) / Sauvegarde
31/08	14,06	Inondations ; Aide et secours ; Dons ; Emotion ; Panne d'électricité	Phénomènes et événements de la crise (2) / Sauvegarde / Soutien
01/09	7,81	Besoin d'aide et de volontaires ; Intensité des inondations ; Refuges	Soutien / Sauvegarde / Informations complémentaires sur le phénomène

Mais incontestablement, ce sont les différentes phases de la seconde crise qui drainent les flux de tweets les plus conséquents : les tweets émis entre le 30 août et le 1er septembre représentent ainsi 58,59% de l'ensemble des tweets de crise de Port Arthur, et 50% des tweets sont émis sur deux jours, le 30 et le 31 août (le flux de tweets émis au quotidien s'engage alors sur une tendance décroissante)[29]. En construisant de nouveau des nuages de mots en fonction de leur nature grammaticale, on peut alors noter trois points essentiels de comparaison avec les résultats de Rockport et des rives du lac Conroe (figure 5.61) :

[29] Il sera cependant difficile de pouvoir comparer les périodes de l'activité tweeting de crise avec les deux autres territoires explorés : les quantités de tweets émis à Rockport peuvent être biaisées en raison de l'évacuation et les rives du lac Conroe enregistrent sept fois moins de tweets qu'à Port Arthur.

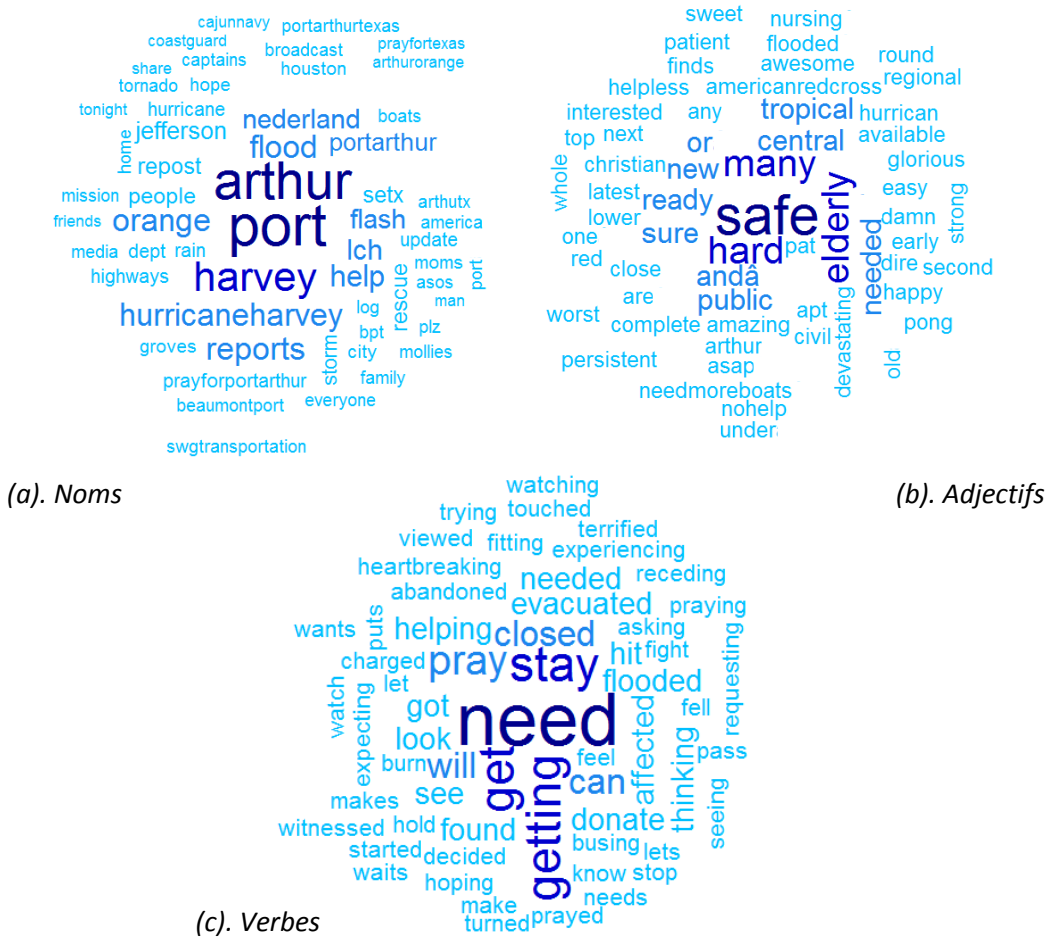


Figure 5.61 : Nuages de mots par nature grammaticale - Port Arthur, 23 août - 1er septembre 2017

- si l'on regarde les noms et adjectifs les plus courants à Rockport (cf. figure 5.59) et à Port Arthur, on constate une certaine proximité puisque les termes de *safe*, *ready* (référence à l'anticipation de la crise) et de *hard,* pour qualifier l'intensité de la crise, s'identifient dans l'événement virtuel des deux villes. Il en est de même pour la mention de l'ouragan (*hurricane* et *harvey*) et des bulletins (*reports*) annonçant la survenue de phénomènes consécutifs aux pluies ou vents (crues éclair, tornades, inondations). A Port Arthur, on retrouve également du vocabulaire exprimant l'intensité des phénomènes (*devastating*, *persistent*, *receding*), mais aussi les sentiments et le vécu des utilisateurs : *helpless*, *heartbraking*, *terrified.* Un utilisateur annonce ainsi que ce phénomène est le pire qu'il ait vécu *"Pray for Port Arthur this is the worst weather I've ever witnessed! Help! We need boats out here!"*.

- En revanche, on identifie ici un fait qui était absent de Rockport et surtout des rives du lac Conroe (qui n'avaient pas été évacuées en prévention) : c'est la forte présence d'un vocabulaire lié aux secours et à la gestion de crise. Le verbe *need* (sous toutes ses déclinaisons : *needs*, *needed*) est le plus fréquent. En fait, bien qu'on ait identifié du vocabulaire empathique et les perpétuels messages de prières à Port Arthur, l'événement virtuel semble ici davantage focalisé sur l'action immédiate : le verbe *need* se trouve ainsi

associé à *boats* ou encore à *help*. En effet, les messages demandant l'intervention des secours ou de l'aide et du ravitaillement pour les sinistrés constituent 22,65% des émissions totales de tweets, enregistrées à Port Arthur. En outre, certains tweets se révèlent des marqueurs de l'utilisation du virtuel par les acteurs officiels de la crise pour organiser la dynamique des secours dans le réel : on trouve un tweet à vocation didactique qui indique les informations exactes à donner pour un habitant demandant du secours via le réseau ("*We need 4 bits of information to dispatch rescue: 1. phone 2.adress 3. individual name*") mais qui ne reçoit pas forcément l'écho attendu ("*2 adults man in wheel chair and wife #elderly #help #needed #asap, #cajunnay, #waconavy*") puisque cet utilisateur préfère l'usage des hashtags à la spatialisation de l'information.

- Enfin, on ne perçoit pas d'amorce d'une phase de résilience virtuelle dans les lieux qui tweetent : les 28 et 29 août, les tweets de crise géolocalisés évoquent une nouvelle préparation pour l'arrivée de la tempête. Après le 30 août, l'événement virtuel reste figé dans cette phase de sauvegarde et de soutien auprès des populations sinistrées, alors que la *FEMA* a déjà répertorié les dégâts aux habitations (puisque nos données sont référencées à cette même date). Après vérification, jusqu'au 3 septembre, aucun tweet géolocalisé émis dans l'aire de Port Arthur ne contient ce vocabulaire identifié à Rockport, que nous avons décrit comme marqueur de l'entrée dans une phase de résilience virtuelle : *rebuild*, *clean*.

De là, doit-on déduire que, parmi les trois territoires explorés jusqu'à présent (et rapportés par la presse comme les plus affectés) par la sémantique de l'événement virtuel, c'est à Port Arthur que le phénomène physique a été le plus dévastateur ? Ou doit-on relier l'activité de crise à la variabilité générale de l'activité virtuelle normale en fonction des territoires? L'ACS fournit des données indiquant le nombre de foyers (*housing units*, qui correspondent aux logements physiques, et non aux foyers fiscaux) par aire urbaine. En rapportant le nombre d'habitations affectées à ces valeurs, il s'avère que c'est dans les trois aires urbaines entourant le lac Conroe que les habitations ont été les plus endommagées : 27,6% des logements sont marqués comme tels par la FEMA (16,8% à Port Arthur et 14,5% à Rockport). Or, autour du lac, un seul tweet mentionne l'intensité de l'inondation ("Massive flooding") et c'est à Rockport qu'on identifie pléthore d'adjectifs et de verbes traduisant l'intensité des dégâts.

Quel(s) indice(s) faut-il alors considérer pour qualifier la gravité d'une crise et comparer des territoires en crise ? A Port Arthur, on enregistre une activité virtuelle existante mais ténue en temps normal, et qui persiste en temps de crise. Autour du lac Conroe, le silence virtuel des situations normales n'est pas brisé par la crise (en dépit de l'intensité des dégâts). Enfin, Rockport, tout comme Port Arthur, est une ville par laquelle l'ouragan a touché terre ; cependant, contrairement à Port Arthur, on ne trouve pas de marqueur de passage à une phase de résilience (pour signifier le début des travaux et mesures destinés à retourner à la situation initiale non perturbée). Au final, si l'on croise les informations du réseau et les chiffres de la FEMA, on pourrait hiérarchiser ces trois territoires en fonction de la gravité apparente (de la crise la plus grave à la moins grave) : les rives du Lac Conroe (indices : silence

virtuel et sévérité des dégâts inventoriés par la FEMA), Port Arthur (indice : forte prégnance du vocabulaire de sauvegarde et de soutien) et Rockport (indice : entrée dans une période de résilience non apparente sur les deux autres territoires).

Pour finir, comme l'exemple de Port Arthur contient un nombre de tweets plus conséquent, nous testons leur composante spatiale : à partir des tweets du 30 et du 31 août 2017, soit les deux journées qui cumulent le plus de tweets de crise géolocalisés, nous cherchons à vérifier si les tweets spatialement proches sont cohérents d'une part, dans leur discours et d'autre part, avec les données des manifestations du phénomène physique (cumuls pluviométriques, NWS) et de l'événement social qui en résulte (dommages aux habitations, FEMA). Les résultats sont cartographiés dans les figures 5.62 et 5.63 : le sens du contenu sémantique des tweets se trouve de nouveau résumé par l'utilisation d'icônes dont la taille varie en fonction du poids représenté par chaque thème identifié dans les tweets inclus dans chaque cluster. Les clusters sont nommés en fonction du thème majoritairement représenté.

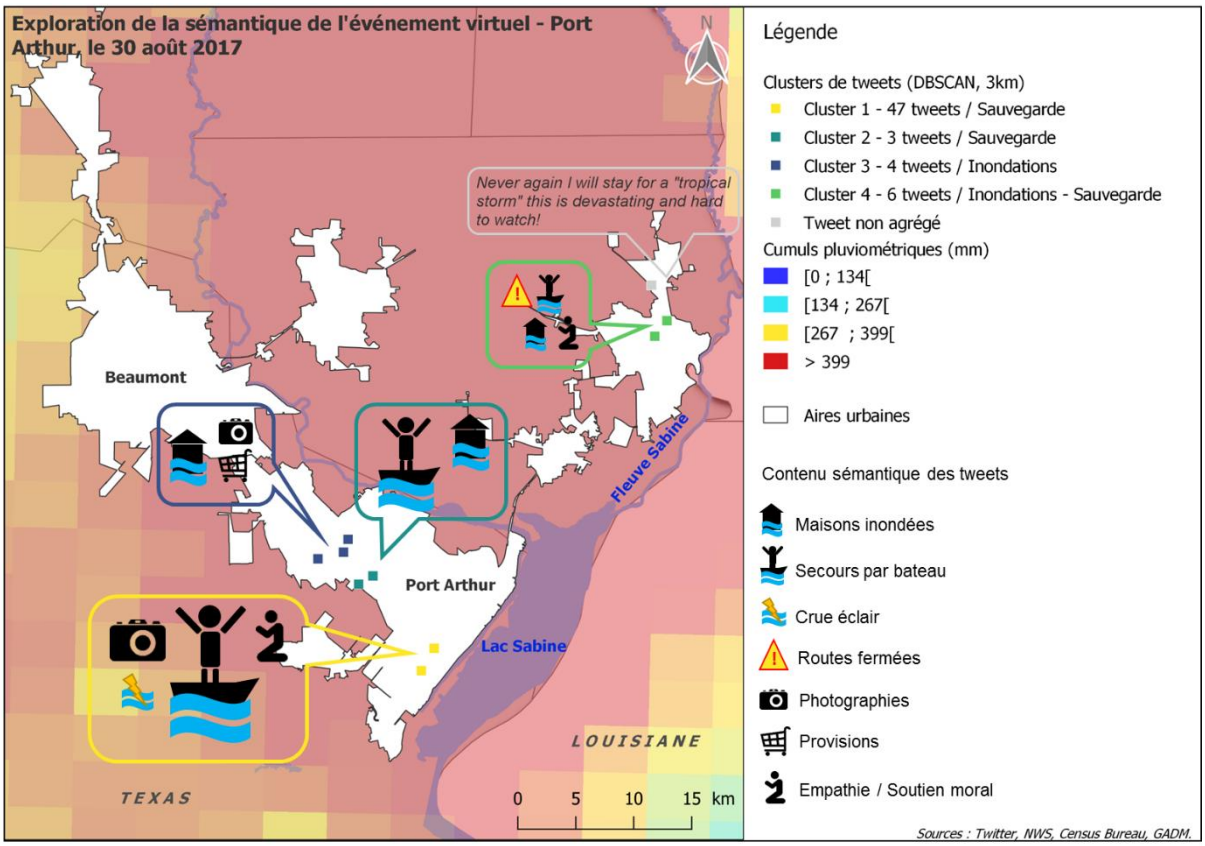


Figure 5.62 : Exploration sémantique de l'événement virtuel clustérisé de Port Arthur le 30 août 2017 (C.Cavalière)

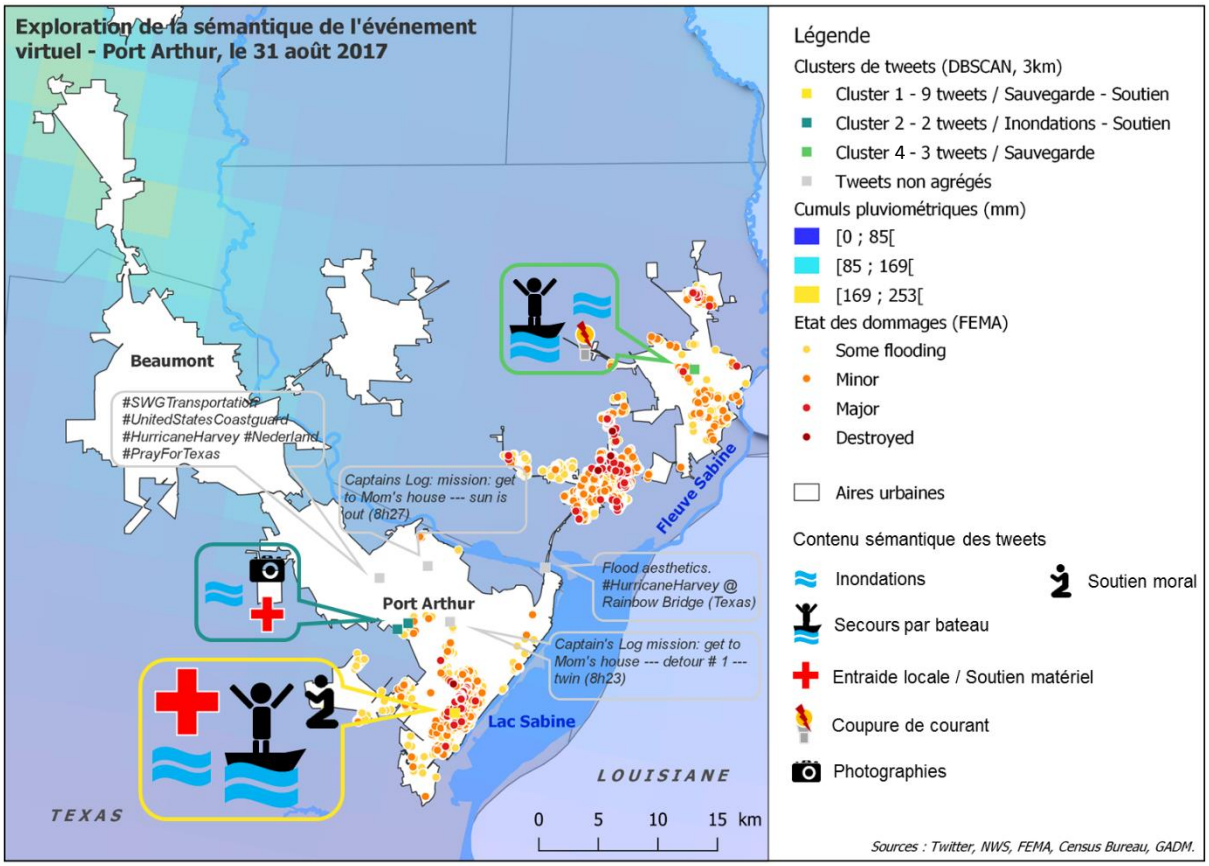


Figure 5.63 : Exploration sémantique de l'événement virtuel clustérisé de Port Arthur le 31 août 2017 (C.Cavalière)

Le 30 août, le phénomène physique s'avère difficile à nuancer, l'ensemble de l'aire urbaine étant affectée par les précipitations les plus intenses (figure 5.62). On distingue néanmoins trois clusters informant d'un besoin d'aide et de bateaux pour évacuer les populations résidentes :

- le cluster 1 (sauvegarde le 30 août, puis sauvegarde et soutien aux sinistrés le 31 août), situé à moins de cinq kilomètres de la rive du lac Sabine, indique la survenue de crues éclair, informe de la présence d'équipes de sauvetage et recèle de tweets diffusant des appels à l'aide des populations sinistrées (*"#Repost We need help out here right now! It's not one boat* [tweet tronqué]" ,"*My friend and her family are in dire need of help. There are two of them and* [tweet tronqué]"). C'est par ailleurs dans ce cluster qu'on identifie les tweets diffusant les consignes relatives aux informations à fournir aux secouristes : *"#HarveyRelief: Port Arthur/Orange Texas adresses needed!*". Si nous constatons que certains utilisateurs demandent de l'aide sans fournir les indications requises et vraisemblablement pour faire secourir des personnes qui n'ont pas les outils ou les moyens de manifester leur détresse sur le média social, d'autres respectent ces critères : "*Need a boat rescue at 2611 East 12th Street, Port Arthur, 77 640. [Nom de la personne à secourir] (74 y/o old)*". Enfin, le discours du cluster concorde avec les données FEMA visibles sur la carte de la figure 5.63 : dans un rayon de trois kilomètres autour des tweets clustérisés, on trouve 70,38% des habitations affectées de l'aire urbaine. Notons

également qu'en dehors d'une alerte aux crues éclair, les mots *flood* ou *flooded* n'apparaissent pas dans cet événement virtuel.

- le cluster 4 (sauvegarde et inondations le 30 août puis sauvegarde le 31 août) correspond au quartier de West Orange et jouxte les rives du fleuve Sabine. On trouve également un vocabulaire de détresse le 30 août "*Need more boats in Port Arthur and Orange, Texas #HurricaneHarvey, #NeedMoreBoats*". Ici, l'événement virtuel semble s'intensifier le 31 août : du hashtag *#NeedMoreBoats*, accompagné d'un vocabulaire exprimant l'inquiétude vis-à-vis des proches, on passe à une plus forte présence d'un vocabulaire de gestion de crise : "*evacuations*", "*rescuers*", "*electricity*", en plus de la mise à disposition d'un numéro d'urgence autre que le 911. Ici, on peut en outre identifier des utilisateurs qui emploient des mots forts, absents dans les tweets des autres clusters ("*helpless*", "*devastating*"). S'agit-il d'un comportement isolé ou le ressenti de la violence des phénomènes physiques est plus intense à West Orange que dans le centre de Port Arthur ? Ou faut-il comprendre que si des utilisateurs tweetent à propos de leur ressenti vis-à-vis du phénomène physique, la situation est moins urgente à West Orange que dans le sud de Port Arthur sur les rives du Lac Sabine ? (car si l'on se réfère aux données FEMA [cf. figure 5.63], West Orange ne concentre que 7,69% des habitations endommagées de l'aire urbaine).

- le cluster 2 (sauvegarde le 30 août, puis inondations et soutien aux sinistrés le 31 août) se situe à une douzaine de kilomètres du lac Sabine et contient un tweet témoin de l'inondation et des évacuations en cours le 30 août ("*Now our civil center is flooded and ppl started to evacuated there! #SMH #HarveyAhoe!*"). Il concorde par ailleurs avec les données de la FEMA : sur la carte du 31 août, on observe bien un îlot de bâtiments endommagés mais ici, le vocabulaire employé dans les tweets n'indique pas un ressenti aussi intense qu'à West Orange.

Pour finir, le cluster 3 (inondations le 30 août) se situe plus à l'intérieur des terres, à une distance moyenne de quinze kilomètres des rives du lac. Ici, le comportement sémantique de l'événement virtuel s'annonce différent : bien qu'un tweet indique la survenue d'une crue éclair le 30 août, l'ensemble des tweets de crise émis pendant cette même journée n'arborent pas la dimension de l'aide d'urgence des lieux précédents et s'avèrent plus hétérogènes dans leur contenu : "*This is the lake of the front yard ☹ I hope these sandbags hold…*", "*Please make Harvey go away now #KalamityJane #MuttsofInstagram #Harvey2017 @ Nederland*". Le 31 août, le cluster 3 n'est plus actif. Ici, la tonalité des tweets émis le 30 août ("*I hope*", "*Please make Harvey go away*"), associée à la dislocation du cluster le lendemain (alors que les précédents clusters persistent) permettent d'envisager un phénomène et des effets moins violents que sur les rives du lac et du fleuve, ce que confirmeraient les données de la FEMA : au nord-ouest de Port Arthur, aucune habitation endommagée n'est enregistrée.

Pour conclure ce nouveau test spatial, on détecte pour ce territoire une certaine cohérence entre la localisation des tweets et de leur contenu par rapport aux données témoins de la FEMA. En outre, les tweets voisins peuvent comporter un discours similaire : le

30 août, 60,6% des tweets inclus dans le cluster 1 évoquent des opérations de secours. En revanche, il reste deux points sensibles :

- la ville de Bridge City (au sud de West Orange) concentre 19,21% des habitations endommagées mais 60,34% des habitations aux dégâts majeurs ou détruites dans l'aire urbaine. Malgré cela, elle reste invisible sur le réseau.
- l'intensité réelle (et non ressentie) d'un phénomène et des événements résultants dans le monde réel reste un paramètre difficile à qualifier par l'événement virtuel : ici, le ressenti sous-entendrait un phénomène réel plus violent dans la partie nord de l'aire urbaine, à West Orange, ce qui ne s'accorde pas avec les lieux concentrant le plus de dégâts d'après la *FEMA*. C'est en effet dans le nord de l'agglomération qu'on trouve des adjectifs témoignant de l'intensité alors que l'activité enregistrée dans le centre de Port Arthur est focalisée sur l'action immédiate[30].

Exploration d'un territoire situé à l'intérieur des terres et traversé par des espaces humides : Lake Charles, Louisiane.

Pour terminer ce test focalisé sur la répétitivité éventuelle des résultats observés dans les premières analyses effectuées sur Rockport et les rives du lac Conroe, nous sélectionnons un second espace situé à l'intérieur des terres et dans un milieu humide : il s'agit ici de l'aire urbaine de Lake Charles (comté de Calcasieu), en Louisiane (figure 5.64). Les informations contenues dans la *SEDB* s'avèrent plutôt réduites : le 27 août, le comté de Calcasieu subit des pluies intenses (entre 15 et 31 pouces, soit entre 381 et 762 mm d'eau). A Lake Charles, 1 572 maisons sont inondées.

[30] La question du tweet comme marqueur d'intensité en un lieu précis est approfondie dans le prochain chapitre, à partir du bilan conclu de cette première exploration.

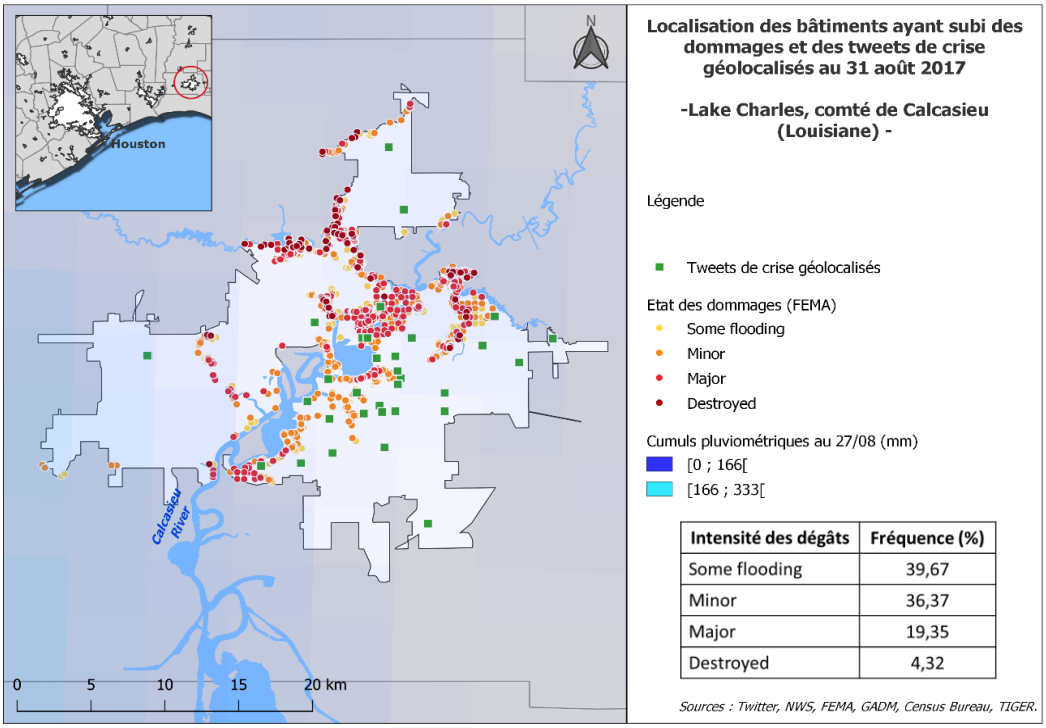


Figure 5.64 : Localisation des tweets et intensité des dommages aux bâtiments - Lake Charles (C.Cavalière)

Dans un premier temps, force est de constater que Lake Charles s'avère finalement épargnée par les précipitations les plus violentes (qui s'abattent en fait sur l'ouest du comté). On pourra néanmoins souligner une incohérence entre les informations incluses dans les sources institutionnelles : les données témoins *NWS* et le contenu de la SEDB annoncent des effectifs différents en ce qui concerne les habitations sinistrées : les données de la FEMA inventorient un total de 2 569 habitations endommagées alors que la SEDB en indique 1 572.

Lorsqu'on s'intéresse à l'événement virtuel enregistré dans cette aire urbaine, le ratio tweets/habitations endommagées reste toujours aussi mince (tableau 5.15). Malgré cela, l'événement virtuel de Lake Charles reste plus important : par rapport aux rives du lac Conroe, les émissions de tweets sont quasiment multipliées par cinq alors que Lake Charles enregistre des dégâts dus aux inondations bien moindres que dans les aires urbaines du lac Conroe (et que la dynamique du phénomène est sans doute identique : un épisode pluvieux intense suivi de crues et d'une élévation du niveau des eaux des lacs). En revanche, on peut s'interroger quant à la pertinence de la localisation des tweets de Lake Charles : alors que les habitations endommagées se répartissent le long des cours d'eau, lacs et zones humides, les tweets se concentrent dans la partie orientale de l'aire urbaine et le tweet le plus proche d'une habitation endommagée reste tout de même distant de 82 mètres. En outre, les territoires longeant les rives des cours d'eau du nord et du nord-est de l'aire urbaine restent invisibles dans l'événement virtuel.

Tableau 5.15 : Comparaison entre l'événement virtuel et les dommages provoqués par le phénomène physique (Lake Charles)

Nombre de tweets de crise (Lake Charles)	92
Bâtiments ayant subi des dommages (FEMA)	2 569
Ratio Nombre de tweets/Nombre de bâtiments endommagés	0,036
Distance minimale Tweet/Bâtiment ayant été endommagé	82,48 mètres
Distance moyenne Tweet/Bâtiment ayant été endommagé	7,76 kilomètres

L'exploration lexicale de la sémantique des tweets de Lake Charles révèle en premier lieu que 21% des tweets sont liés à l'émission automatique de bulletins météorologiques ("*Lake Charles LA Sun Aug 27th PM forecast: tonight thunderstorm LO 71 Monday thunder storms HI 79*") ; par ailleurs, 13% des tweets filtrés dans l'événement virtuel se révèlent ambigus, c'est-à-dire qu'on ne sait pas quel sens leur attribuer : "*Gonna take a bite out of #HurricaneHarvey*", "*#HurricaneHarveyPartu with these two crazies*[31]". Pout autant, on peut une fois de plus identifier différentes phases de réponse à la crise via l'événement virtuel (tableau 5.16) :

Tableau 5.16 : Thèmes des tweets de crise géolocalisés de Lake Charles et phases identifiées

Date	Fréquence Tweet (%)	Thèmes	Phase identifiée dans les tweets
24/08	9,09	Annonce orages ; Préparation	Anticipation
25/08	7,79	Annonce orages ; Préparation	Anticipation
26/08	11,69	Annonce orages ; Préparation ; Prières Texas	Anticipation
27/08	16,88	Fortes pluies ; Alertes et crues éclair en cours ; Prières Houston	Alerte / Phénomènes de la crise en cours
28/08	23,38	Tempête tropicale, deuxième jour de pluie ; Inondations et alertes aux crues éclair ; Comportements ; Evacuations ; Alerte aux tornades ; Prières	Alerte / Phénomènes de la crise en cours / Sauvegarde
29/08	11,69	Comportement ; Aide dans les refuges ; Phénomène terminé / crues éclair	Phénomène de la crise en cours / Sauvegarde / Soutien
30/08	10,39	Mise à disposition de pétrole et gaz ; Comportements ; Routes inondées	Evénements consécutifs aux phénomènes / Soutien / Sauvegarde
31/08	6,49	Entraide ; Amélioration des conditions météorologiques ; Prières	Soutien
01/09	2,60	Prières	Soutien moral

[31] Dans ce tweet, il s'agit vraisemblablement de l'organisation d'une *hurricane party* (faute de frappe) entre amis.

- contrairement aux aires urbaines du lac Conroe, la phase d'anticipation est plus longue puisqu'on trouve des messages de préparation dès le 24 août (au total, ces messages émis pendant cette phase d'anticipation représentent plus de 28% de l'événement virtuel de Lake Charles) : "*Bracing for Hurricane Harvey, Sulphur, Louisiana*", "*Me and my little new friend!! I've pumped up! Prep for floodings*".

- On constate ici un décalage entre les informations de la SEDB et les temporalités de l'événement virtuel : les premiers tweets informant de la survenue des pluies intenses sont bien détectés dès le 27 août ("*Raining a lot. That's all folks! Lake Charles, Louisiana*") mais l'événement virtuel indique que le phénomène pluvieux se poursuit le lendemain ("*Second rainy day in Lake Charles, Louisiana*") et c'est d'ailleurs le 28 août qui draine le flux de tweets le plus volumineux. On repère néanmoins une incohérence interne au réseau quant à la fin du phénomène physique : alors qu'un tweet annonce une fin de phénomène le 29/08 "*After Harvey […]*", d'autres, qui arborent un caractère officiel, indiquent des crues éclair toujours en cours : "*At 5:27 AM, 1 E Lake Charles [Calcasieu CO, LA] broadcast media reports flash flood #lch*".

- Une nouvelle fois, l'organisation de la gestion de crise (sauvegarde et soutien) se manifeste dans le réseau, par la diffusion d'informations relatives aux refuges, aux routes coupées et à la mise à disposition de carburant pour les sinistrés. Urgence et émotion restent cependant moins perceptibles qu'à Rockport et Port Arthur : les messages demandant de l'aide pour une évacuation urgente sont quasiment absents malgré les inondations et le lexique empathique ne se retrouve que dans 7% des messages.

Sur l'aire urbaine de Lake Charles, on pourra également noter qu'on retrouve un type de tweets qu'on avait identifié lors de l'exploration lexicale des tweets émis en réponse aux phénomènes récurrents et qui était absent des territoires de l'ouragan explorés jusqu'ici : le tweet exprimant des comportements et actions individuelles ou collectives. Ici, on trouve ainsi des tweets témoignant de l'annulation d'événements collectifs habituels ("*#Repost ChristianworlDLC Due to inclement weather, our church offices will be closed Wednesday*"), mais également des tweets indiquant des comportements individuels a priori inadaptés : "*Maybe not the best idea to go out with all this rain, but we got tired of being in the house all* [tweet tronqué]". Par ailleurs, on pourra toujours noter le même silence virtuel sur la période de résilience : on ne détecte aucun vocabulaire indiquant le retour des populations et leur entrée dans une phase de nettoyage et de reconstruction.

Bilan de l'exploration lexicale en réponse à un phénomène extrême rare

Nous avons ici exploré deux territoires du littoral (Rockport et Port Arthur) et deux territoires traversés par des espaces humides à l'intérieur des terres (rives du lac Conroe et aire urbaine de Lake Charles). L'association de l'événement virtuel et des informations de la SEBD permettent, dans un premier temps, d'entrevoir deux types de manifestations physiques différentes : alors que Rockport semblerait davantage affectée par les vents violents et l'onde de tempête (réseau et SEDB évoquaient ainsi des toitures et arbres arrachés et le mot *flood*/*flooding* était absent du réseau), Port Arthur, Lake Charles et Lake Conroe subissent des épisodes pluvieux intenses entraînant l'élévation du niveau des eaux des rivières et lacs voisins, ayant sans doute pour conséquence des inondations plus importantes qu'à Rockport (l'absence d'un vocabulaire de nettoyage, reconstruction pourrait alors s'expliquer par une décrue plus longue). Néanmoins, quels que soient les types de phénomènes ayant affecté ces territoires, on peut distinguer deux points communs dans les quatre événements virtuels :

- le pic de tweets de l'événement virtuel a lieu pendant la période de gestion de crise (sauf à Rockport où ce pic était enregistré pendant la phase de résilience mais cette observation peut être biaisée par l'évacuation préventive de la ville du littoral puisque seuls les utilisateurs étant restés sur place ont pu tweeter pendant la période de crise).
- L'événement virtuel décrit les manifestations physiques du phénomène et enregistre les modes d'organisation des populations affectées et acteurs de la crise.

Pour autant, on constate une série de points à nuancer : en premier lieu, événement virtuel et données officielles témoins ne sont pas toujours concordants. Nous avons en effet observé que localisation et effectifs de tweets émis dans différents territoires ne reflètent pas dans tous les cas avec exactitude les lieux affectés et l'intensité réelle des dégâts. En conséquence, et pour compléter les propos annoncés dans le bilan précédent relatif aux événements récurrents, les tweets de crise géolocalisés ne semblent pas assimilables à des marqueurs d'intensité fiables en un lieu réel.

En outre, le contenu et le poids d'un événement virtuel généré en réponse à un phénomène extrême rare s'avère difficilement prévisible. Dans les quatre territoires explorés, nous avons identifié des événements virtuels au contenu sémantique variable : un événement virtuel mineur à Lake Conroe alors que davantage de dégâts sont enregistrés par rapport à Lake Charles, dont l'événement virtuel se révèle bien plus important en termes d'effectifs de tweets ; un événement focalisé sur l'émotion et la reconstruction à Rockport ; de la même manière, un événement enregistrant une prépondérance de l'action d'urgence à Port Arthur (qui constitue par ailleurs le lieu dans lequel on a pu observer une correspondance entre proximité des tweets géolocalisés aux quartiers affectés et discours contenu dans les tweets).

Bilan de l'exploration lexicale en réponse à un phénomène extrême rare (suite)

Pour terminer, on pourra signaler la persistance de lieux affectés invisibles du réseau mais également de phases temporelles manquantes : malgré l'enregistrement de tweets témoignant d'une anticipation des populations et autorités avant la crise, les phases d'entrée en vigilance et alerte (marquée par les mots *watching* et *warning*) n'ont été identifiées qu'à Lake Charles en Louisiane. En outre, si nous avons indiqué la possibilité de distinguer différentes phases dans l'organisation de la réponse sociale au phénomène par l'exploration sémantique du réseau, et ce malgré un nombre restreint de tweets, ceux-ci restent trop peu nombreux pour restituer le récit de dynamiques complexes : bien qu'on puisse identifier des lieux concentrant les appels à l'aide pour évacuer (comme le centre de Port Arthur), on ne capture que les bribes d'un récit : il n'y a pas de retweets d'un message particulier, pas de réponse à un utilisateur précis et la seule trajectoire identifiée concerne l'utilisateur de Port Arthur qui annonce, en trois tweets, son parcours pour rejoindre le domicile de sa mère, situé dans le centre de la ville, en bordure du Lac Sabine.

Conclusion du chapitre 5

Dans l'introduction de ce chapitre, nous avions soumis la question suivante : *le tweet de crise géolocalisé constitue-t-il un témoin systématique des interactions entre phénomènes naturels et populations affectées dans le monde réel ?* Cette question a ainsi été abordée en fonction de deux types de phénomènes naturels d'origine hydrométéorologique : les épisodes récurrents de pluies et d'inondations ainsi que par le phénomène extrême rare que constitue l'exemple de l'ouragan Harvey.

Dans un premier temps, la réponse virtuelle à la survenue d'un phénomène naturel réel reste, quels que soient le type de phénomène considéré et son intensité, influencée par le profil des populations et des territoires : ainsi, la métropole reste le territoire le plus virtuellement actif en période de crise et, *a contrario*, la majorité des territoires affectés s'avèrent peu documentés voire absents de l'événement virtuel. Les analyses statistiques mettent en évidence un second comportement général : quels que soient les profils de populations et de territoires considérés, une minorité de lieux concentrent la majorité de l'activité virtuelle géolocalisée de crise. On pourra néanmoins retenir l'existence de dynamiques engendrées par le phénomène rare seul : certains comtés qui se montraient peu (voire pas) actifs pendant les phénomènes enregistrés au printemps 2016, se sont réveillés lorsqu'ils ont été frappés pour l'ouragan en août 2017 (c'est notamment de cas des comtés de Jefferson [villes de Port Arthur et de Beaumont] et d'Orange [villes de Bridge City et d'Orange]). Alors que ces territoires témoignent d'une sensibilité face au phénomène extrême rare, les populations des territoires métropolitains restent sensibles face à tous types de phénomènes naturels.

De la même manière, quels que soient les types de phénomènes considérés, on peut distinguer différentes phases de l'événement virtuel (et ce, même malgré un nombre restreint de tweets de crise géolocalisés). L'analyse sémantique révèle ainsi l'existence de différents thèmes (qui se succèdent ou s'entrecroisent en fonction du temps) : l'anticipation de la crise (avec les mentions de comportements individuels ou collectifs et la diffusion de consignes), l'alerte, le signalement des conditions environnementales en temps réel (via le rapport de phénomènes météorologiques et d'inondations en cours, ou encore des événements qui leur sont associés comme une route fermée, des maisons inondées, *etc.*), les mesures de sauvegarde et de soutien matériel aux sinistrés (opérations de secours, évacuations, recherche de bénévoles, collecte de dons, mise à disposition de biens matériels pour les sinistrés). Les messages de prières ou affichant une tonalité empathique sont quant à eux, inclus dans les différentes phases de l'événement virtuel, quel que soit son type. Pour autant, la plupart des messages géolocalisés affichent une consonance officielle (qu'ils concernent la diffusion d'alerte, de consignes, ou encore de conditions météorologiques) ; en fait, peu de messages provenant d'individus au cœur de crise (et tweetant à titre privé) traduisent directement un vécu et un ressenti par rapport au phénomène en cours. En conséquence, même si ces faibles quantités de tweets de crise géolocalisés indiquent l'existence de

tendances générales qui varient en fonction du temps, leur contenu n'est pas suffisant pour reconstituer les facettes d'un récit de crise cartographiable des populations dans un territoire en crise.

Par ailleurs, en dépit de la capacité de l'événement virtuel à se scinder en différentes phases, il semble que ses autres caractéristiques ne soient pas prédictibles : quel que soit le phénomène considéré, la localisation des tweets de crise n'est pas toujours voisine des phénomènes et événements inventoriés dans le réel. Ainsi, en avril 2016, un phénomène d'inondation était invisible du réseau sur le comté de Jefferson ; de même, pendant l'ouragan, les aires urbaines situées sur les rives du lac Conroe (qui témoignent des plus fortes proportions d'habitations sinistrées ou complètement détruites parmi les territoires explorés) figurent parmi les moins actives sur le réseau. A l'inverse, dans le quartier de Kingwood (aire métropolitaine de Houston), dans lequel on avait détecté une activité virtuelle soudaine le 31 août 2017, les tweets de crise étaient localisés dans le voisinage direct des propriétés sinistrées et leur lexique se montrait cohérent. De même, à Port Arthur, les tweets clustérisés les plus proches du fleuve et du lac Sabine se montraient similaires en mentionnant des appels à l'aide et des opérations de secours dans les zones inondées. Au contraire, les clusters les plus éloignés du fleuve et du lac contenaient moins de tweets, mais leur contenu s'avérait plus hétéroclite. Enfin, pour le cas de Rockport, nous avons introduit l'idée d'une résilience virtuelle pour qualifier l'existence de témoins sémantiques caractéristiques d'une situation de résilience (nettoyage et reconstruction) mais les tweets mentionnant ces opérations étaient en réalité distants des lieux de concentration des habitations affectées. Au final, nous disposons d'un événement virtuel certes précisément géolocalisé, mais dont on ne sait pas si le discours est articulé, dans tous les territoires, à la situation locale.

En conséquence, face à la diversité des comportements virtuels constatés, il semble que la localisation des poches d'activité virtuelle de crise ne constitue pas, dans tous les cas, un reflet objectif de l'intensité d'une crise en un territoire donné : en effet, à de multiples reprises, nous avons pu relever des dissonances entre les données témoins localisant les lieux affectés et la géolocalisation des poches de l'activité virtuelle de crise (dans le sud de l'aire métropolitaine de Houston en avril 2016 ou encore à Rockport et à Lake Charles où des lieux affectés étaient vide de tweets de crise). Quels indices considérer alors comme marqueurs révélateurs de l'intensité d'une crise ?

- le ou les thèmes inclus dans la sémantique de l'événement virtuel : près du lac Sabine à Port Arthur, nous avons vu que l'événement virtuel des 30 et 31 août 2017 était en majorité focalisé sur les demandes d'intervention des secours ; peut-on alors considérer que les lieux enregistrant une activité virtuelle homogène, répartie en une poignée de thèmes prépondérants, subissent une crise plus intense que les lieux dans lesquels on identifie des contenus hétérogènes ? (par exemple la description des conditions environnementales ou du ressenti du phénomène, c'est-à-dire l'information qui était présentée ci-avant comme rare, ou encore la diffusion des messages de prières ou de soutien moral).

- le silence virtuel ou la disparition de l'activité tweeting : dans les différents territoires, nous avons constaté l'existence de quartiers témoignant de fortes concentrations d'habitations sinitrées (par les données de la FEMA) mais vides de tweets de crise géolocalisés. Il serait logique de considérer le silence comme indice d'intensité dans la mesure où les individus en situation d'urgence ne vont pas perdre du temps sur les réseaux sociaux (et d'ailleurs, on pouvait voir, à Port Arthur, que certains tweets géolocalisés requéraient l'intervention des secours pour des tiers). En revanche, dans le cas des rives du lac Charles, ce silence en situation de crise paraît difficile à mesurer étant donné que ce territoire est déjà peu inscrit sur le réseau en temps normal.

6. Le tweet de crise géolocalisé, un marqueur spatio-temporel pertinent des dynamiques et des paramètres du phénomène réel ?

Dans ce dernier chapitre, nous changeons d'échelle spatiale pour nous focaliser sur l'espace métropolitain, c'est-à-dire l'individu statistique hors-normes qui concentre les plus fortes quantités de tweets géolocalisés, quelle que soit la situation (normale ou perturbée). Dans le chapitre précédent, qui constituait la base de *primo-connaissances*, la métropole se présentait en effet comme un milieu sensible et réactif face aux phénomènes récurrents ou extrêmes. En revanche, la géographie virtuelle du milieu métropolitain semblait suivre les logiques identifiées à plus petite échelle : par l'étude de la crise du 16 au 21 avril 2016, on avait d'ores-et-déjà distingué des lieux d'émission de lieux invisibles de l'aire métropolitaine de Houston, tout comme nous avions identifié des comtés plus ou moins inscrits sur le réseau en fonction du profil des populations.

Avec ce changement d'échelle, nous nous focalisons sur l'étude des lieux de réactivité d'un territoire particulier, la métropole. Le premier objectif de cette nouvelle série d'analyses consiste ainsi à discerner les facteurs qui expliciteraient la présence ou l'absence de tweets géolocalisés dans des lieux particuliers. Le deuxième objectif consiste à tester la pertinence du tweet comme marqueur, au niveau de ses trois composantes : dans le chapitre précédent, nous avions observé que la première loi de Tobler n'était pas systématiquement vérifiée, qu'on considère les tweets de crise géolocalisés entre eux ou qu'on les compare à la localisation et à l'intensité des dégâts. Dans la métropole, constate-t-on le même comportement spatial que dans les milieux non métropolitains ou peut-on identifier une significativité spatiale et temporelle dans la distribution des tweets ? De même, nous envisageons de comparer la dynamique du réel et la dynamique virtuelle en variant la résolution temporelle afin d'identifier les pas de temps qui apportent du sens dans l'analyse sémantique des tweets de crise à grande échelle.

Enfin, nous souhaitons évaluer le tweet de crise géolocalisé comme marqueur d'un paramètre social de la crise, à savoir la vulnérabilité des individus dans les différents territoires de la métropole : précédemment, nous avons présenté les paramètres spatiaux et quantitatifs des tweets de crise géolocalisés comme des marqueurs peu fiables de l'intensité d'une crise en un lieu donné, en comparant lieux de l'activité tweeting et données offrant un aperçu des dégâts causés aux habitations. En revanche, le contenu sémantique ainsi que le silence relatif ou la brusque apparition d'un événement virtuel, même mineur, pouvaient se montrer comme des témoins plus pertinents. Peut-on alors considérer le tweet de crise géolocalisé comme un nouveau moyen d'appréhender, par la cartographie et le calcul de paramètres

statistiques associés à l'analyse sémantique, la variabilité spatiale de la vulnérabilité en fonction des réponses virtuelles locales face au phénomène ?

Ce dernier chapitre est structuré comme suit : la première partie est articulée autour de l'identification de facteurs explicatifs de l'activité tweeting dans les milieux métropolitains. La deuxième partie examine la question des logiques spatio-temporelles de la distribution de l'activité virtuelle de crise dans les différents lieux de la métropole ainsi que son contenu sémantique. Le dernier axe explore la mise en relation d'indices statistiques avec la sémantique des tweets de crise.

6.1. Les lieux et les facteurs de l'activité virtuelle dans le territoire métropolitain

Dans le chapitre précédent, nous avons mis en évidence que les territoires concentrant les plus fortes quantités de tweets correspondent à des individus statistiquement hors-normes, quel que soit le milieu concerné (urbain comme hyper-rural). Dans cette première section, nous abordons la question des logiques de distribution spatiale de l'activité tweeting ainsi que les biais susceptibles d'influencer la répartition des tweets géolocalisés dans les différents territoires des métropoles, en nous référant à l'unité de recensement la plus fine : le *Census Tract*. Nous recourons à deux aires métropolitaines : dans le souci d'examiner la répétitivité des observations, San Antonio constitue la ville témoin et Houston, l'objet d'étude principal. Il s'agit en outre de répondre à la question suivante : la spatialisation des émissions de tweets géolocalisés du milieu métropolitain en période de crise est-elle prédictible ?

6.1.1. Quelle distribution spatiale de l'activité virtuelle de crise dans les milieux métropolitains ?

6.1.1.1. Logique de spatialisation de l'activité virtuelle de crise en réponse à des phénomènes extrêmes récurrents

Dans le chapitre précédent, nous avons constaté que la logique de spatialisation de l'activité tweeting de crise ne semblait pas, dans tous les cas, se positionner comme l'écho de la spatialisation et de l'intensité d'un phénomène physique. Qu'observe-t-on alors à l'échelle locale, dans les milieux métropolitains qui se présentent comme les foyers de l'activité tweeting ? La figure 6.1 représente les tweets de crise géolocalisés et clustérisés émis dans l'aire métropolitaine de San Antonio entre le 16 et le 21 avril 2016 (dates qui correspondent à l'événement virtuel mis en évidence dans le chapitre précédent). Ces tweets sont

accompagnés de données témoins destinées à spatialiser les paramètres du phénomène physique :

- les cumuls pluviométriques du *NWS*, agrégés à 5 km, indiquent que seule la partie nord de l'aire métropolitaine a subi les précipitations les plus violentes (entre 74 et 143 mm d'eau, la normale pour le mois d'avril étant de 53 mm d'eau[1]) ;

- les tweets émis par les automates associés au réseau de stations de l'*USGS* se localisent principalement dans le centre et la moitié est de l'aire métropolitaine.

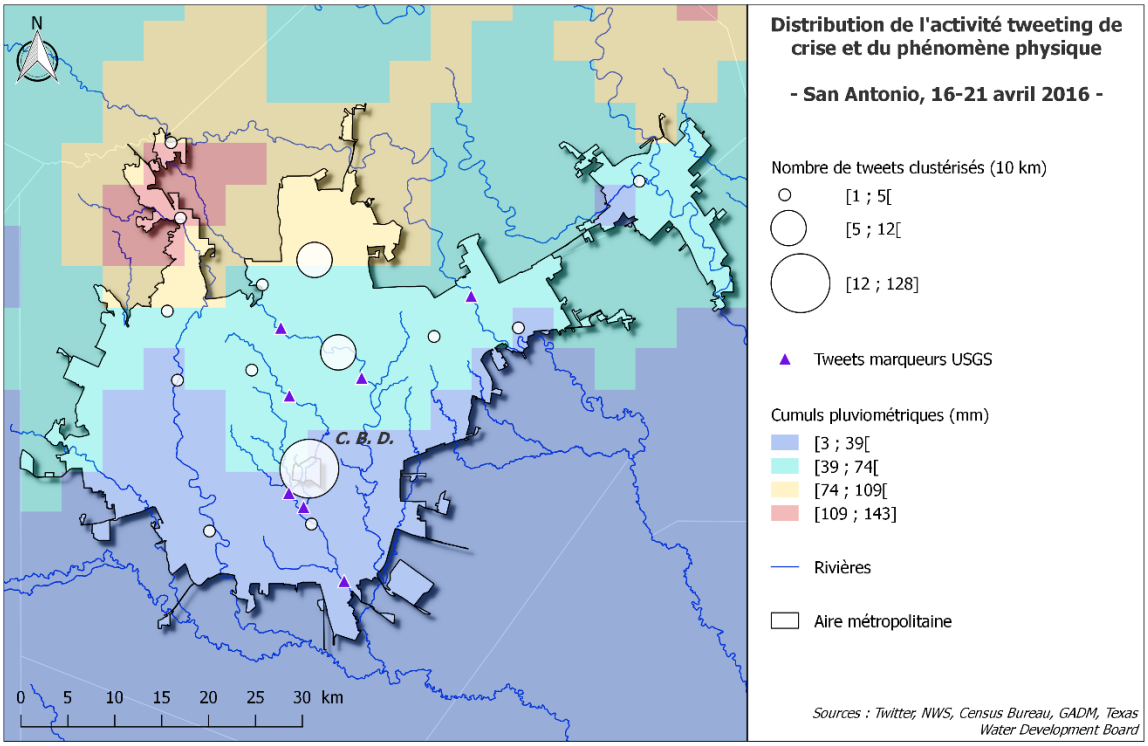


Figure 6.1 : Paramètres du phénomène physique et distribution des tweets de crise géolocalisés dans la métropole témoin de San Antonio, avril 2016 (C.Cavalière)

La correspondance entre tweets de crise géolocalisés et paramètres physiques du phénomène n'est pas évidente : 63,7% des tweets émis en réponse au phénomène sont concentrés dans le *CBD* (qui cumule entre 3 et 73 mm d'eau sur les six jours). A l'inverse, seuls 5,2% de ces tweets sont localisés dans les territoires enregistrant au moins 74 mm d'eau. Le tweet le plus proche d'un marqueur *USGS* de précipitations localement intenses survenues dans le dernier quart d'heure écoulé se situe à 927 m : il est émis le même jour que le tweet marqueur mais cinquante minutes après et le sens de son discours reste ambigu : "*Begin the flood of the Internet*". En outre, la distance moyenne des tweets de crise aux tweets marqueurs de l'*USGS* reste de 14 km et d'autre part, la moitié sud-sud-ouest de la métropole

[1] Source : https://en.wikipedia.org/wiki/San_Antonio#Climate (Consulté pour la dernière fois le 06/06/2019)

reste en marge de l'événement virtuel (les émissions de ce territoire enregistrées dans la période perturbée ne représentent que 9% des émissions totales de l'aire métropolitaine).

Dans l'aire métropolitaine de Houston, pour la même perturbation, les logiques de l'activité tweeting de crise s'avèrent plus contrastées (figure 6.2) : aucun territoire de l'agglomération ne se révèle totalement vide de tweets de crise bien qu'on observe toujours le même décalage entre intensité des précipitations et foyers les plus actifs en termes d'émissions.

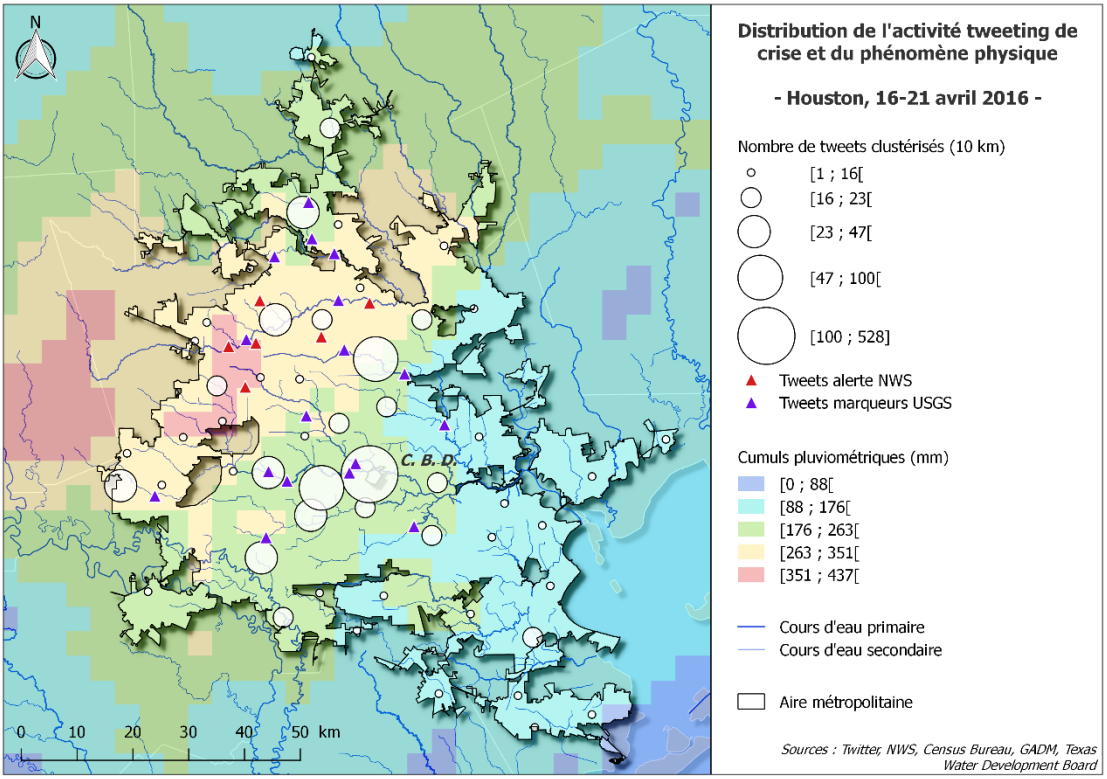


Figure 6.2 : Paramètres du phénomène physique et distribution des tweets de crise géolocalisés dans la métropole d'étude de Houston, avril 2016 (C.Cavalière)

A Houston, les zones enregistrant les précipitations les plus intenses (soit plus de 351 mm d'eau) ne contiennent que 2% des émissions de tweets. Les émissions les plus conséquentes sont perceptibles dans la zone cumulant entre 176 et 262 mm d'eau (soit un cumul toutefois deux à trois fois plus important que la normale du mois d'avril qui correspond à une hauteur d'eau de 84 mm[2]). Dans cette même zone, on trouve le *CBD* qui tout comme à San Antonio, concentre les émissions les plus volumineuses dans sa proximité directe. On observe quelques foyers dans la partie nord de l'agglomération mais la partie est-sud-est reste globalement moins inscrite sur le réseau (les tweets de crise de ces territoires moins actifs représentent 10% de l'activité virtuelle globale de l'aire métropolitaine).

[2] Source : https://en.wikipedia.org/wiki/Climate_of_Houston (Consulté pour la dernière fois le 06/06/2019)

Par ailleurs, tout comme à San Antonio, la proximité d'un tweet de crise géolocalisé quelconque à un tweet d'alerte émis depuis un compte officiel ne constitue pas une réponse virtuelle à cette alerte (tableau 6.1). A Houston, le tweet le plus proche d'une alerte émise par le compte du *NWS* se trouve à 600 mètres ; bien qu'il témoigne directement d'un nouvel épisode pluvieux qui semble inquiéter son auteur, il est émis trois jours après l'alerte. Cette situation se répète en ce qui concerne le tweet le plus proche d'une alerte *USGS* aux pluies intenses : celui-ci se trouve localisé à plus de 300 mètres du tweet d'alerte, est émis presque deux jours après l'alerte et n'évoque pas les conditions météorologiques en cours. On constate donc de nouveau l'existence probable d'un épisode silencieux pendant la situation de crise.

Tableau 6.1 : Caractéristiques des tweets les plus proches des alertes virtuelles

Source de l'alerte	Distance du tweet le plus proche (m)	Texte	Intervalle temporel (heures)	Distance moyenne des autres tweets (km)
NWS	601	*No no not this rain again today*	78,26	36,5
USGS	321	*Ray of hope after the great flood in Houston #Houston #Texas #flood #busstop #metro*	46,56	33,3

6.1.1.2. Logiques de spatialisation de l'activité de crise en réponse à des phénomènes extrêmes rares

Dans le cas du phénomène rare, observe-t-on les mêmes tendances de répartition spatiale de l'événement virtuel ? En d'autres termes, constate-t-on une répétitivité ou une variabilité des espaces de l'activité tweeting de crise en fonction des types de phénomènes ? La cartographie de l'activité virtuelle de crise agrégée en cluster (distance de 10 km) révèle la tendance de concentration des foyers d'activité les plus conséquents dans des lieux identiques (figure 6.3).

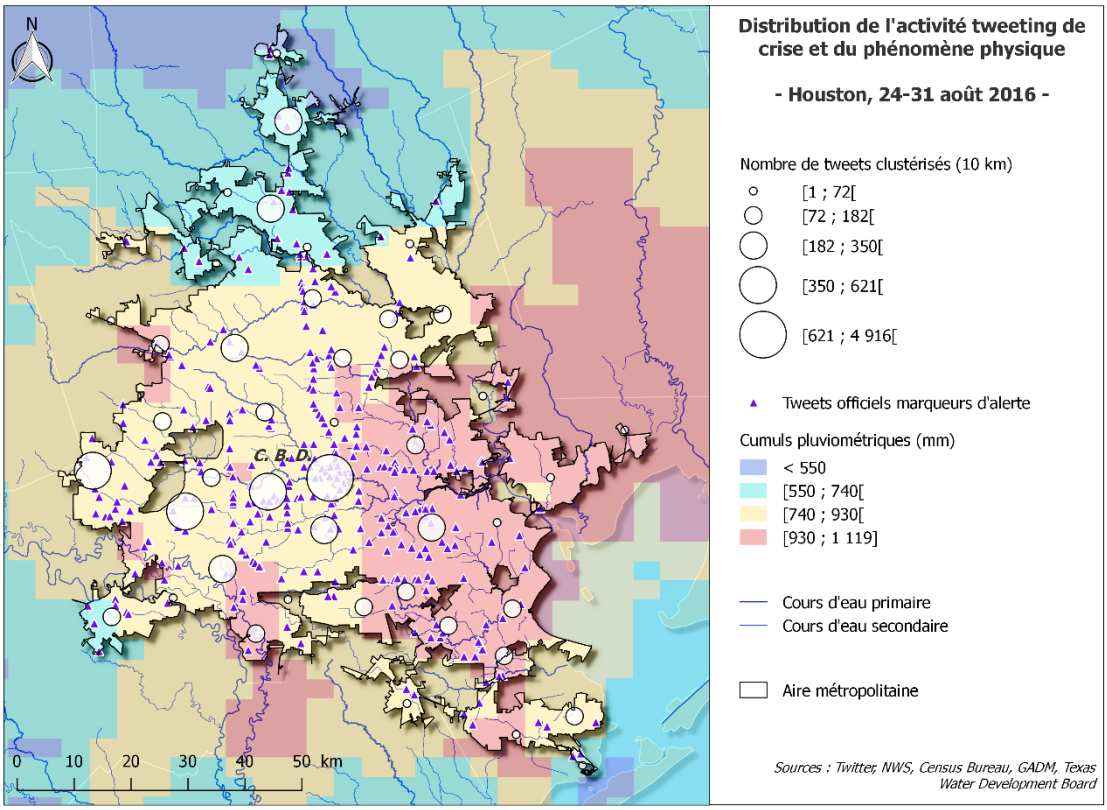


Figure 6.3 : Paramètres du phénomène physique et distribution des tweets de crise géolocalisés dans la métropole d'étude de Houston, ouragan Harvey – 2017 (C.Cavalière)

On retrouve en effet deux foyers majeurs d'émission de crise dans le *CBD* et sa périphérie occidentale, des foyers secondaires dans les parties nord et ouest de l'aire métropolitaine, ainsi qu'une activité plus ténue dans l'est et le sud-est. Les espaces ayant été touchés par les cumuls de précipitations records (cumul ≥ 930 mm d'eau en sept jours) enregistrent cette fois 20,8% de l'activité tweeting.

L'analyse de l'aire métropolitaine par mailles à une échelle spatiale plus fine (taille de la maille définie à 1 km de côté) permettra néanmoins de préciser ce constat (tableau 6.2) : à grande échelle, plus de 80% des mailles restent vides de tweets géolocalisés, quel que soit le phénomène considéré. Seules 3% des mailles tweetent de manière systématique, c'est-à-dire qu'elles enregistrent une activité tweeting à la fois pendant la perturbation d'avril 2016 et pendant le passage de l'ouragan. Si ces mailles ne constituent qu'une minorité des individus spatiaux, elles concentrent en revanche la majorité des émissions de tweets de crise géolocalisés (85% pendant la perturbation d'avril 2016 et 63% des émissions pendant l'ouragan). La réponse virtuelle au phénomène rare s'avère néanmoins plus diffuse que la réponse au phénomène récurrent : 13,6% des mailles actives apparaissent pendant l'ouragan Harvey et celles-ci cumulent tout de même 37% de l'ensemble des tweets de crise géolocalisés non rattachés à des acteurs institutionnels.

Tableau 6.2 : Comparaison de l'activité détectée dans les mailles de 1 km de côté

Mailles actives en avril 2016 et en août 2017		**Mailles invisibles pendant l'ouragan 2017**	**Mailles actives apparues pendant l'ouragan, août 2017**
3%		83,47%	13,60%
% Tweets inclus dans ces mailles, 16-21/04/2016	% Tweets inclus dans ces mailles, 24-31/08/2017	**Mailles invisibles en avril 2016 et pendant l'ouragan 2017**	% Tweets inclus dans les nouvelles mailles actives, ouragan 2017
85%	63%	81,82%	37%

A ce niveau de l'analyse, il est difficile de confirmer l'existence d'une logique de spatialisation de l'activité tweeting propre aux périodes de crise : à l'échelle de l'aire métropolitaine, la réponse virtuelle aux phénomènes et événements du réel semble indépendante de l'intensité physique. En outre, on observe une absence d'agitation marquée dans les lieux qui concentrent les tweets officiels témoins de l'alerte. Face à ces résultats, nous proposons de tester un nouveau facteur susceptible d'expliquer la distribution spatiale des tweets géolocalisés de crise : la spatialisation de l'activité virtuelle normale (c'est-à-dire l'ensemble des émissions virtuelles géolocalisées pendant une période donnée). L'activité de crise dépend-elle de l'existence d'une activité virtuelle en temps normal (auquel cas, l'activité de crise s'annoncerait comme une activité "opportuniste[3]") ?

6.1.2. Identification des facteurs de l'activité virtuelle dans les milieux métropolitains

6.1.2.1. Spatialisation de l'activité virtuelle globale dans les aires métropolitaines

Activité virtuelle globale de San Antonio, avril 2016.

Dans un premier temps, nous avons représenté la distribution spatiale des densités de tweets géolocalisés émis dans l'aire métropolitaine de San Antonio en avril 2016 (figure 6.4 en page suivante).

[3] Autrement dit, l'activité virtuelle de crise se trouve localisée dans les lieux polarisant l'activité virtuelle habituelle et on observe simplement une modification des sujets de tweeting.

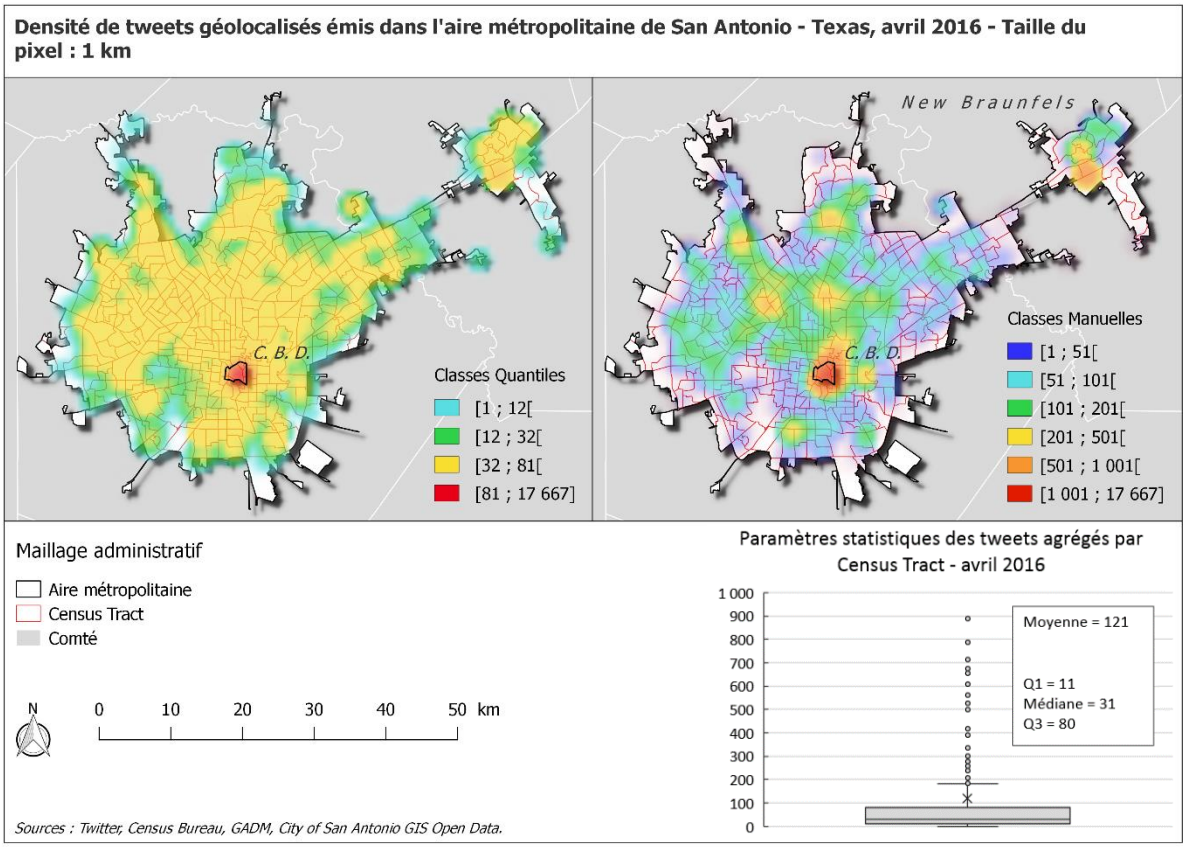


Figure 6.4 : Densités de tweets géolocalisés dans l'aire métropolitaine de San Antonio et paramètres de distribution statistique - avril 2016 (C.Cavalière)

Pour ce mois d'avril 2016, ce territoire enregistre un total de 48 162 tweets géolocalisés dont 35% sont concentrés dans le *CBD*. La carte de chaleur représentant les densités de tweets géolocalisés indique en effet, quelle que soit la méthode de discrétisation utilisée, l'existence d'un foyer principal de concentration des émissions dans les Census Tracts recoupant le *CBD* alors que les émissions diminuent drastiquement lorsqu'on s'éloigne de ce centre : elles deviennent quasi nulles dans les marges de l'aire métropolitaine (bien qu'on constate également des lieux vides de tweets au sein de l'aire métropolitaine, en particulier dans le nord-est et le sud-ouest). Une discrétisation manuelle plus fine permet néanmoins de vérifier l'existence de foyers d'émissions secondaires, dont le plus actif se situe dans la terminaison nord-est de l'aire métropolitaine, dans la ville de New Braunfels : l'activité virtuelle reste cependant plus discrète que dans le *CBD* puisque ce foyer ne contient que 5% des émissions totales de tweets géolocalisés. En outre, après vérification, cette activité apparaît liée aux émissions d'un compte automatique, vraisemblablement associé à une station météorologique[4].

[4] Il émet des messages de ce type : "*Wind 4.3 MPH SSE. Barometer 29.849 IN, rising slowly. Temperature 75.2 °F. Rain today 0.00in. Humidity 86%*".

Les paramètres statistiques indiquent toujours la même tendance : la moyenne reste influencée par les valeurs extrêmes de trente-sept Census Tracts (soit 9,3% des unités de l'aire métropolitaine) qui concentrent 70% de l'ensemble des tweets géolocalisés alors que 75% de ces Census Tracts contiennent tout au plus 80 tweets géolocalisés.

Activité virtuelle globale de Houston, avril 2016.

Nous recherchons ici une éventuelle répétitivité des tendances observées sur l'aire métropolitaine témoin en exécutant des analyses identiques sur l'aire métropolitaine d'étude de Houston. En avril 2016, Houston enregistre un total de 141 334 tweets géolocalisés (soit trois fois plus qu'à San Antonio) émis dont 18,67% se trouvent dans le CBD. Les foyers d'émission de tweets géolocalisés sont ici plus diffus : en effet, si l'on observe la carte des densités de tweets géolocalisés (figure 6.5), le CBD apparaît certes toujours comme unique haut-lieu de l'activité tweeting, mais la classification manuelle révèle un foyer majeur qui se diffuse du CBD vers le nord et l'ouest de l'aire métropolitaine ainsi que des foyers ponctuels secondaires localisés dans l'ensemble des marges de cette même aire.

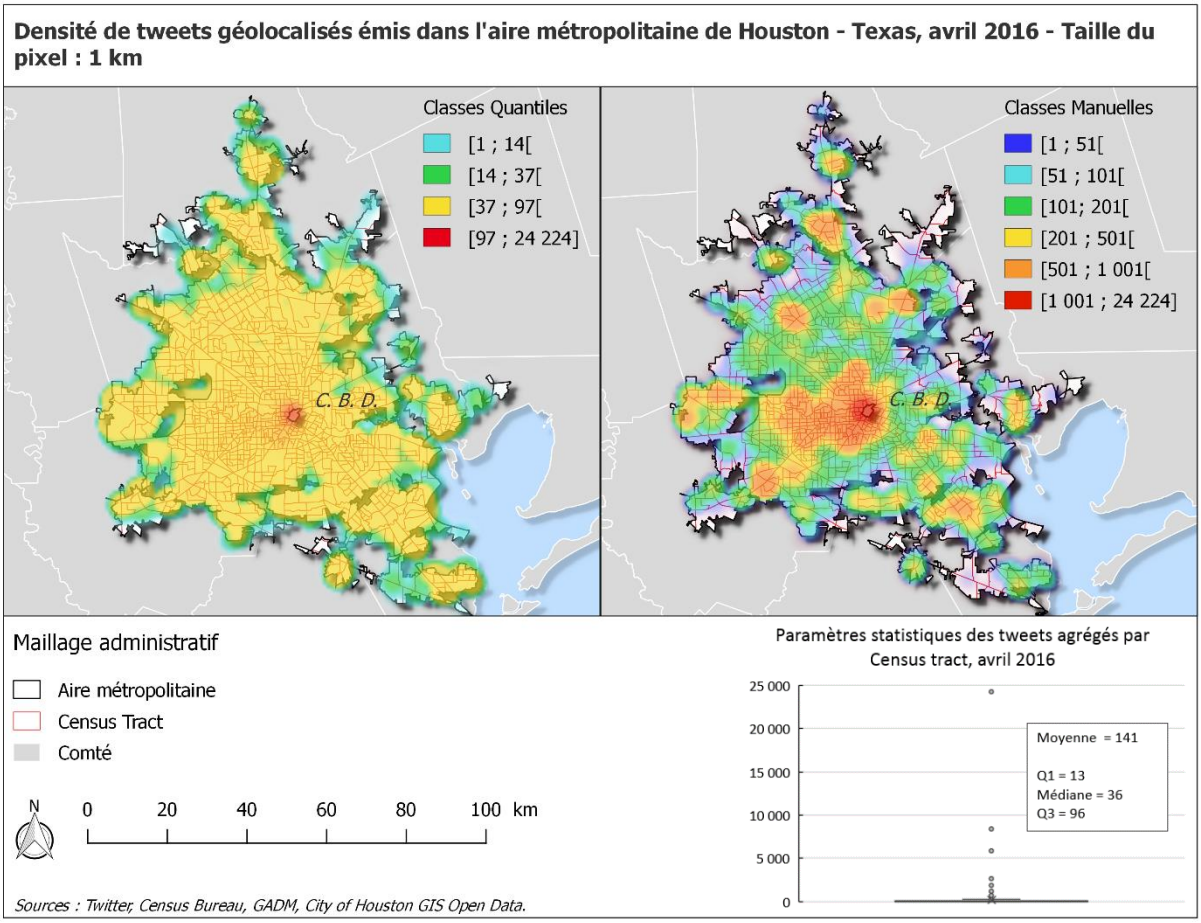


Figure 6.5 : Densités de tweets géolocalisés dans l'aire métropolitaine de Houston et paramètres de distribution statistique - avril 2016 (C.Cavalière)

Pour autant, les paramètres statistiques restent ancrés dans la tendance générale : 75% des Census Tracts contiennent tout au plus 96 tweets géolocalisés alors que 109 individus hors-normes (soit 10,88% des Census Tracts de l'aire métropolitaine) concentrent 68,93% de l'ensemble des émissions de tweets géolocalisés en avril 2016.

Des cellules d'activité virtuelle identiques en fonction de la situation locale ?

Les lieux enregistrant une activité tweeting géolocalisée en situation normale sont-ils alors ceux dans lesquels on enregistre une activité virtuelle de crise ? Le tableau 6.3 indique en effet cette tendance : en avril 2016, aucune cellule inactive en temps normal ne s'est réveillée pendant la période perturbée.

Tableau 6.3 : Comparaison des structures de l'activité tweeting en temps normal et en période de crise

Activité virtuelle par CT	San Antonio	Houston
% de CT actifs en temps normal	96,4	95,9
% de CT actifs pendant la crise	12,3	29,8
% de CT actifs en permanence	12,3	29,8
% de CT invisibles en permanence	3,5	4,1

A San Antonio, tout comme à Houston, l'ensemble des Census Tracts dans lesquels on détecte une activité de crise correspondent aux unités d'ores-et-déjà actives en temps normal[5]. De même, on constate de nouveau qu'une minorité des individus concentrent une majorité des tweets de crise géolocalisés : à San Antonio, 98 Census Tracts (soit 25%) affichent des émissions de tweets géolocalisés supérieures à la valeur du troisième quartile en situation normale (Q3 = 80). Seules deux de ces 98 unités concentrent 68% des émissions de tweets de crise géolocalisés pour l'événement virtuel du 16 au 21 avril 2016. A Houston, 250 Census Tracts (soit de nouveau 25% des Census Tracts) affichent des émissions de tweets géolocalisés supérieures à 97 (valeur de Q3) en situation normale. Pendant le même événement virtuel, 14 de ces Census Tracts très actifs en temps normal cumulent déjà 49% des émissions de tweets géolocalisés.

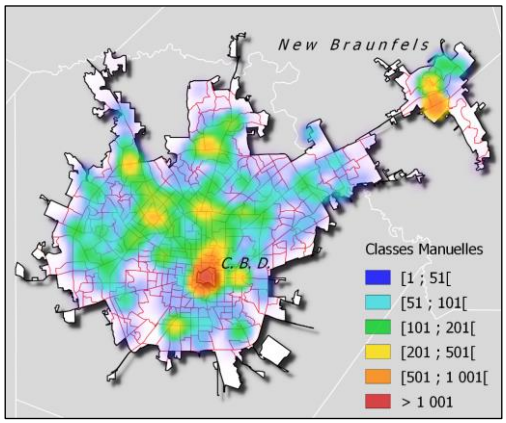
[5] Le test n'a pas été mené pour l'événement virtuel lié à l'ouragan Harvey car nous n'avons pas collecté l'ensemble des tweets bruts émis à l'été 2017 et il est possible que les structures de l'activité virtuelle brute du mois d'avril 2016 aient pu évoluer dans l'intervalle de 14 mois.

6.1.2.2. Recherche de facteurs explicatifs de la distribution des tweets géolocalisés

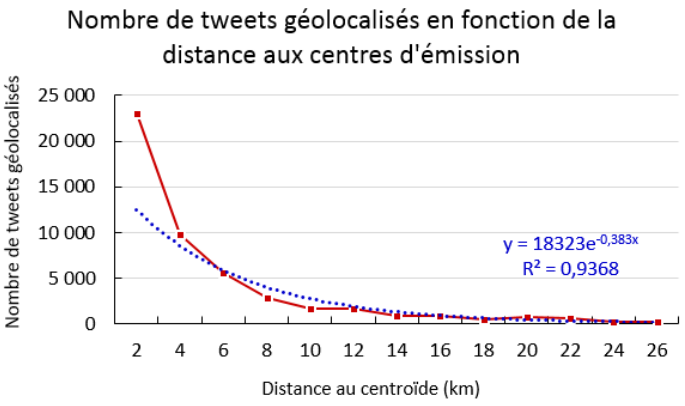
Si les lieux de réactivité qui concentrent les émissions de tweets géolocalisés en période de crise correspondent aux lieux d'activité normale, peut-on mettre en évidence l'existence de facteurs explicatifs de la distribution de l'activité tweeting normale ? Si oui, peut-on également les observer dans la distribution des tweets de crise géolocalisés ?

Caractérisation de l'activité tweeting normale en fonction de lieux, San Antonio.

Nous focalisons ici notre attention sur les lieux qui captent l'essentiel de l'activité tweeting. En effet, l'observation des densités de tweets géolocalisés de San Antonio laissait transparaître des centres concentrant l'essentiel de l'activité virtuelle ainsi qu'une décroissance de cette activité en fonction de l'éloignement au centre (cf. figure 6.4, reportée en page suivante dans la figure 6.6.a). On pourra alors émettre une nouvelle hypothèse relative aux logiques de distribution spatiale des tweets géolocalisés, fondée sur un modèle centre/périphérie. Le test est effectué à partir des huit foyers de l'activité tweeting métropolitaine, identifiés dans la figure 6.4. Un foyer d'activité tweeting est défini comme suit : d'après la partition établie et afin d'atténuer l'effet aberrant du *CBD*, les foyers correspondent aux pixels de 1km² concentrant au moins 201 tweets géolocalisés. Pour chaque semis de tweets inclus dans un foyer, on en génère le centre de gravité ; enfin, on calcule le nombre de tweets émis dans un rayon (pas de deux kilomètres) autour de chaque centroïde. La figure 6.6.b présente la courbe résultante du nombre de tweets géolocalisés émis en fonction de la distance au centre de gravité de chaque foyer d'émission.



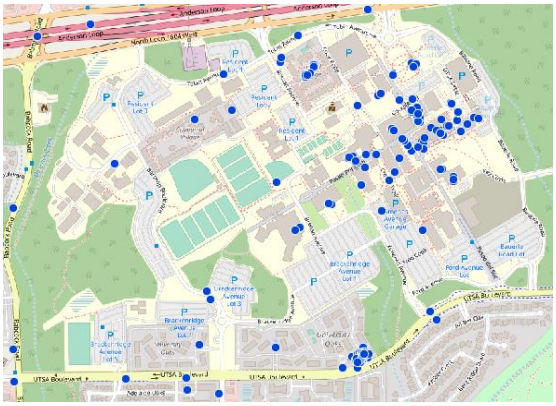
(a). Densités de tweets géolocalisés émis dans l'aire métropolitaine de San Antonio en avril 2016



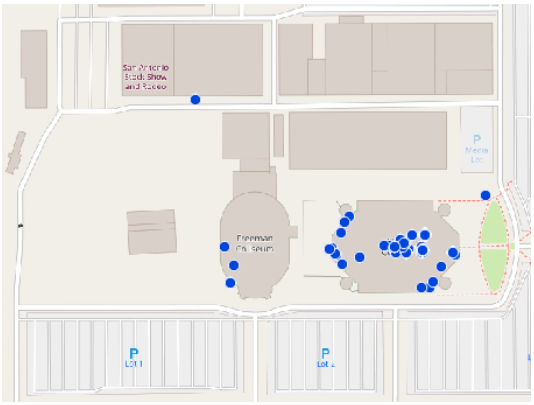
(b). Modélisation du nombre de tweets de crise géolocalisés émis en fonction de la distance à un foyer de l'activité virtuelle géolocalisée, San Antonio, avril 2016

Figure 6.6 : Comportement spatial de l'activité tweeting géolocalisée en fonction des foyers identifiés dans l'aire métropolitaine de San Antonio (C.Cavalière)

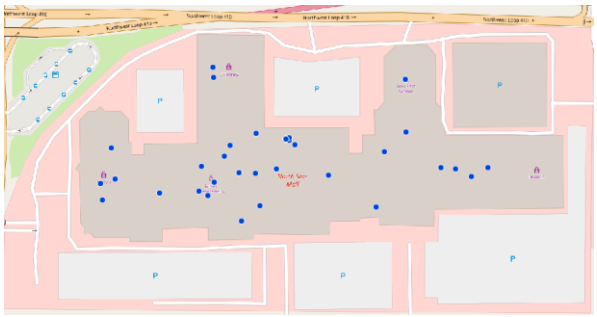
La courbe témoigne bien d'une décroissance progressive des émissions de tweets en fonction de la distance au centre ; en fait, on peut mettre en exergue l'existence d'une relation exponentielle décroissante permettant de prédire les quantités de tweets géolocalisés en fonction de la distance à un centre, d'équation $y = 18\,232e^{-0,383x}$ et dont la qualité d'ajustement s'avère forte (R^2 = 93%). Ces lieux concentrant l'activité sont-ils particuliers ? Dans chaque foyer d'émission identifié, on peut en effet distinguer des objets qui polarisent les émissions de tweets (figure 6.7) comme les établissements de santé, les universités et les objets et lieux de culture et de loisirs (théâtres, musées, complexes sportifs, parcs, *etc.*), quelle que soit la localisation du foyer.



(a). Campus universitaire, nord-ouest du CBD



(b). Complexe sportif, est du CBD

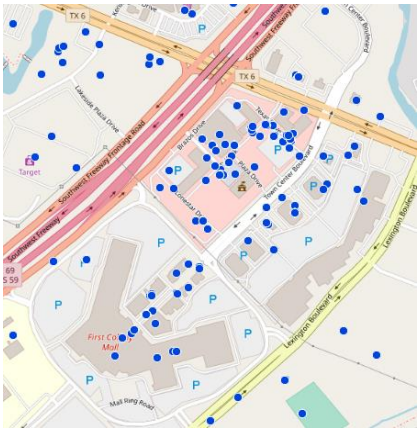


(c). Centre commercial, nord de la métropole

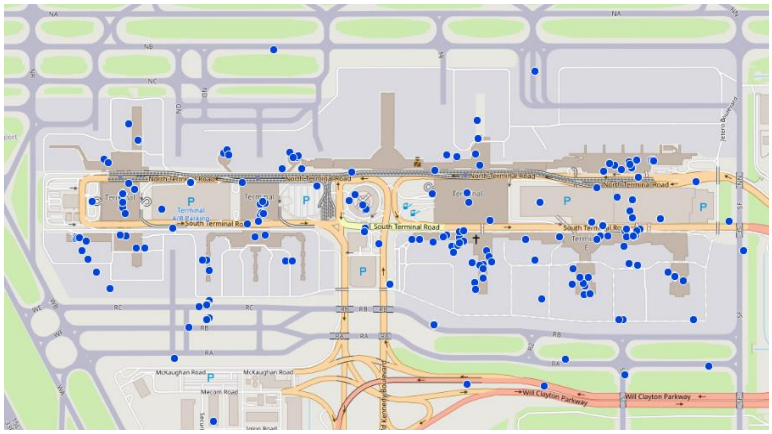
Figure 6.7 : Formes de l'activité tweeting normale, San Antonio, avril 2016 (fond de carte OpenStreet Map)

Caractérisation de l'activité tweeting normale en fonction d'un réseau, Houston.

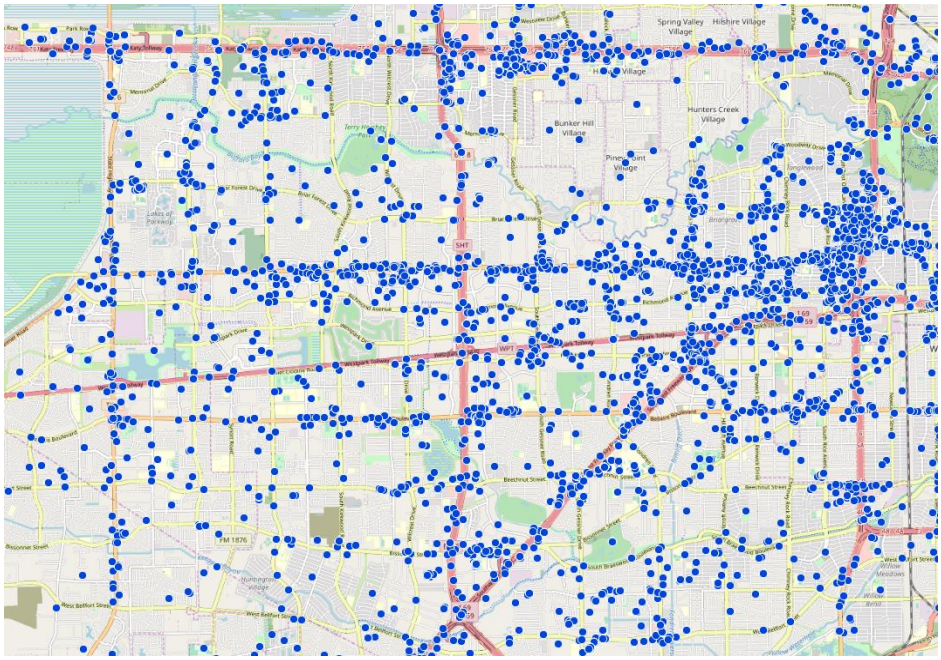
A Houston, la situation se révèle plus complexe qu'à San Antonio : une logique centre-périphérie se manifeste dans les foyers détectés dans les marges de l'aire métropolitaine mais, dans son centre, celle-ci s'avère imperceptible. Dans les foyers périphériques, on peut identifier, tout comme à San Antonio, des lieux d'intérêts qui concentrent l'activité tweeting normale (cf. figure 6.8.a et 6.8.b : l'aéroport intercontinental, les centres commerciaux, les établissements de santé, *etc.*) mais en ce qui concerne le foyer d'activité le plus conséquent, qui englobe tout le centre-ouest de la métropole, une nouvelle logique se manifeste : l'activité tweeting semble se concentrer le long des axes routiers (figure 6.8.c).



(a). Centre commercial, sud-ouest du CBD



(b). Aéroport Intercontinental George Bush, nord-est du CBD

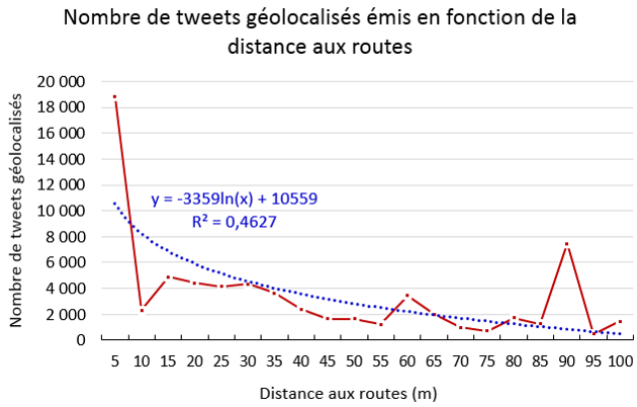


(c). Activité dans le voisinage des axes routiers, centre-ouest du CBD

Figure 6.8 : Formes de l'activité tweeting normale, Houston, avril 2016 (fond de carte OpenStreet Map)

Dans le centre-ouest de l'aire métropolitaine de Houston, nous testons alors l'effet polarisant des routes sur l'activité tweeting en temps normal, en sélectionnant un pas de distance de cinq mètres, dans les cent premiers mètres de part et d'autre des routes. Le problème reste que les données TIGER utilisées afin de représenter ces routes, sont constituées de traits simples : en conséquence, quatre séries de quatre lignes simples ne modélisent pas fidèlement l'espace plein que représente, dans le réel, une *Interstate* américaine de quatre séries de quatre voies (soit entre 90-95 mètres de large). Les grandes avenues qui connectent l'ensemble des quartiers sont larges de 30 à 40 mètres et pour finir, les rues qui parcourent l'intérieur des quartiers du centre de la métropole sont larges d'une quinzaine de mètres. Les résultats sont consignés dans la figure 6.9 (le graphique [a] affiche

les effectifs de tweets géolocalisés en fonction du pas de distance de cinq mètres ; le tableau [b] présente les fréquences cumulées croissantes des effectifs de tweets géolocalisés relevés en fonction de la distance aux routes, agrégés sur un pas de dix mètres).



(a). Modélisation du nombre de tweets de crise géolocalisés émis en fonction de la distance à une voie de circulation, centre de Houston, avril 2016

Distance à la route (m)	Fréquences cumulées croissantes
≤ 10	30,6%
≤ 20	44,1%
≤ 30	56,4%
≤ 40	65,1%
≤ 50	69,9%
≤ 60	76,7%
≤ 70	81,0%
≤ 80	84,5%
≤ 90	97,2%
≤ 100	100%

(b). Fréquences cumulées croissantes des effectifs de tweets géolocalisés en fonction de la distance aux routes, pas de dix mètres, centre de Houston, avril 2016

Figure 6.9 : Activité tweeting normale à Houston, en fonction de la distance aux routes, avril 2016

Dans le cas du centre ville de Houston, on peut tenter de modéliser une relation d'ordre logarithmique négative qui indiquerait une décroissance de l'activité tweeting en fonction de la distance à la route. Pour autant, le réseau routier ne constitue pas l'unique facteur explicatif de la distribution des tweets géolocalisés sur centre ($R^2 = 0,46$ donc la relation est de moindre intensité que le modèle mis en évidence précédemment à San Antonio) ; en effet, on constate des pics de tweets à certaines distances, le plus fort étant détecté à une distance de 80 à 90 mètres des routes. Après exploration, cette poche d'activité est vraisemblablement due à la présence d'un objet particulier du territoire, en l'occurrence le *Houston Police Officers Memorial*, foyer d'activité virtuelle géolocalisée concentrant 7 902 tweets (soit 95% des tweets situés entre 80 et 90 mètres des routes). Notons tout de même que 51,2% des tweets géolocalisés émis en avril à Houston sont situés à moins de 100 mètres d'une route et que parmi ces tweets, 56,4% sont localisés à moins de 30 mètres d'une route. Au final, dans le cas de Houston, deux modèles explicatifs de l'activité virtuelle géolocalisée coexistent sans doute : une logique centre/périphérie autour d'objets polarisant les émissions et une logique de distribution linéaire articulée autour du réseau routier.

6.1.2.3. Quelles logiques pour l'activité virtuelle géolocalisée de crise ?

Deux logiques semblent donc prévaloir, à cette échelle, dans la localisation de l'activité tweeting normale : la présence d'objets d'intérêts qui polarisent les émissions ainsi que les axes de communication. Les modèles mis en évidence sont-ils alors transposables à l'activité virtuelle de crise ? Précédemment, nous avons vu que les lieux qui réagissent en période de crise correspondent aux lieux qui sont déjà actifs au quotidien. Mais peut-on modéliser une logique de spatialisation des tweets de crise géolocalisés, non pas en fonction de l'intensité (comme nous avons vu que ce paramètre n'avait pas de réponse virtuelle quantitative significative) mais d'objets d'intérêt relatifs à la crise sur les territoires métropolitains ?

Nous avons effectué un test sur la métropole de Houston, pour laquelle nous avons pu télécharger un jeu de données non fondé sur l'intensité : il s'agit de la localisation des sites ouverts (refuges, aide médicale, pharmacie, ravitaillement, soins aux animaux domestiques, lieux de recrutement de volontaires, *etc.*) où les habitants pouvaient se protéger ou obtenir soins et vivres pendant le passage de l'ouragan Harvey. Le pas de distance retenu est ici de dix mètres (on ne trouve aucun tweet de crise géolocalisé à moins de cette distance des sites). La figure 6.10 présente ainsi les effectifs de tweets de crise géolocalisés en fonction de la distance aux sites d'aide et de refuge :

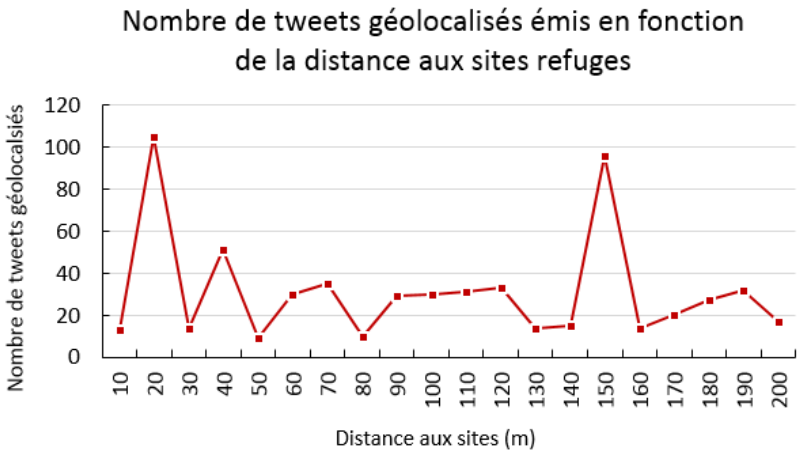


Figure 6.10 : Activité virtuelle de crise, en fonction de la distance aux sites refuges, ouragan Harvey

Dans la métropole en situation de crise, il s'avère en fait que les objets d'intérêt liés à la crise testés ne constituent pas des facteurs polarisant de l'activité tweeting : d'après la figure 6.10, il n'existe aucune relation apparente entre les émissions de tweets géolocalisés de crise et la distance à un site refuge. En outre, seul 0,56% des tweets géolocalisés se trouvent à une distance inférieure ou égale à 200 mètres de ces sites. Face à ces résultats qui semblent indiquer des difficultés à généraliser les résultats d'un test, répété dans le même lieu mais dans un contexte différent, nous testons simplement l'existence, pour l'activité virtuelle de

crise liée à l'ouragan, d'une dépendance spatiale de la distribution des tweets géolocalisés. Le test d'autocorrélation est mené à l'échelle des *census tracts* (dans une unité statistiquement homogène afin d'éviter les risques de biais induits par une agrégation spatiale des tweets par maille). La visualisation de l'existence éventuelle d'une structuration spatiale des valeurs exprimant le nombre de tweets de crise dans chaque *census tract* par le diagramme de Moran n'est pas évidente (figure 6.11). On repère d'emblée les individus aux valeurs fortes par rapport à la valeur moyenne dans un voisinage qui leur ressemble mais, en raison de la grande quantité de valeurs proches de 0, la lisibilité graphique reste médiocre et le résultat difficile à interpréter :

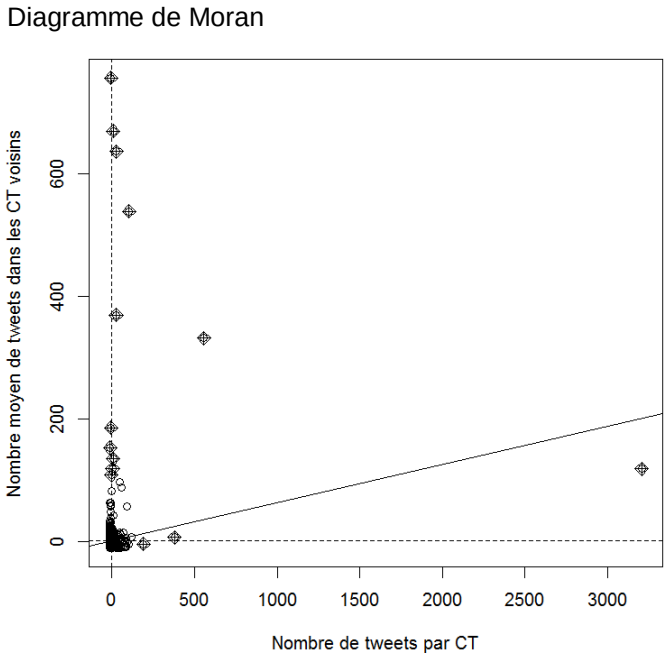


Figure 6.11 : Diagramme de Moran, test d'autocorrélation spatiale

Néanmoins, le calcul de l'indice global de Moran confirme l'hypothèse nulle : $I_w = 0{,}062$. On peut alors considérer la distribution des tweets géolocalisés de crise comme une distribution aléatoire. Le test est toutefois répété en retirant les individus aux valeurs extrêmes (seuls les *census tracts* qui regroupent moins de cent tweets de crise géolocalisés sont conservés pour ce second test). L'absence de structure particulière dans la distribution des valeurs est ici davantage perceptible (figure 6.12).

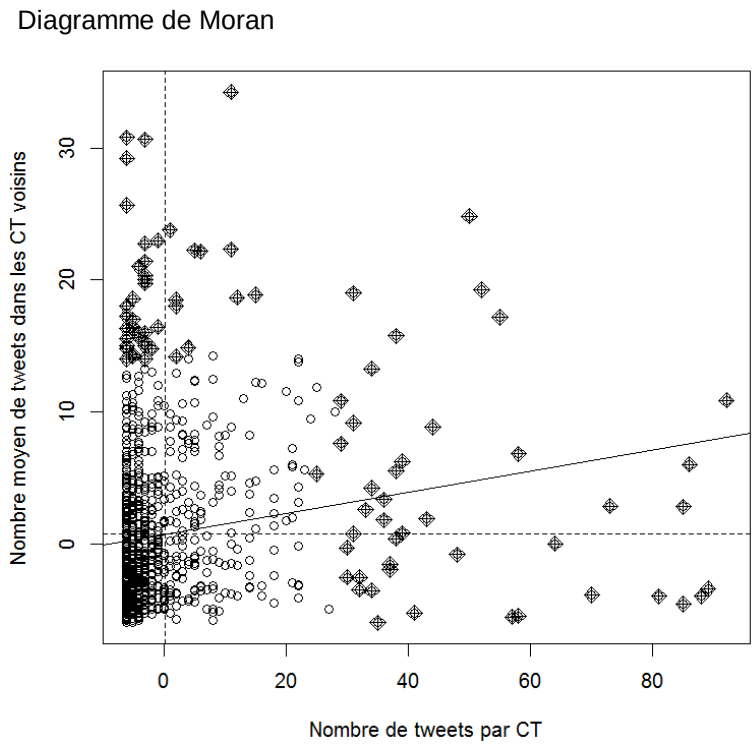


Figure 6.12 : Diagramme de Moran, test d'autocorrélation spatiale sans les individus présentant des valeurs extrêmes

Le calcul de l'indice de Moran confirme de nouveau l'existence d'une distribution aléatoire, indépendante des valeurs dans les unités de voisinage : I_w = 0,08.

Bilan des analyses relatives à la spatialisation des tweets géolocalisés

A ce stade des analyses, la composante spatiale des tweets géolocalisés n'apparaît toujours pas prédictive, quel que soit le contexte (normal ou perturbé) du terrain d'étude. On ne peut ni modéliser, ni standardiser des comportements spatiaux liés à l'utilisation du média social géolocalisé : il apparaît en effet que les lieux actifs en temps normal sont enclins à réagir par temps de crise mais dans des proportions variables (pour rappel, lors du phénomène de pluies intenses/inondations survenu en avril 2016, seuls 12,3% des *census tracts* actifs participaient à l'événement virtuel de crise à San Antonio alors qu'ils étaient 29,8% à Houston). Mais surtout, les facteurs de spatialisation de l'activité tweeting qu'on peut identifier en temps normal (objets d'intérêt locaux, réseaux de communication) semblent caducs en temps de crise.

Bilan des analyses relatives à la spatialisation des tweets géolocalisés (suite)

En conséquence, si l'on peut identifier des facteurs explicatifs de la localisation de l'activité virtuelle géolocalisée dans un contexte précis, il s'avère périlleux de les généraliser dans un contexte différent. Par ailleurs, on peut toujours s'interroger sur la validité de la première loi de Tobler (lorsqu'on compare objets du territoire en crise et tweets de crise géolocalisés, il s'avère que ces objets ne polarisent pas l'activité numérique) et sur l'adaptabilité des méthodes statistiques traditionnelles au caractère imprévisible des traces numériques géolocalisées émises dans différents contextes.

6.1.3. Définir des profils de populations productrices de l'activité virtuelle géolocalisée

6.1.3.1. Populations productrices de contenus géolocalisés de la métropole témoin, San Antonio

Si l'on ne peut pas caractériser de manière systématique les facteurs de réactivité et de sensibilité qui favorisent l'émergence de tweets géolocalisés en des lieux précis des territoires, peut-on au moins, sur notre terrain d'étude, déterminer les caractéristiques socio-démographiques des populations productrices de tweets géolocalisés de l'aire métropolitaine ? Pour tenter d'éclairer ce point, nous avons recherché l'existence éventuelle de variables influençant non pas la localisation des tweets mais l'adhésion et la participation à la création de contenus géolocalisés sur Twitter. En effet, si le propre du tweet est d'être créé en situation de mobilité dans le cadre des activités quotidiennes, les données démographiques, sociales et économiques sont produites au domicile des individus. Nous n'avons donc pas utilisé les tweets mais défini la localisation supposée du domicile des utilisateurs, en ayant recours à la méthode décrite par (Kounadi et al., 2015), c'est-à-dire la recherche du barycentre des émissions de chaque utilisateur ayant tweeté au moins une fois entre 1h et 6h30 le matin, pendant le mois d'avril 2016 (soit 2 637 tweets géolocalisés pour 552 utilisateurs différents). Les variables socio-démographiques test sont les suivantes6 : le revenu médian par foyer, le nombre d'individus ayant un diplôme de licence (Bachelor) ou plus, le nombre d'individus ayant au moins un diplôme de licence et disposant d'un accès Internet à très haut débit, le nombre d'individus dont l'âge est compris entre 20 et 44 ans, le

[6] Les critères considérés dans le choix des variables prises en compte sont les suivants : les revenus par foyer (disposer d'un smartphone avec une connexion à l'Internet mobile pouvant représenter un certain coût financier) qui sont susceptibles d'être influencés par l'origine ethnique des individus, mais aussi par leur niveau de diplôme (qui lui-même influence le degré d'utilisation des technologies de l'information et de la communication ainsi que des plateformes du Web social [Eskenazi *et al.*, 2017]) ; et enfin, l'âge des individus, étant donné que plus de 50% des utilisateurs de Twitter aux Etats-Unis sont âgés de 18 à 45 ans (Source : https://www.statista.com/statistics/192703/age-distribution-of-users-on-twitter-in-the-united-states/ [Consulté le 6/11/2019]).

nombre d'individus disposant d'un smartphone ainsi que la répartition ethnique des individus (nombre d'individus d'origine caucasienne, afro-américaine, indienne et asiatique).

La cartographie des utilisateurs (figure 6.13) indique la même tendance générale que la spatialisation des tweets : les *census tracts* du *CBD* cumulent 37,5% des utilisateurs. En revanche, on identifie des foyers ponctuels secondaires d'utilisateurs en particulier dans le nord de la métropole. Malgré tout, les *census tracts* n'enregistrant qu'un ou deux utilisateurs représentent 72% du territoire.

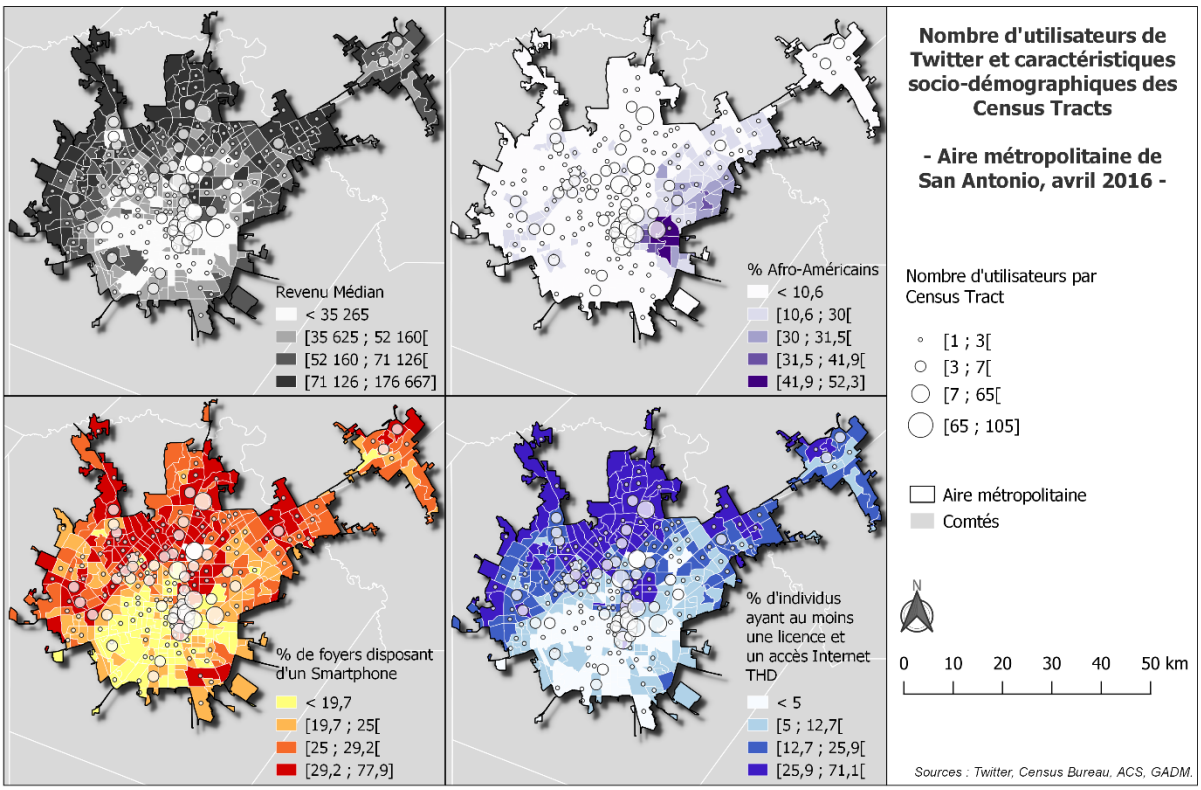


Figure 6.13 : Utilisateurs et caractéristiques socio-démagraphiques des Census Tracts, aire métropolitaine de San Antonio (C.Cavalière)

Les variables socio-démographiques représentées sur la figure 6.13 mettent en évidence les structures suivantes[7] :

- les plus hauts revenus sont concentrés dans la moitié nord de l'aire métropolitaine, et en particulier dans ses marges, ainsi que dans le *CBD*. Les populations moins favorisées se localisent plutôt au sud du *CBD*, dans une couronne d'une quinzaine de kilomètres ; on retrouve néanmoins des îlots de populations moins aisées dans la partie nord de la métropole.

- La répartition des individus disposant d'un smartphone et des individus titulaires d'au moins un diplôme de licence et disposant d'un accès à Internet très haut débit, suit cette

[7] L'âge n'est pas représenté sur la carte : les populations âgées entre 20 et 44 ans sont réparties de manière assez homogène dans l'ensemble des unités de recensement. En conséquence, on ne distingue aucune rupture spatiale marquée, contrairement aux revenus, niveaux de diplômes et accès aux TIC.

même logique : ceux-ci se concentrent davantage dans la partie nord de la métropole et dans le *CBD*, alors que le sud apparaît moins diplômé et moins connecté.

- La répartition ethnique des populations indique toujours cette tendance à la ségrégation des populations afro-américaines qui se concentrent plutôt dans l'est de la métropole.

- La localisation des utilisateurs ne révèle pas de rupture marquée en fonction des caractéristiques socio-démographiques des unités de recensement : 26% des utilisateurs vivent dans des unités dont le revenu médian par foyer s'élève à plus de 71 126$ en 2016 ; dans les unités dont le revenu médian est inférieur à 35 625$, on trouve tout de même 22% des utilisateurs. De la même manière, les unités dont au moins 29,2% des foyers disposent d'un smartphone concentrent 51,4% des utilisateurs mais on en trouve 12,3% dans les unités dont moins de 19,6% des foyers possèdent un smartphone.

Face à l'absence de rupture socio-spatiale marquée dans la répartition des utilisateurs, nous avons soumis l'ensemble de ces données à une ACP normée afin de vérifier ou d'invalider l'existence de liens entre profils socio-démographiques des individus et utilisateurs de Twitter (liens qui ne paraissent pas évidents sur la cartographie). Le résultat n'apporte que de minces précisions : notre attention s'est focalisée sur les quatre premières composantes, soit 81% de la variance totale (figure 6.14).

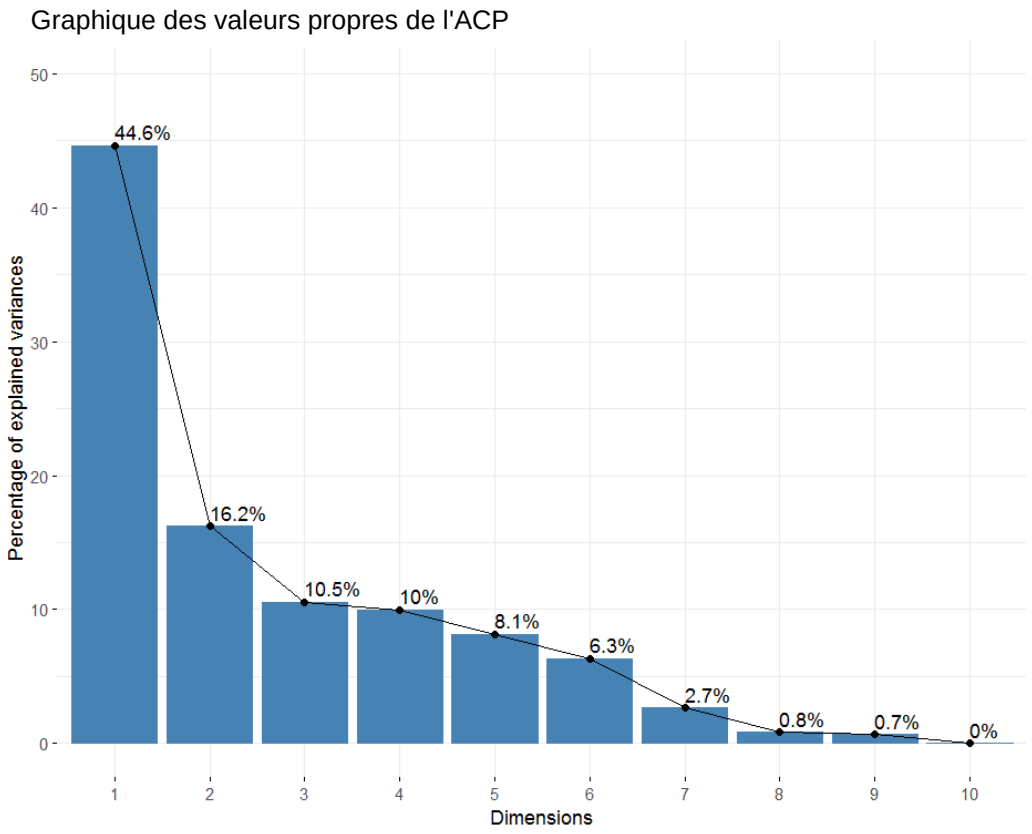


Figure 6.14 : Graphique des valeurs propres des composantes

La figure 6.15 représente ensuite la qualité de la représentation et la contribution des variables :

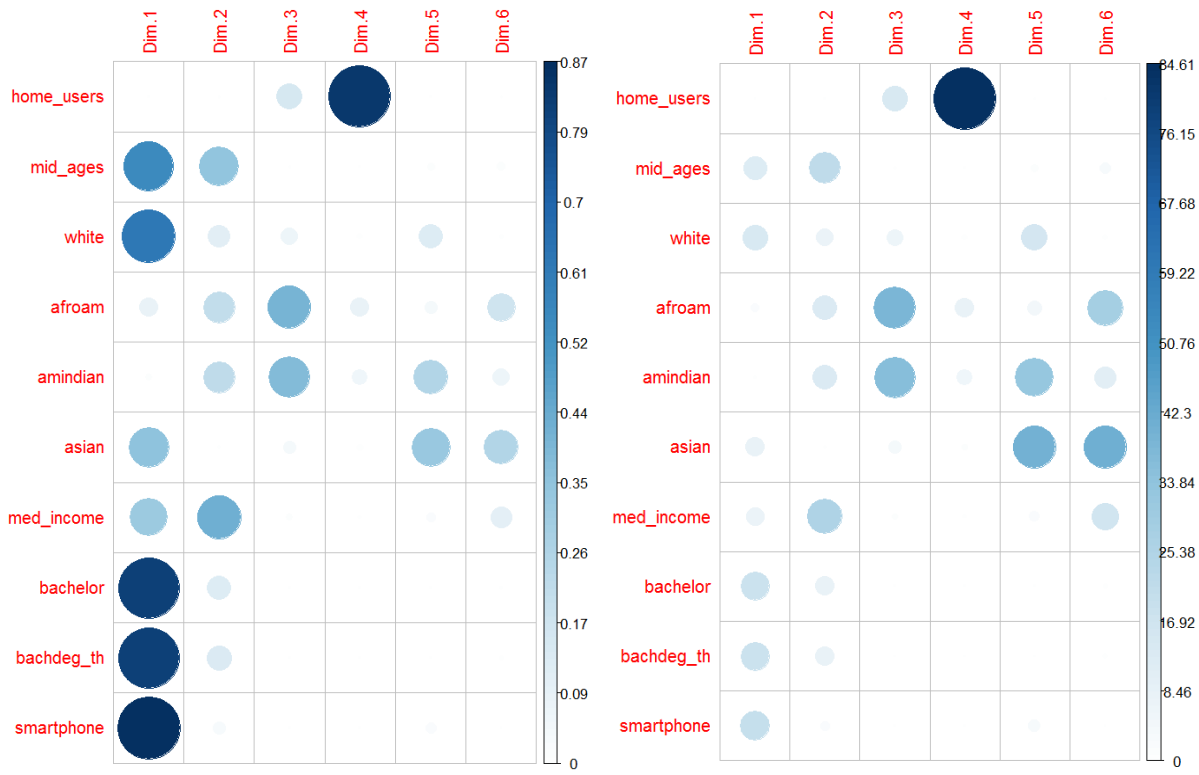


Figure 6.15 : Matrice de la qualité de représentation (cos², gauche) et des contributions (en %, droite) des variables aux composantes

Pour les quatre premières composantes, les valeurs des cos² indiquent que la variable *home_users* (soit le nombre supposé d'utilisateurs habitant dans une unité de recensement) est bien représentée sur la seule composante 4 : cos² = 0,84 ; elle contribue d'ailleurs à 84,6% à cette composante (figure 6.15). Sur cette même composante, on trouve également les variables *afroam* (nombre d'habitants d'origine afro-américaine) et *amindian* (nombre d'habitants d'origine amérindienne), mais d'une qualité de représentation et d'une contribution moindres : les valeurs des cos² sont respectivement de 0,07 et de 0,05 et les pourcentages de contribution, de 7,9 et de 5,5.

Les trois autres composantes retenues représentent (dans l'ordre décroissant des valeurs cos²) :

- composante 1 : les individus disposant d'un Smartphone (*smartphone*), les individus diplômés d'au moins une licence (*bachelor*) et les mêmes individus bénéficiant en plus d'une connexion très haut débit (*bachdeg_th*) ; la classe d'âge 20-44 ans (*mid_ages*) et la population blanche (*white*). Revenu médian (*med_income*) et population d'origine asiatique (*asian*) sont également perceptibles sur cet axe mais ne contribuent respectivement qu'à hauteur de 6 et 7%.

- composante 2 : le revenu médian par unité et la classe d'âge 20-44 ans. Diplômés/Diplômés disposant d'une connexion très haut débit et populations d'origines afro-américaine, amérindienne et blanche se retrouvent mais avec une qualité de représentation bien plus faible.

- composante 3 : trois variables se distinguent mais la qualité de leur représentation reste faible : populations d'origines afro-américaine ($cos^2 = 0,4$) et amérindienne ($cos^2 = 0,38$) ainsi que nombre supposé d'utilisateurs habitant dans un *census tract* ($cos^2 = 0,15$).

Pour terminer, les cercles de corrélations représentant les variables ont été réalisés en deux temps : le premier cercle croise les deux premières composantes (figure 6.16).

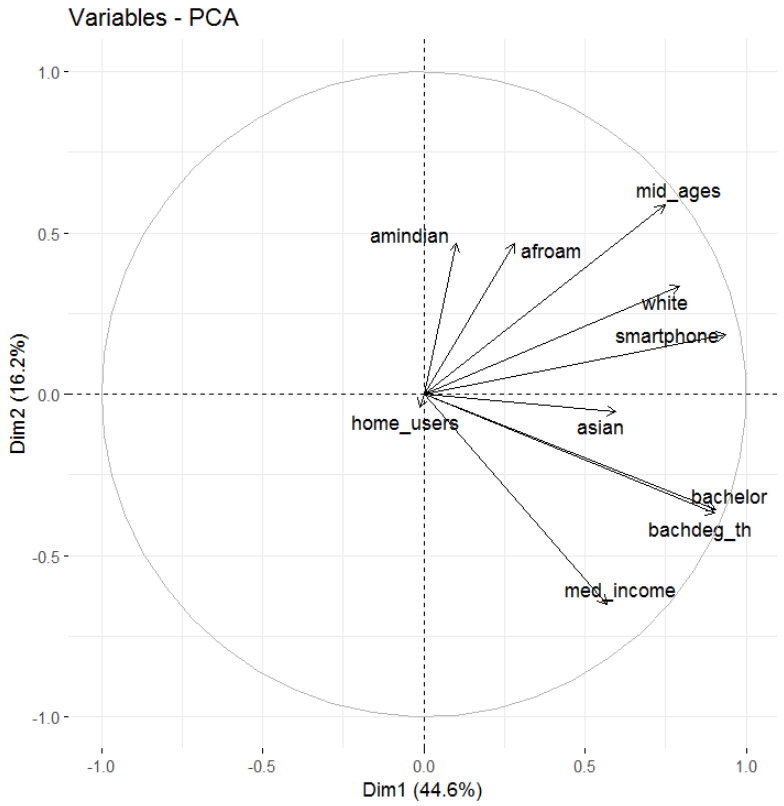


Figure 6.16 : Cercle des corrélations – Composantes 1 et 2

Nous pouvons d'ores-et-déjà noter que, pour les deux premières composantes (soit 60,8% de la variance totale), la variable *home_users* se révèle indépendante, puisqu'elle apparaît au centre du cercle (sa contribution à l'axe et d'ailleurs négligeable : 0,003%). La composante 1 ne note pas de rupture nette entre des groupes de variables. Nous pouvons cependant noter que les populations d'origines amérindienne et afro-américaine se trouvent en marge des variables marquant l'obtention d'un diplôme universitaire, l'accès aux TIC et la classe d'âge 20-44 ans. La composante 2 confirme cette tendance : il y a corrélation négative entre revenu médian d'une part et populations d'origines amérindienne et afro-américaine d'autre part.

Les cercles de corrélations suivants croisent les trois premières composantes avec la quatrième (puisque la contribution de la variable *home_users* ne se manifeste que sur cette composante). Les composantes 3 et 4 (figures 6.17 et 6.18) sembleraient indiquer une légère corrélation négative entre les lieux supposés de domiciliation des utilisateurs d'une part, et les populations d'origines afro-américaine et amérindienne d'autre part ; mais étant donné les faibles pourcentages de variance prise en compte par ces axes (respectivement 10,5% et 10%) et la faible qualité de représentation des variables relatives aux origines ethniques, il serait hasardeux de confirmer l'existence de corrélations entre l'activité virtuelle géolocalisée et ces deux groupes ethniques. En revanche, la composante 2 marque une corrélation négative entre le revenu médian d'une part, et la classe d'âge 20-44 ans, les populations d'origines afro-américaine et amérindienne d'autre part (figure 6.19).

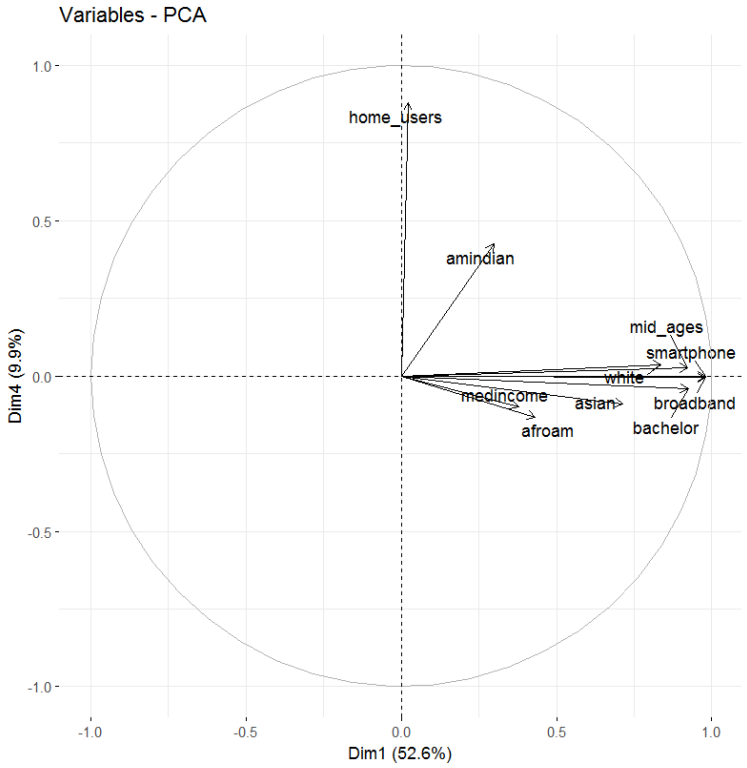


Figure 6.17 : Cercle des corrélations – Composantes 1 et 4

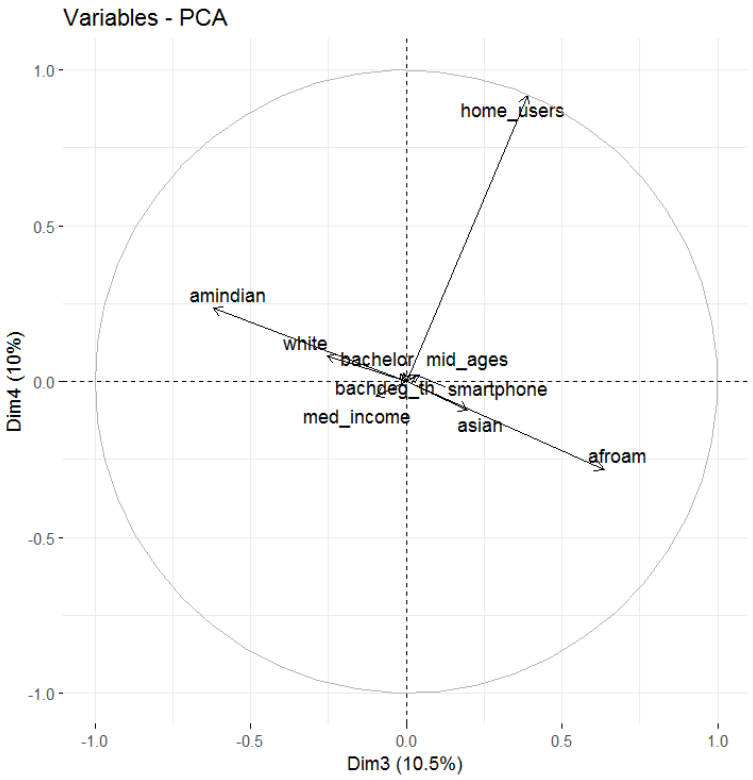


Figure 6.18 : Cercle des corrélations - Composantes 3 et 4

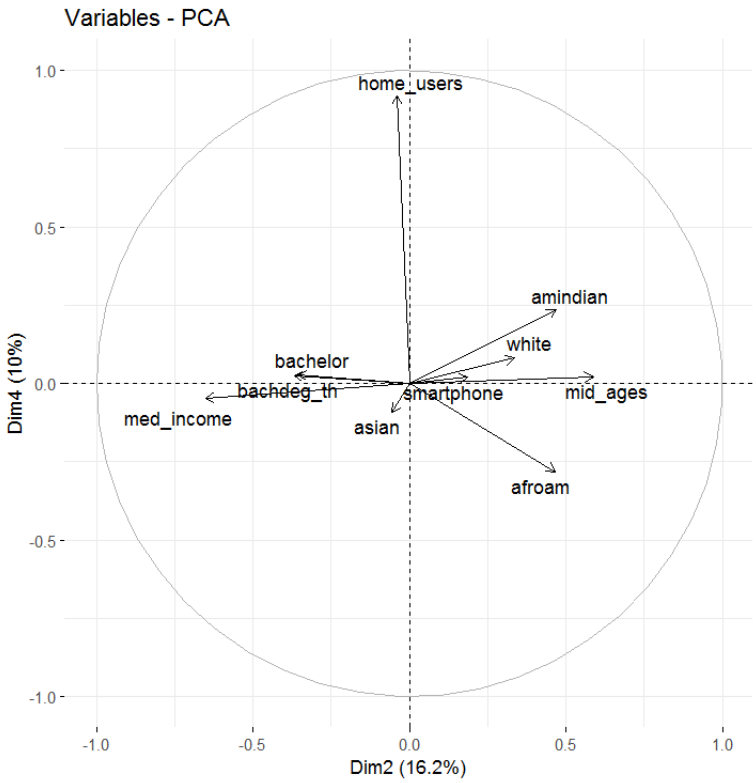


Figure 6.19 : Cercle des corrélations – Composantes 2 et 4

Au final, on ne sait pas si certaines populations sont exclues de la production de contenus virtuels géolocalisés sur Twitter. En outre, le fait de posséder un smartphone n'apparaît pas dépendant des revenus mais peut être lié au niveau de diplôme et associé à la population d'origine caucasienne (cf. composante 1, figures 6.16 et 6.17). En revanche, on ne peut pas conclure sur le point essentiel du profil de la population qui tweete en activant la géolocalisation : en effet, même si l'on parvient à regrouper ou opposer certains groupes de variables socio-démographiques, la variable *home_users* tend à l'indépendance statistique.

6.1.3.2. Populations productrices de contenus géolocalisés de la métropole d'étude, Houston

Les logiques de localisation des foyers d'émission de tweets géolocalisés s'avérant différentes de San Antonio, pouvons-nous cette fois mettre en évidence l'existence de relations entre variables socio-démographiques et participation à la création de contenus numériques géolocalisés ? Nous réutilisons la méthode de détermination du domicile de l'utilisateur pour l'aire métropolitaine de Houston et relevons un total de 7 888 tweets géolocalisés émis entre 1h et 6h30 le matin, par 1 476 utilisateurs. La localisation de ces utilisateurs nuance les observations constatées à San Antonio (figure 6.20) :

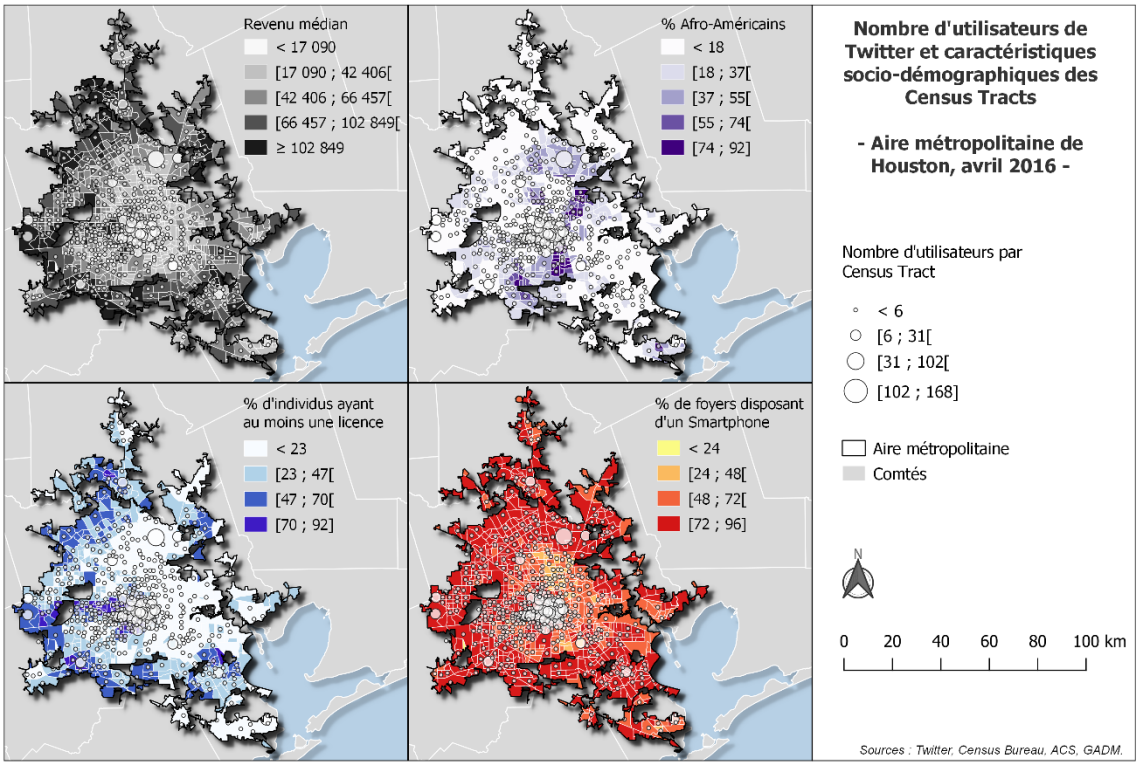


Figure 6.20 : Utilisateurs et caractéristiques socio-démagraphiques des Census Tracts, aire métropolitaine de Houston (C.Cavalière)

- le foyer le plus conséquent reste localisé dans les quartiers du centre-ouest de la métropole mais, contrairement à San Antonio, les utilisateurs sont quantitativement moins concentrés : le *CBD* ne rassemble ici que 23% des utilisateurs (37,5% à San Antonio) et seuls 37% des *census tracts* ne contiennent qu'un ou deux utilisateurs (72% à San Antonio) ;

- même si la localisation des utilisateurs nocturnes semble plus diffuse spatialement et enregistrée jusque dans les marges de l'aire urbaine, 51,7% des *census tracts* de Houston n'enregistrent pas d'utilisateur (ils étaient 43,3% à San Antonio) ;

Comme à San Antonio, la distribution spatiale des variables met en évidence les tendances suivantes :

- dans le centre de la métropole, les hauts revenus se concentrent dans les quartiers centre et centre-ouest ; se dessine ensuite une première couronne d'unités aux populations moins aisées dans un rayon d'une vingtaine de kilomètres autour du *CBD*. On trouve de nouveau des revenus médians élevés dans une seconde couronne localisée dans les marges de la métropole, entre 30 et 40 kilomètres du *CBD*. En outre, la répartition spatiale des individus diplômés de l'enseignement supérieur réfléchit la distribution des revenus médians.

- Globalement, les populations d'origine afro-américaine se concentrent dans les quartiers sud, nord et nord-est de la métropole : pour les *census tracts* proches du centre de la métropole (entre 3 et 20 kilomètres du *CBD*), il s'agit des quartiers les moins aisés et qui comptabilisent peu d'utilisateurs de Twitter.

- Le smartphone se répartit dans l'ensemble de l'aire métropolitaine : dans 94% des *census tracts*, un foyer sur deux dispose d'au moins un smartphone. En revanche, les 6% restants correspondent à ces quartiers défavorisés proches du centre (le taux de disposition d'un smartphone par foyer le plus bas est de 28%).

- à Houston, les relations entre variables socio-démographiques et répartition des utilisateurs s'avèrent sans doute plus marquées qu'à San Antonio : 40% des utilisateurs se trouvent dans des unités dont le revenu médian annuel dépassait 66 457$ en 2016. Les *census tracts* inclus dans les quartiers défavorisés et dont les revenus médians annuels font partie des plus faibles (< 17 090$ par foyer en 2016) de la métropole regroupent quant à eux 0,3% des utilisateurs. La relation entre le taux de foyers disposant d'un Smartphone et le nombre d'utilisateurs par unité de recensement apparaît encore plus nette : 82% des utilisateurs sont enregistrés dans un *census tract* dont au moins 72% des foyers disposent d'un smartphone.

Les données sont de nouveau soumises à une ACP normée afin d'identifier des liens éventuels dans l'ensemble des variables ajoutées : les quatre premières composantes cumulent 86% de la variance et sont par conséquent retenues pour la suite de l'analyse (figure 6.21).

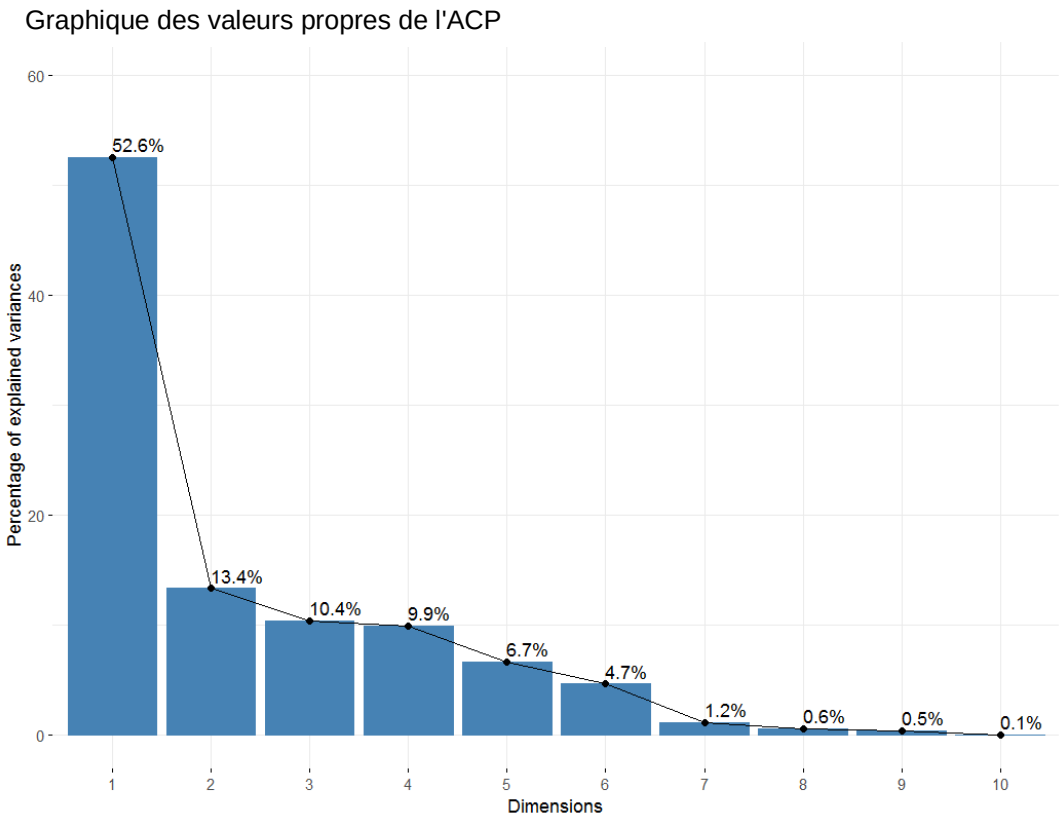


Figure 6.21 : Graphique des valeurs propres des composantes

Dans le cas de Houston, la qualité de représentation (figure 6.22 en page suivante) de la variable *home_users* (soit le nombre supposé d'utilisateurs résidents par *census tract*) s'avère de nouveau très faible sur trois des quatre composantes (les valeurs des cos² oscillent entre 0,00049 sur la composante 1 et 0,77 sur la composante 4). Pour autant, elle affiche une contribution (figure 6.22 en page suivante) de 20% à la composante 3, sur laquelle on retrouve également les populations d'origines amérindienne (52%) et afro-américaine (10%), ainsi que de 77% à la composante 4, à laquelle les populations amérindiennes contribuent également, mais à hauteur de 18% seulement.

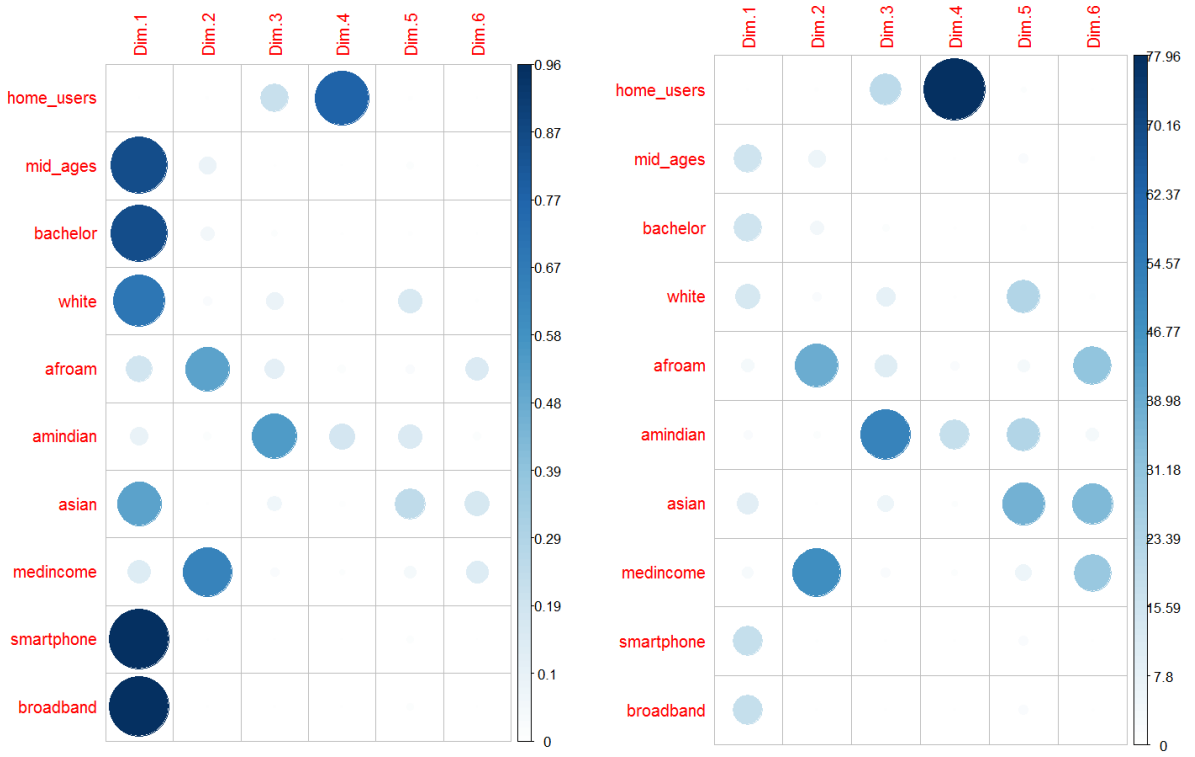


Figure 6.22 : Matrice de la qualité de représentation (cos², gauche) et des contributions (en %, droite) des variables aux composantes

Les deux premières composantes de l'analyse, qui cumulent 65,9% de la variance totale, représentent les variables suivantes :

- composante 1 : les valeurs des cos² dépassent 0,7 pour les variables *mid_ages* (nombre d'habitants âgés entre 20 et 44 ans), *bachelor* (nombre d'individus étant diplômés d'au moins une licence), *white* (nombre d'individus d'origine caucasienne), *smartphone* (nombre de foyers disposant d'au moins un smartphone) et *broadband* (nombre de foyers disposant d'un accès Internet à très haut débit). En outre, ces variables contribuent à la définition de l'axe 1 dans des proportions équivalentes (entre 13 et 20%).
- composante 2 : les valeurs des cos² dépassent 0,5 pour seules deux variables : *afroam* (nombre d'habitants d'origine afro-américaine) et *medincome* (revenu médian annuel par foyer). Elles contribuent à la définition de cet axe à 38% (*afroam*) et 42% (*medincome*).

Tout comme à San Antonio, la localisation présumée du domicile des utilisateurs contribue aux composantes 3 et 4, qui s'avèrent également de moindre poids dans l'analyse (respectivement 10,4% et 9,9% de variance prise en compte). De même, c'est une nouvelle fois sur la composante 4 que cette variable *home_users* présente les plus hautes valeurs de qualité de représentation et de contribution : mais sur cette même composante, la prise en compte des autres variables reste faible (population d'origine amérindienne) voire nulle (pour l'ensemble des autres variables du jeu de données). Observe-t-on alors les mêmes tendances qu'à San Antonio par les cercles des corrélations ? La position des variables sur les quatre

premières composantes met effectivement en évidence des structures analogues à la métropole témoin :

- la composante 1 (figure 6.23) témoigne de la non-significativité (et donc de l'indépendance) de la variable *home_users* dont la coordonnée est de 0,02. En revanche, elle corrèle les variables *bachelor*, *smartphone*, *broadband*, *mid_ages*, et, dans une moindre mesure, *white* et *asian* (nombre d'individus d'origine asiatique) ;

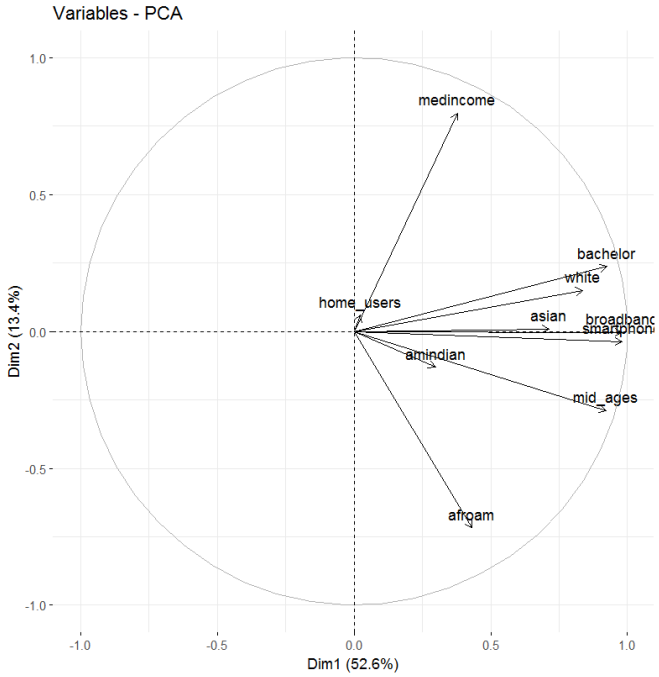


Figure 6.23 : Cercles des corrélations - Composantes 1 et 2

- la composante 2 marque une corrélation négative nette entre le revenu médian annuel par foyer et les populations d'origine afro-américaine (cf. figures 6.23 et 6.24) ; la variable *home_users* reste indépendante.

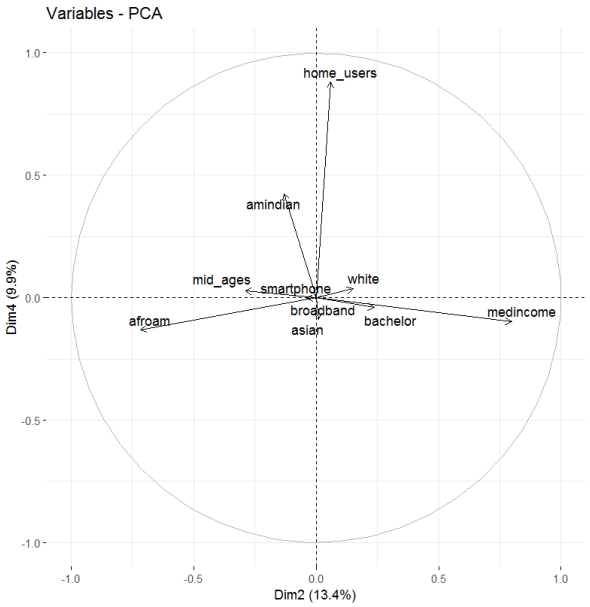


Figure 6.24 : Cercles des corrélations - Composantes 2 et 4

- la composante 3 semble marquer une corrélation négative entre la variable *home_users* et les populations d'origine amérindienne (figure 6.25) ;

- la composante 4 marque une corrélation négative plus ténue entre la variable *home_users* et les populations d'origine afro-américaine (figure 6.25) mais, tout comme dans le cas d'étude de San Antonio, le poids de cette composante dans l'analyse reste très faible.

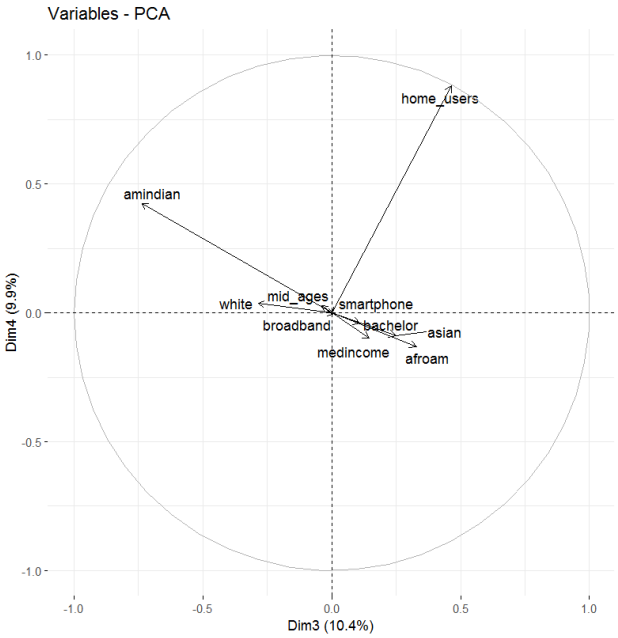


Figure 6.25 : Cercles des corrélations - Composantes 3 et 4

Pour conclure les essais de caractérisation de profils des populations créatrices de contenus géolocalisés sur Twitter, on pourra retenir que le fait d'être présent et actif sur le réseau en utilisant la géolocalisation ne semble pas, dans les deux cas étudiés, dépendre étroitement de facteurs socio-démographiques particuliers (puisque nous n'avons trouvé aucune corrélation positive ou négative forte entre la variable *home_users* et l'ensemble des variables test).

6.1.3.3. Les lieux de l'activité virtuelle géolocalisée sont-ils représentatifs des populations qui les fréquentent ?

Etant donné les nouvelles difficultés à identifier des profils de populations qui émettent des tweets géolocalisés, nous cherchons, dans un dernier test, à savoir si le lieu qui concentre des émissions de tweets géolocalisés fédère certaines catégories de populations. Nous nous intéressons donc à la mobilité des utilisateurs dans les lieux de l'activité tweeting normale de Houston. Le test est articulé autour de cette question : peut-on considérer que les utilisateurs qui fréquentent et tweetent en un lieu donné présentent les mêmes caractéristiques socio-démographiques que les populations recensées dans le lieu en question ? En d'autres termes, nous cherchons à savoir si la mobilité des utilisateurs et le fait de fréquenter et de tweeter en des lieux précis sont contraints par une ségrégation socio-spatiale.

A partir d'une couche vectorielle représentant les différents quartiers du centre de l'aire métropolitaine de Houston, nous sélectionnons les utilisateurs dont nous avons déterminé le domicile supposé dans le paragraphe précédent. Pour chaque utilisateur, nous sélectionnons ensuite l'ensemble des tweets géolocalisés émis en avril 2016 et construisons la matrice des quartiers d'origine et de destinations. Les cartes établies présentent ainsi les différents quartiers fréquentés et "tweetés" par des utilisateurs domiciliés dans un quartier différent (la taille des flèches varie en fonction du nombre d'utilisateurs ayant tweeté au moins une fois dans un quartier différent du quartier de résidence).

Au final, à Houston, on ne constate pas l'existence de mobilités spatiales contraintes par des critères socio-démographiques : on observe en effet que les populations originaires des quartiers précaires fréquentent et tweetent aussi bien dans des quartiers dont le profil est similaire à leur quartier d'origine, que dans des quartiers plus aisés, et quelle que soit la distance des différents quartiers. Les figures 6.26 et 6.27 (en page suivante) représentent ainsi les flux d'utilisateurs géolocalisés (reconstitués à partir des différents lieux dans lesquels ils ont tweeté), respectivement originaires des quartiers modestes de *North* et de *Southeast Houston* ; ces flux ont pour destinations d'autres quartiers modestes comme l'*Eastend* ou encore le *Northeast Houston*, mais encore des quartiers plus aisés comme *Greater Memorial*, *Rice Military* ou *River Oaks*.

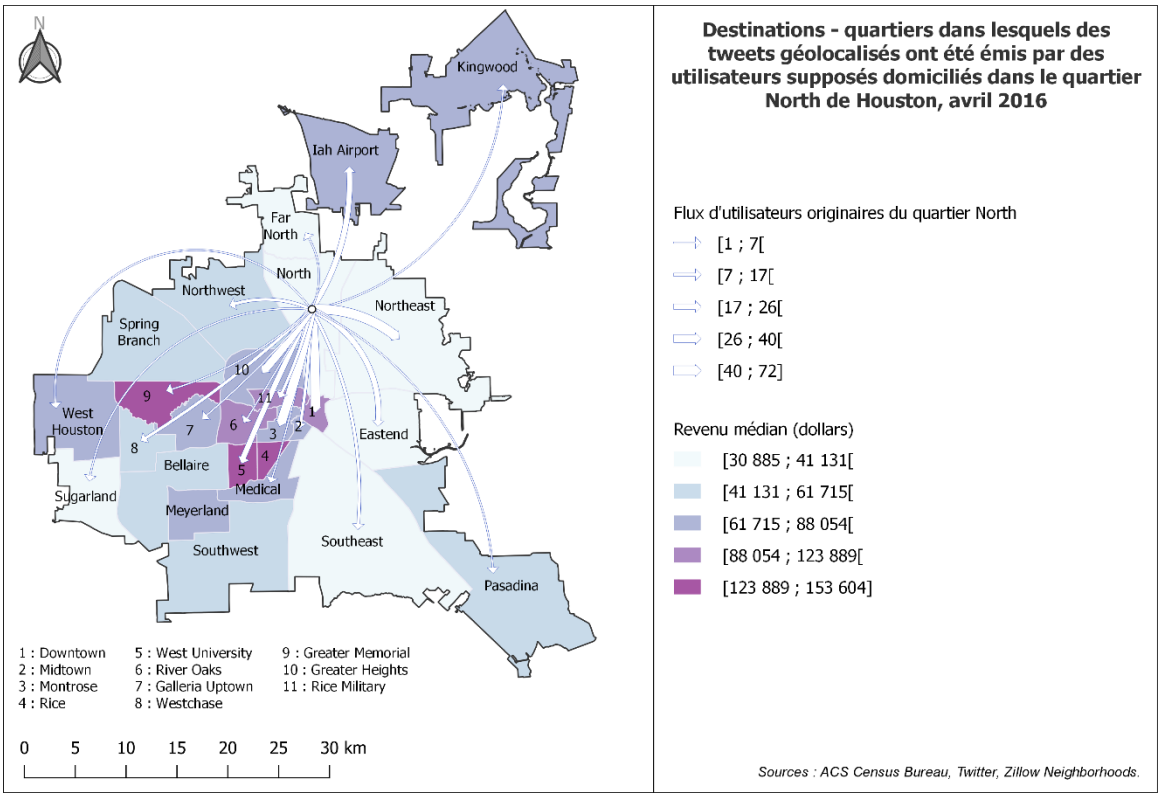


Figure 6.27 : Quartiers fréquentés par des utilisateurs originaires d'un quartier aux populations modestes du centre de Houston, North (C.Cavalière)

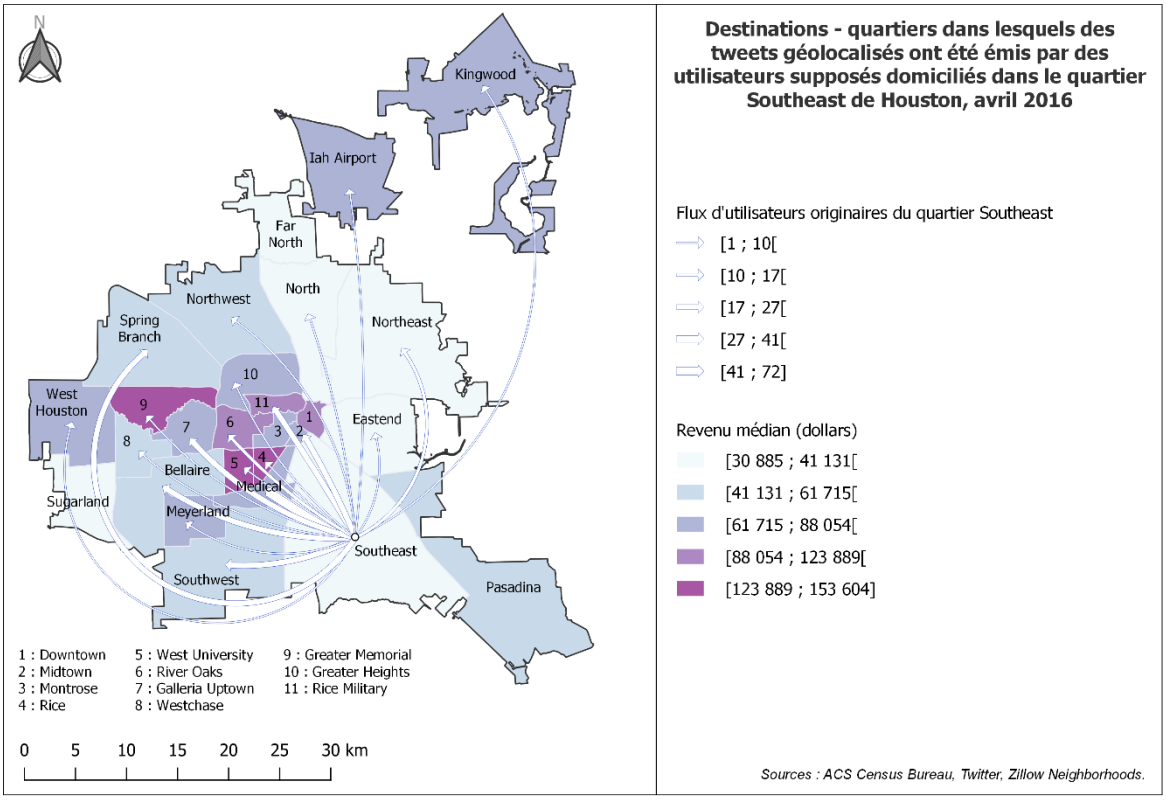


Figure 6.26 : Quartiers fréquentés par des utilisateurs originaires d'un quartier aux populations modestes du centre de Houston, Southeast (C.Cavalière)

Si l'on s'intéresse ensuite aux utilisateurs originaires de quartiers plus aisés, le même phénomène est perceptible (il n'y a pas de comportement spatial indiquant un "entre-soi"). Les figures 6.28 et 6.29 représentent ainsi les utilisateurs respectivement originaires des quartiers plus aisés de *Montrose* (limitrophe du *CBD*) et de *River Oaks* : même si, pour ces utilisateurs, les destinations les plus fréquentes restent des quartiers limitrophes du quartier d'origine, ou situés dans un rayon d'une dizaine de kilomètres, ils s'avèrent actifs dans tous types de quartiers, modestes ou aisés.

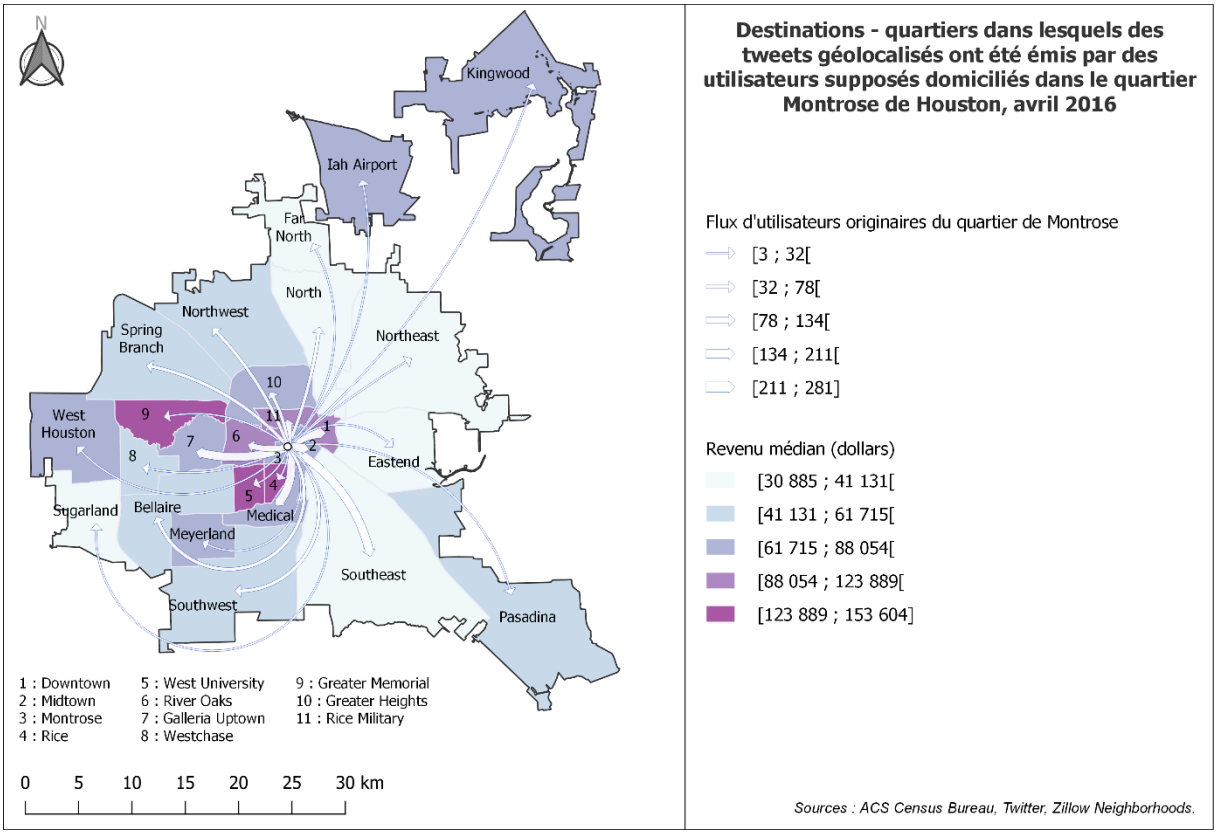


Figure 6.28 : Quartiers fréquentés par des utilisateurs originaires d'un quartier aux populations aisées du centre de Houston, Montrose (C.Cavalière)

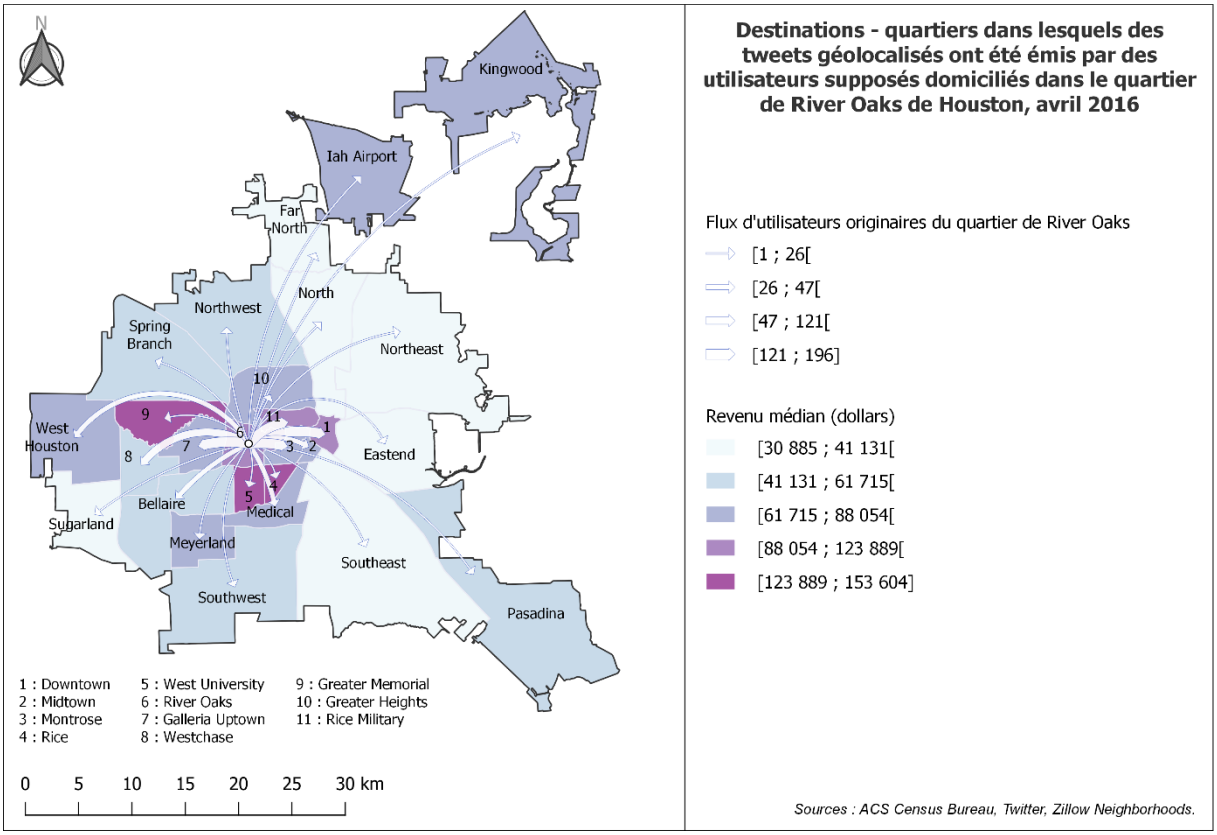


Figure 6.29 : Quartiers fréquentés par des utilisateurs originaires d'un quartier aux populations aisées du centre de Houston, River Oaks (C.Cavalière)

Bilan des tests de détermination de profils de populations productrices de tweets géolocalisés

Dans notre cas d'étude (et contrairement à [Li *et al*., 2013]), les analyses statistiques ne permettent pas de supposer l'existence d'un ou de plusieurs profils types de populations productrices de tweets géolocalisés.

Dans un second temps, et puisqu'il s'avérait difficile de déterminer des profils précis d'utilisateurs, nous avons effectué la démarche en sens inverse (on ne se focalise plus sur un domicile supposé mais sur l'activité virtuelle quotidienne) : nous avons souhaité savoir si les utilisateurs qui tweetent en un lieu donné (et qui fréquentent donc ce lieu en situation de mobilité) présentent le profil socio-démographique des populations recensées en ce lieu. Les tests apportent une réponse négative à cette question puisque les lieux d'activité virtuelle des utilisateurs géolocalisés n'apparaissent pas contraints par des barrières sociales informelles. Au final, sur Houston, on peut penser que n'importe quel individu est susceptible de participer et ce, en dépit de son profil socio-démographique.

6.2. Les logiques d'émergence spatio-temporelle de l'activité virtuelle géolocalisée de crise

Puisqu'il apparaît, dans notre cas d'étude, périlleux de rechercher des facteurs explicatifs de la spatialisation globale de l'activité virtuelle de crise ou encore de mettre en évidence des profils de populations créatrices de cette activité, nous nous intéressons maintenant au tweet de crise géolocalisé par lui-même, en termes d'émergence spatio-temporelle de l'activité virtuelle de crise. Dans la métropole, quels lieux créent l'événement virtuel et selon quelles temporalités ? Observe-t-on une agitation subite ou une agitation régulière et diffuse sur l'ensemble du territoire métropolitain ? Les tweets émis dans une certaine proximité spatiale et temporelle sont-ils cohérents dans leur discours ?

6.2.1. Réactivité temporelle face à l'alerte virtuelle

6.2.1.1. Réactivité temporelle à l'alerte du phénomène récurrent

Dans l'aire métropolitaine, nous cherchons dans un premier temps à mettre en évidence l'éventuelle existence d'une réactivité temporelle à l'émission de l'alerte, dans le voisinage direct des lieux d'émission de ces alertes. Celles-ci sont représentées par deux types de tweets :

- les tweets associés aux stations météorologiques et stations de jaugeage de l'*USGS*, qui indiquent la survenue de pluies violentes ou de crues locales, dans le quart d'heure écoulé avant l'émission du message : *"#USGS08068400 - Panther Br at Gosling Rd, The Woodlands, TX, Heavy rain (0.56 in/hr) over the last 15 minutes* " (pour la survenue de pluies intenses : 0,56 in/h = 14,22 mm/h) ; *"#USGS08073700 - Buffalo Bayou at Piney Point TX, is above NWS flood stage (52ft)*" (pour signaler qu'à la station localisée à *Piney Point Village* [quartier résidentiel de *Greater Memorial*, cf. figures 6.27 à 6.29] le long du *Buffalo Bayou*, la hauteur des eaux est supérieure au seuil de danger[8] fixé par le *NWS*, en l'occurrence 15,84 mètres).

- les tweets associés aux comptes du *NWS* : même si le tweet est géolocalisé en un point précis du territoire, le contenu du message concerne un espace plus large, comme une banlieue résidentielle de l'aire métropolitaine : "*Severe thunderstorm warning including Jersey Village TX, Bunker Hill Village TX until 3.30 AM".*

Les graphiques ci-après (figures 6.30 et 6.31) représentent les différences des temporalités d'émission entre les tweets de ces organismes officiels, considérés comme marqueurs virtuels de mise en alerte réelle, et les autres tweets géolocalisés émis dans un

[8] Définition du *flood stage* : en un point donné du lit du cours d'eau, il s'agit de la hauteur d'eau seuil, au-delà duquel l'élévation des eaux va engendrer un risque pour les personnes, biens et activités. Source : https://www.weather.gov/aprfc/terminology (Consulté pour la dernière fois le 24/06/2019).

rayon de cinq kilomètres : ces tweets représentent 63,2% des tweets non institutionnels émis le 19 avril 2016 (journée pour laquelle on enregistrait le plus de tweets officiels : 5 pour le *NWS* et 45 pour l'*USGS*). En-dessous de cette distance, on sélectionne trop peu de tweets : 4,1% à moins de 1 km et 7,6% à moins de 2 km).

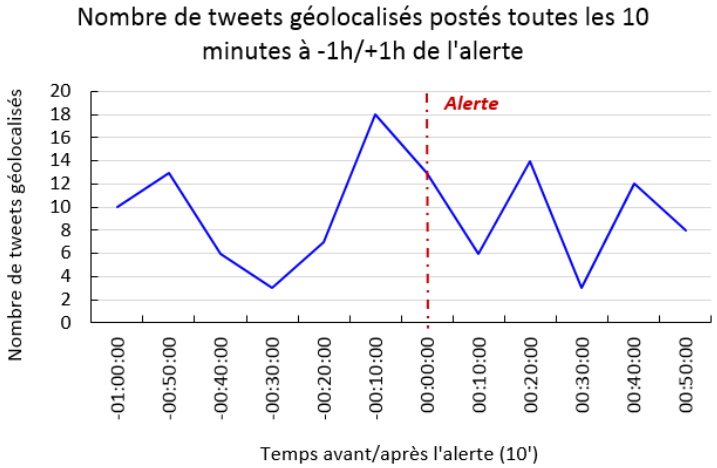


Figure 6.30 : Emissions de tweets géolocalisés à moins de 5 km d'une alerte et à ± 1h de l'alerte

La figure 6.30 (nombre de tweets géolocalisés émis à une heure de l'alerte, agrégés par dix minutes) n'indique pas d'épisode de réactivité temporelle localisée suite à l'émission d'une alerte virtuelle : en effet, si l'on constate un pic d'émissions à moins de dix minutes avant l'alerte, celles-ci s'avèrent en baisse dans les dix premières minutes après l'alerte. Si l'on modifie la résolution temporelle, on identifie la même tendance (figure 6.31) : les tweets émis à ± 1 minute de l'alerte, agrégés par dix secondes, ne représentent qu'une maigre poignée (soit un total de 5 tweets). Les émissions les plus "importantes" sont détectées dans les vingt secondes qui précèdent l'alerte, et l'émission post-alerte temporellement la plus proche est enregistrée à plus de cinquantes secondes.

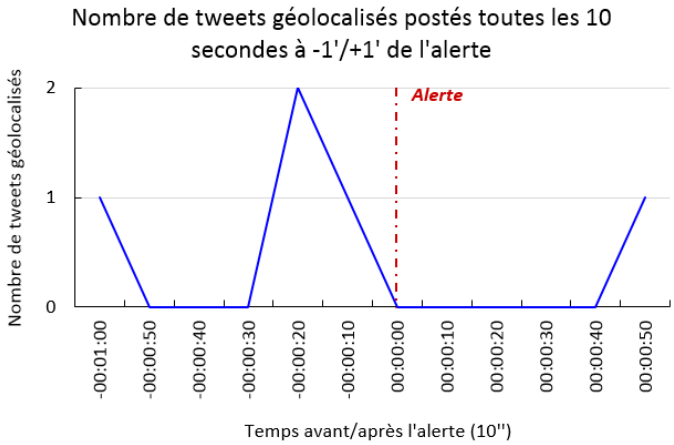
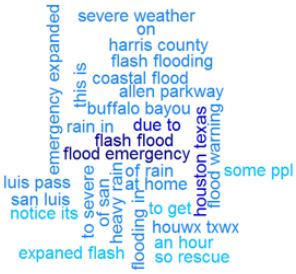


Figure 6.31 : Emissions de tweets géolocalisés à moins de 5 km d'une alerte et à ± 1' de l'alerte

Le comportement collectif de réactivité face à l'émission d'une alerte virtuelle n'est ici pas manifeste, puisque les pics d'émissions conséquents interviennent avant même l'émission des alertes virtuelles. Dans un second temps, nous interrogeons la cohérence des discours entre l'alerte et les émissions géolocalisées de crise non institutionnelles qui n'apparaissent pas comme des réponses directes à l'alerte. Le tableau 6.4 indique les types d'alerte identifiés dans les tweets marqueurs officiels ainsi que les thèmes (ou le lexique) rencontrés dans les tweets (toujours localisés à moins de 5 km des tweets alerte), aux deux pas de temps retenus précédemment (± 1h de l'alerte puis ± 1 minute de l'alerte).

Tableau 6.4 : Thèmes des tweets géolocalisés marqueurs d'alerte et autres, à ± 1h et ± 1' de l'émission de l'alerte virtuelle

<table>
<tr><th colspan="4">Discours des Tweets officiels</th></tr>
<tr><th colspan="2">NWS</th><th colspan="2">USGS</th></tr>
<tr><td colspan="2">Alerte aux orages violents
Alerte tornade</td><td colspan="2">Pluies intenses
Niveau de l'eau en-dessous/au-dessus du Flood Stage</td></tr>
<tr><th colspan="4">Discours des autres tweets géolocalisés</th></tr>
<tr><th>- 1h</th><th>+ 1h</th><th>-1'</th><th>+1'</th></tr>
<tr>
<td> 
</td>
<td> 
</td>
<td>Pluies intenses (pouring rain)
Nombreuses alertes inondations
Maison non inondée
Rues inondées
Maintien d'une activité professionnelle
Recherche d'aliments
Expression "Great flood of Houston"</td>
<td>#Houston
#Flooding
#EyeOpener</td>
</tr>
</table>

En fait, si l'on examine le contenu lexical des tweets géolocalisés spatialement et temporellement proches des tweets marqueurs d'une alerte, on pourra constater les trois faits suivants :

- l'ensemble des phénomènes du réel ne sont pas forcément relayés dans des temporalités proches de l'émission de l'alerte quand on s'intéresse à l'événement virtuel des usagers lambda : dans la temporalité horaire explorée ici, le phénomène de tornade annoncé par le *NWS* est totalement occulté par les pluies/crues/inondations[9].

- On observe une cohérence entre les types d'alerte et le discours des tweets : "*many flooding warnings*", "*rain still pouring*", "*heavy rain*", "*flash flooding*" ;

[9] Le phénomène de tornade apparaît bien plus tardivement dans l'événement virtuel : une première mention de tornade est constatée 1h30 après l'émission de l'alerte ; la seconde est enregistrée 19h14 après cette même alerte.

- pour autant, on peut s'interroger sur la significativité du tweet marqueur d'alerte officielle dans l'événement virtuel : en effet, dans l'heure précédant l'émission des tweets *NWS* et *USGS*, l'événement virtuel témoigne déjà d'un vocabulaire d'alerte et de gestion de crise : "*high water rescue*", "*flooding warnings*". En outre, la diffusion de ces messages ne semble pas bouleverser dynamique et contenu du réseau : comme les graphiques des figures 6.30 et 6.31 l'ont montré, les émissions sont plus fortes avant l'alerte. Le contenu sémantique paraît toutefois prendre en compte les tweets marqueurs officiels : "*flash flood emergency expanded*" émis entre 9 et 26 minutes après l'alerte indique le maintien de la situation de crise.

6.2.1.2. Réactivité temporelle à l'alerte du phénomène rare

Dans le cas particulier de l'ouragan, pouvons-nous alors observer un effet déluge de tweets, ou *a minima*, une réactivité temporelle dans les espaces proches des tweets marqueurs d'alerte (effet qui jusqu'alors n'a pas été perceptible) ? Nous répétons l'expérience précédente, mais sélectionnons ici les tweets géolocalisés émis le 27 août 2017 à moins de 1 km des tweets marqueurs d'alerte (pour rappel, le nombre de tweets de crise géolocalisés non institutionnels émis dans la métropole entre le 16 et le 21 avril 2016 était de 1 234 ; pendant l'ouragan, ce même paramètre était de 9 405 entre le 24/08/2017 et le 31/08/2017). Dans un premier temps, étant donné que l'ouragan a drainé 7,6 fois plus de tweets que le phénomène récurrent dans l'aire métropolitaine, nous avons agrégé les tweets par minute pour visualiser les tendances d'émissions à ± 1h des tweets d'alerte (figure 6.32).

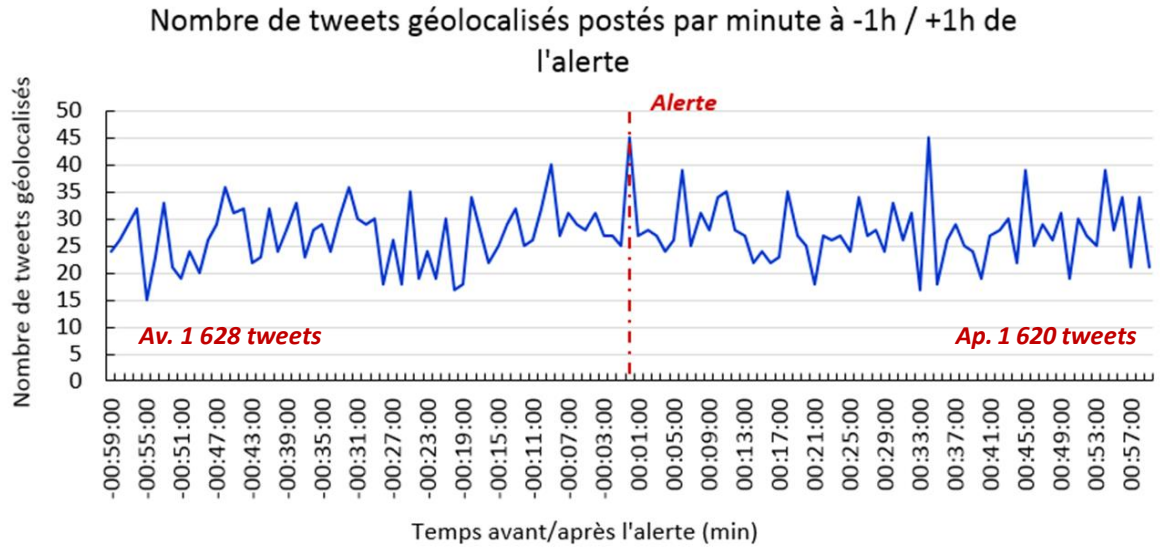


Figure 6.32 : Emissions de tweets géolocalisés à moins de 1 km d'une alerte et à ± 1h de l'alerte

Cette fois-ci, les émissions sont quantitativement équivalentes, qu'elles soient enregistrées avant ou après la diffusion des messages d'alerte (pouvant de nouveau remettre

en cause leur significativité étant donné qu'on enregistre déjà une activité antérieure à l'alerte dans leur voisinage proche). En revanche, on détecte, dans le cas précis de l'ouragan, un pic d'activité virtuelle concomitant à la diffusion de l'alerte virtuelle : ce pic intervient en effet dans la première minute qui suit la création du tweet marqueur (on pourra également noter un second pic équivalent en termes quantitatifs, détecté 32 minutes après le tweet d'alerte).

Par une résolution temporelle plus précise (de l'ordre de la minute, les tweets étant de nouveau agrégés par pas de dix secondes), on pourra détecter le premier pic suivant l'alerte entre [20 et 30[secondes après l'émission (figure 6.33).

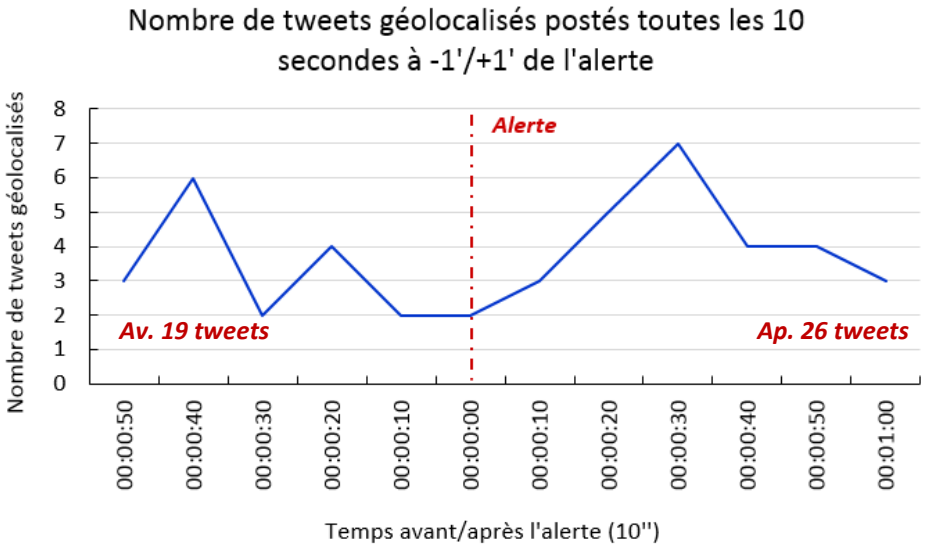
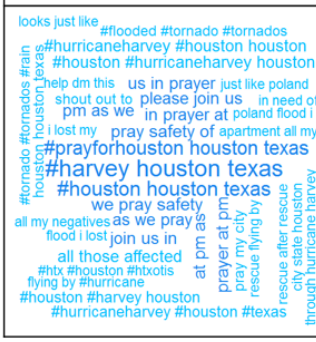
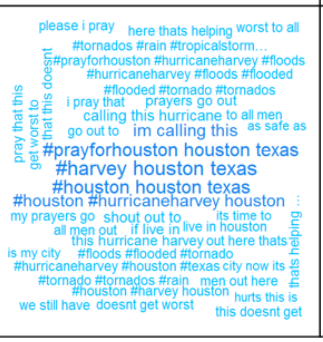
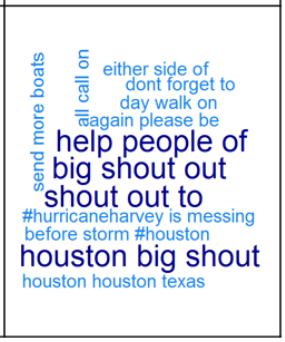


Figure 6.33 : Emissions de tweets géolocalisés à moins de 1 km d'une alerte et à ± 1' de l'alerte

Dans le cas du phénomène rare, on peut alors observer un effet de réactivité (bien qu'on reste quantitativement dans des ordres de grandeur assez faibles, d'où le non recours au terme *déluge*) dans des temporalités et des lieux voisins de l'alerte. Mais une fois de plus, l'activité restant continue et oscillant entre pics et creux sur l'ensemble de la plage horaire, cet effet est-il dû à l'émission des tweets d'alerte ? Nous procédons alors de nouveau à un examen lexical du contenu de l'ensemble des tweets (tableau 6.5).

Tableau 6.5 : Thèmes des tweets géolocalisés marqueurs d'alerte et autres, à ± 1h et ± 1' de l'émission de l'alerte virtuelle

Discours des Tweets officiels			
Pluies intenses Vents violents Orages		Buffalo Bayou en crue Montée des eaux Inondations	
Discours des autres tweets géolocalisés			
- 1h	***+ 1h***	***-1'***	***+1'***
			

Le phénomène rare se caractérise cette fois-ci par la présence prégnante des messages adressant des prières (en témoigne la récurrence du hashtag *#PrayForHouston*), messages qu'on ne peut pas systématiquement attribuer à des personnes directement affectées ou se trouvant dans une situation critique. Pour autant, nous pouvons dresser un constat identique au précédent : dans l'heure, voire dans la minute même, antérieure à l'émission des tweets d'alerte, on détecte déjà un vocabulaire propre à la gestion de crise : "*rescue flying*", "*alert emergency*", "*list of shelters*", "*if have boats*". En outre, on identifie le participe "*flooded*" indiquant que les conséquences du phénomène sont déjà visibles dans le monde réel.

En conséquence, quels que soient le type et l'intensité du phénomène physique, l'événement virtuel se trouve déjà engagé dans une dynamique de gestion de crise, au moins pour les tweets géolocalisés proches des tweets d'alerte, avant la diffusion de ces alertes.

6.2.2. Emergence spatio-temporelle et cohérence de l'activité virtuelle géolocalisée

6.2.2.1. Détection de périodes temporelles exploratoires

Dans le paragraphe précédent, nous avons constaté une variabilité temporelle de la réponse virtuelle aux phénomènes et événements résultants du réel ; en revanche, le sens des temporalités d'émission des tweets officiels s'avérait limité. Avant d'entamer les analyses spatio-temporelles et sémantiques localisées à l'échelle de l'aire métropolitaine, nous avons besoin de définir les périodes horaires pendant lesquelles nous les focaliserons. Cette tâche

est effectuée à partir de la visualisation graphique des émissions de tweets géolocalisés par tranche horaire. Désormais, en raison des résultats précédents concernant les temporalités peu significatives des tweets officiels, seuls les tweets non inventoriés comme institutionnels sont pris en compte dans la délimitation de périodes d'étude. La délimitation des périodes d'étude prend en compte les critères suivants :

- l'existence de pics d'émissions qui se distinguent des émissions continues pour les journées pendant lesquelles la ville est directement frappée par la perturbation ;
- l'existence de pics d'émissions dans les journées qui suivent le passage de la perturbation ;
- les périodes continues qui alternent pics et creux d'émissions.

La figure 6.34 représente les effectifs de tweets non institutionnels émis par heure entre le 18 et le 21 avril 2016. Conformément aux critères énoncés ci-avant, nous définissons, comme périodes exploratoires, les plages horaires suivantes sur des périodes continues :

- le 18/04 de 6h à 12h ;
- le 19/04 de 2h à 7h ;
- le 20/04 de 11h à 19h ;
- le 21/04 de 10h à 14h.

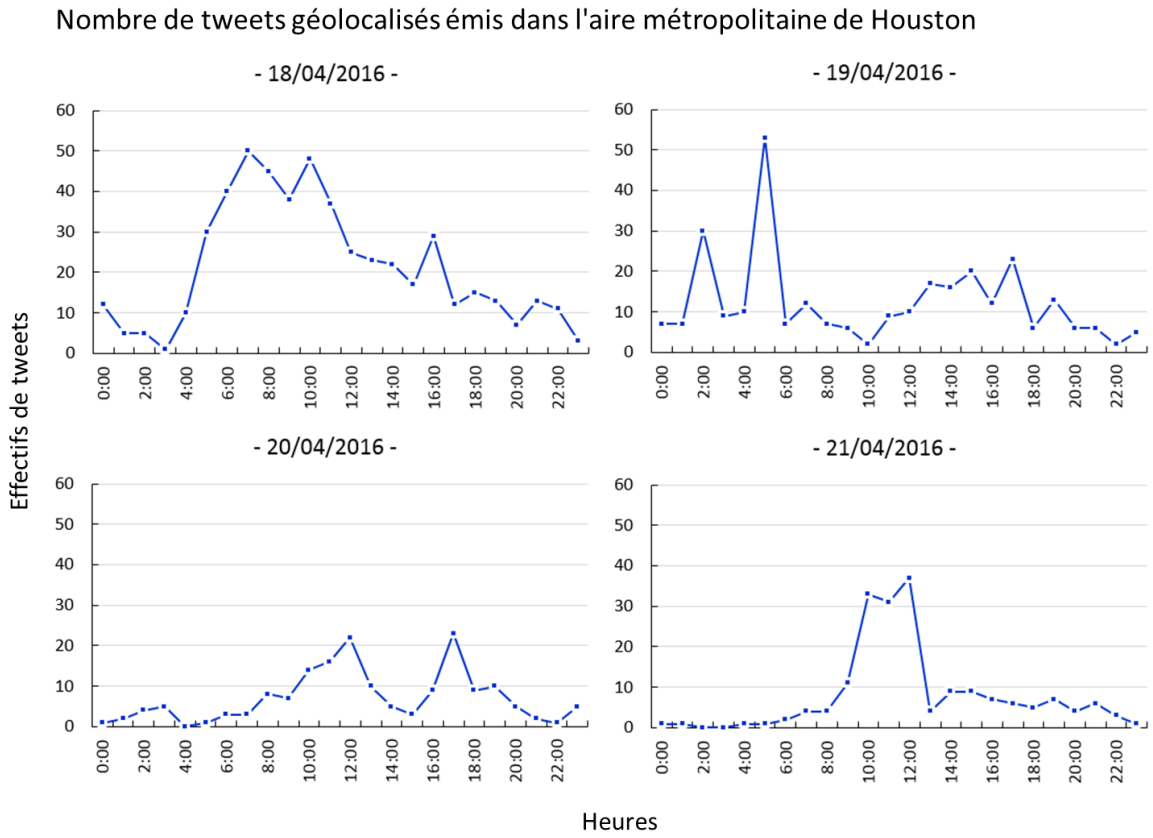


Figure 6.34 : Nombre de tweets géolocalisés émis par heure dans l'aire métropolitaine de Houston, 18-21/04/2016

La figure 6.35 représente les effectifs de tweets non institutionnels émis par heure pendant le passage de l'ouragan Harvey sur la métropole de Houston, entre le 25 et le 30 août 2017 (la ville n'est physiquement touchée qu'entre le 26 et le 29 août mais elle reste le lieu le plus actif pendant toute la durée du phénomène). Dans le chapitre précédent, nous avions constaté une certaine homogénéité temporelle concernant les périodes pendant lesquelles on enregistrait un événement virtuel (pour rappel, on considérait alors l'émergence d'un événement lorsque le nombre de tweets émis dépassait la médiane) : quels que soient les jours, on observait ces événements entre 8h et 21-22h (soit les horaires d'activité normale). Nous nous demandions alors si les temporalités de l'activité tweeting de crise suivaient la logique normale ou les dynamiques temporelles du phénomène physique et des événements réels résultants. A Houston, cette même tendance est perceptible (figure 6.35) : même si l'activité générale augmente à partir du 27 août, l'activité la plus conséquente reste enregistrée entre 8h et 19h, quel que soit le jour considéré.

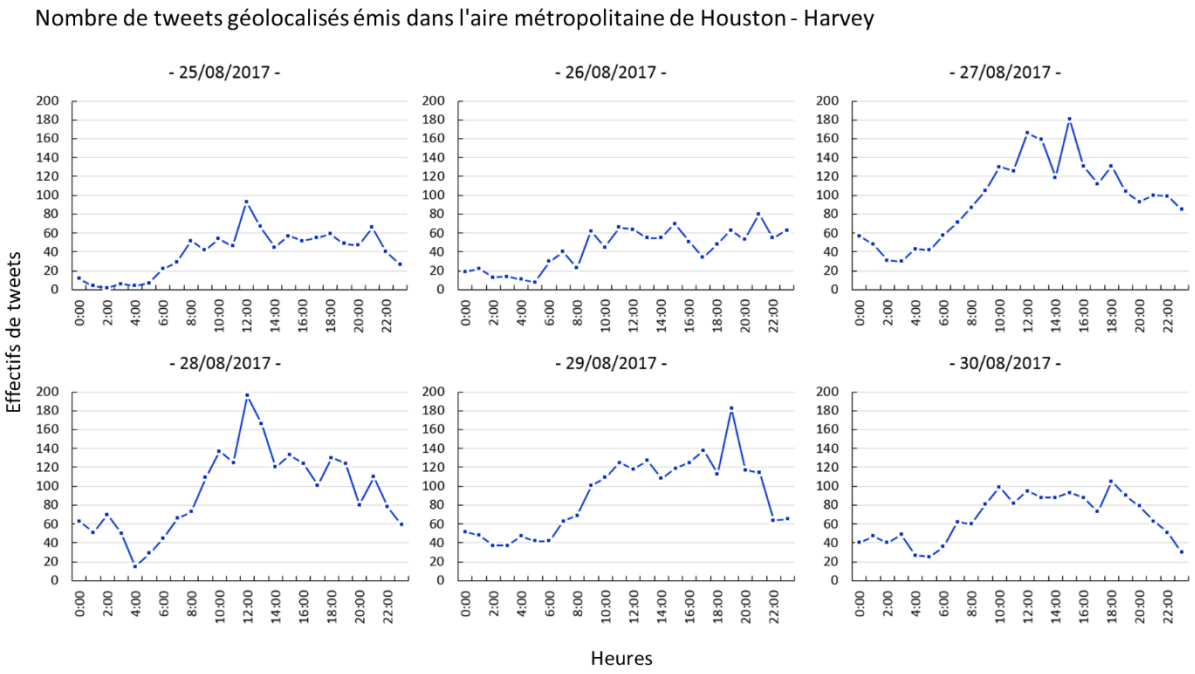


Figure 6.35 : Nombre de tweets géolocalisés émis par heure dans l'aire métropolitaine de Houston, 25-30/08/2017

6.2.2.2. Emergence et cohérence de l'événement virtuel en réponse à un phénomène récurrent

Test de la LDA pour l'analyse sémantique des tweets de crise géolocalisés.

Une LDA a été exécutée, afin de clustériser les mots fréquemment associés dans les tweets, pour chaque journée comprise entre le 18 et le 21 avril 2016 (on peut en effet supposer que les thèmes présents dans l'événement virtuel changent en fonction du temps : une LDA exécutée sur l'ensemble des tweets émis du 18 au 21 avril risquerait de ne pas fournir

des résultats assez précis). Le premier constat est le suivant : en dépit de l'échelle temporelle retenue, il y a une difficulté manifeste à rattacher les *topics* constitués, quel que soit le nombre de *topics* paramétrés dans l'analyse, à un thème précis ; par exemple, les mots *flood*, *floodings* et *Houston* sont quasiment présents dans l'ensemble des *topics* constitués. La figure 6.36 présente deux tests effectués pour la journée du 18 avril, avec respectivement 10 et 7 *topics.*

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
[1,]	"rain"	"houston"	"flooding"	"bayou"	"houston"	"rain"	"flood"
[2,]	"#houston"	"texas"	"due"	"#houston"	"flooded"	"now"	"flash"
[3,]	"texas"	"#houstonflood"	"closed"	"#houstonflood"	"city"	"weather"	"harris"
[4,]	"flood"	"#houston"	"houwx"	"storm"	"flooding"	"houston"	"county"
[5,]	"#flood"	"water"	"today"	"buffalo"	"flood"	"heavy"	"houston"
[6,]	"today"	"safe"	"houstonflood"	"oak"	"rain"	"wind"	"emergency"
[7,]	"one"	"home"	"isd"	"white"	"morning"	"mph"	"warning"
[8,]	"#houstonweather"	"everyone"	"weather"	"park"	"texas"	"get"	"houwx"
[9,]	"#flooding"	"day"	"flooded"	"flooding"	"abc13"	"flooded"	"rain"
[10,]	"288"	"#flood"	"severe"	"flooded"	"getting"	"work"	"reports"

(a). LDA exécutée avec sept topics de dix mots

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
[1,]	"flooded"	"rain"	"today"	"day"	"#houston"	"flood"	"houston"	"flooding"	"raining"	"closed"
[2,]	"houston"	"water"	"texas"	"rain"	"#houstonflood"	"flash"	"texas"	"houwx"	"houstonflood"	"due"
[3,]	"city"	"flooded"	"oak"	"today"	"texas"	"harris"	"#rain"	"houston"	"home"	"weather"
[4,]	"rain"	"night"	"white"	"flood"	"#flooding"	"county"	"#thunderstorm"	"weather"	"flooding"	"today"
[5,]	"flooding"	"texas"	"#houston"	"south"	"houston"	"houston"	"now"	"rain"	"everyone"	"stay"
[6,]	"good"	"last"	"#flood"	"north"	"storm"	"emergency"	"subdivision"	"now"	"getting"	"severe"
[7,]	"#houston"	"houston"	"bayou"	"flooded"	"bayou"	"warning"	"car"	"hounews"	"just"	"isd"
[8,]	"streets"	"one"	"#houstonweather"	"area"	"morning"	"reports"	"flooded"	"news"	"work"	"amp"
[9,]	"drive"	"lake"	"#houstonflood"	"san"	"buffalo"	"tornado"	"#houstonflood"	"wind"	"bayou"	"safe"
[10,]	"make"	"apartments"	"still"	"houston"	"park"	"houwx"	"#fairgreen"	"mph"	"hope"	"rain"

(a). LDA exécutée avec dix topics de dix mots

Figure 6.36 : Les dix premiers mots des topics créés pour deux tests de LDA

Ces *topics* permettent néanmoins d'identifier un vocabulaire minoritaire, qui peut s'avérer révélateur des perturbations engendrées et de la question des mesures de résilience : pour la journée du 18 avril, on pouvait ainsi détecter un *topic* dans lequel se trouvait le vocabulaire suivant : "*flood*", "*insurance*", "*departure*", "*delays*" faisant très vraisemblablement référence aux procédures d'indemnisation post-inondations (qu'on peut rattacher à une procédure administrative de résilience) ainsi qu'aux retards des vols au départ de l'aéroport intercontinental de Houston (perturbations consécutive aux intempéries). En dehors de ces indices épars (qu'on peut repérer en faisant varier le nombre de *topics* à créer ainsi que le nombre de mots à considérer dans chaque *topic*), il s'avère difficile d'identifier des ruptures permettant de discriminer les différents *topics* (et quel que soit le paramétrage de la LDA) : ici, on retrouve le mot *flood* (sous toutes ses déclinaisons) dans la quasi totalité des *topics* générés. En outre, peu de *topics* apportent un sens cohérent : si l'on reprend la figure 6.36.a, les *topics* qui fournissent une information cohérente seraient les topics 2 (adopter un comportement approprié face aux conditions environnementales), 3 (infrastructures fermées en raison des inondations), 4 (inondations des *Buffalo Bayou* et *White Oak Bayou*) et 7 (alerte aux crues éclair et signal de phénomènes en cours). A l'inverse, les autres *topics* ressemblent davantage à un enchevêtrement d'informations, comme le *topic* 6 dont le contenu, qui ressemble d'abord à un bulletin météorologique, se retrouve associé à des mots comme "*get*" et "*work*".

Le second problème rencontré reste que la fonction utilisée afin d'étiqueter chaque tweet en fonction du *topic* auquel il a la plus forte probabilité d'appartenir ne s'avère pas fiable sur ces jeux test peu volumineux. Le tableau 6.6 indique ainsi, pour chaque journée, le pourcentage de tweets de crise dont on connaît le *topic* d'appartenance le plus probable :

Tableau 6.6 : Pourcentage de tweets étiquetés par la LDA pour chaque journée de crise

Date	% de tweets étiquetés par LDA
18/04	70,4
19/04	7,1
20/04	1,1
21/04	9,1

Pourquoi un tel résultat est-il observé ? Nous pouvons d'abord préciser que même si les tweets du 18 avril ont été étiquetés à hauteur de 70%, tout tweet n'est pas nécessairement rattaché au *topic* le plus pertinent. Voici quelques exemples de tweets (à partir d'une LDA effectuée sur une base de sept *topics*, cf. figure 6.36.a) : "*Flash Flood extended from 12.30 to 1.30pm I guess I'm staying home*" est attaché au *topic* 1 ; il pourrait être relié aux *topics* 2 ("*home*") ou 7 ("*flash*" et "*flood*"). De même, "*Adline is cancelling school for Monday, April 18 due to severe weather, stay safe and dry every one!*" est classé dans le *topic* 1 alors qu'il contient "*severe weather*" du *topic* 3 et "*safe*" du *topic* 2. Le facteur explicatif le plus probable de ce manque de cohérence reste la trop grande hétérogénéité sémantique du contenu des tweets. En fait, la LDA est souvent utilisée pour des textes construits selon un plan précis et des idées hiérarchisées, par un unique auteur ; le tweet est rapide, sans contrainte de rigueur et, au vu de l'ensemble des résultats, on peut penser qu'il n'existe pas de *message standard*, lexicalement semblable aux autres, dès qu'on sort des tweets institutionnels qui affichent un langage cohérent et énonçant généralement une seule idée (l'émission d'une alerte, une route fermée, *etc*.). Par exemple, si l'on reprend le tweet précédent ("*Flash Flood extended from 12.30 to 1.30pm I guess I'm staying home*"), il entremêle une information à consonance officielle "*Flash flood extended from 12.30 to 1.30pm"* ainsi qu'une information personnelle "*I guess I'm staying home"* ; or, si cette information personnelle est intéressante d'un point de vue géographique, dans la mesure où elle révèle un comportement adapté à des conditions environnementales perturbées, le fait qu'elle associe deux informations à tonalités différentes semble introduire un biais dans la discrimination sémantique des tweets par l'algorithme de la LDA.

De ce fait, pour les cartographies présentées dans les pages suivantes, seuls les tweets émis pendant la journée du 18 avril sont représentés par la LDA ; la sémantique des jours suivants sera représentée par des nuages de mots de co-occurrences lexicales.

Cartographie des périodes temporelles identifiées.

La dynamique spatio-temporelle de l'événement virtuel permet de distinguer les comportements spatiaux de l'activité virtuelle de crise et d'observer la variabilité des lieux d'émission dans une même journée ainsi que d'un jour à l'autre. Cette dynamique est ici représentée par la cartographie des émissions quotidiennes de tweets de crise géolocalisés, agrégés par heure, dans les figures 6.37 à 6.40 (NB : dans la figure 6.37, qui représente les tweets de crise émis le 18 avril 2016, ceux-ci sont classés en fonction des *topics* identifiés par la LDA paramétrée à sept *topics* [cf. figure 6.36.a] ; les tweets des journées suivantes ne sont pas classés étant donné les résultats peu pertinents de la LDA). Pour chaque journée, les plages horaires cartographiées correspondent aux quatre périodes temporelles mises en évidence dans le paragraphe 6.2.2.1 (pour rappel, il s'agit des créneaux : 6h-12h le 18 avril, 2h-7h le 19 avril, 11h-19h le 20 avril et 10h-14h le 21 avril).

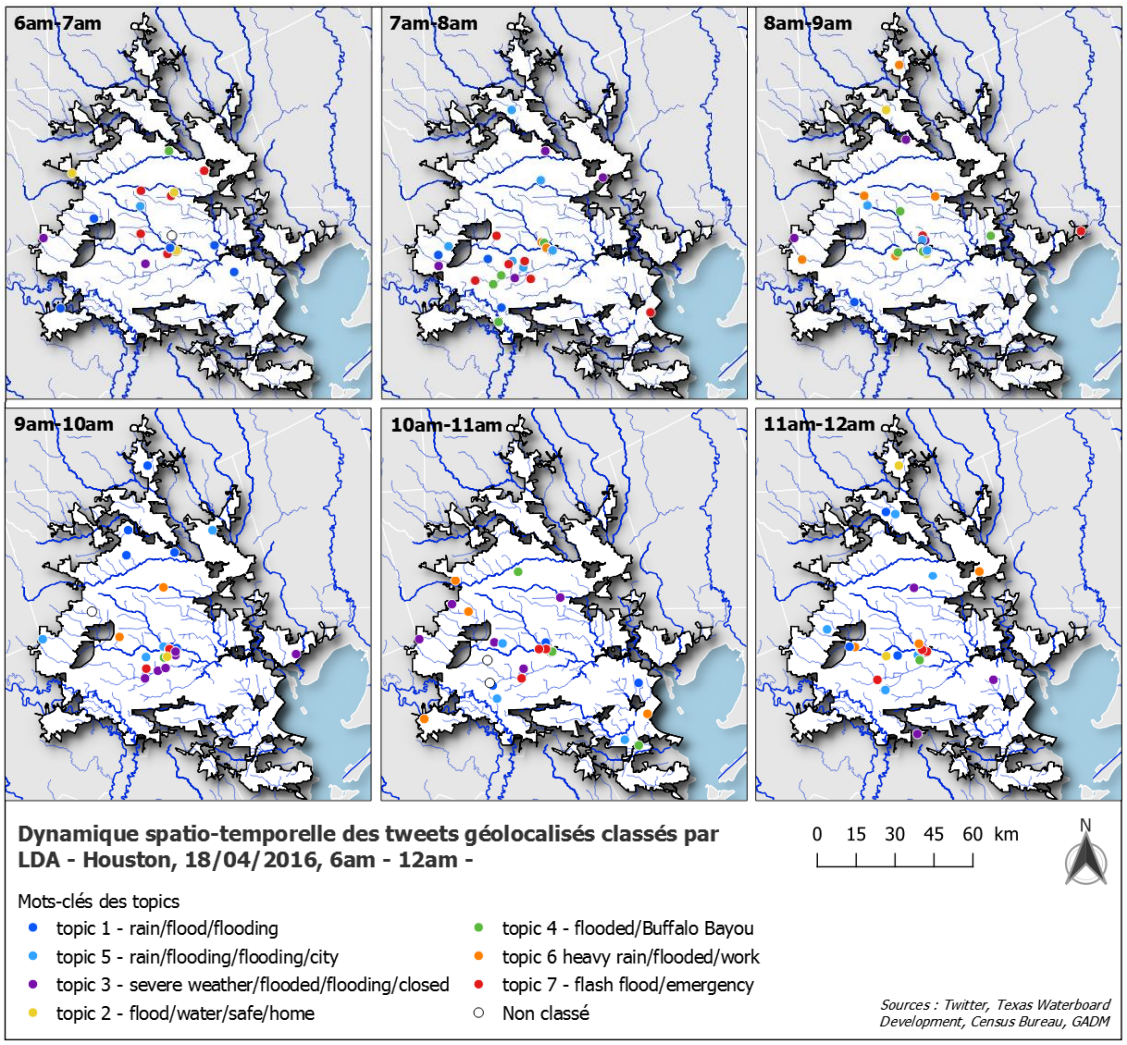


Figure 6.37 : Tweets de crise géolocalisés émis le 18/04/2016 entre 6h et 12h à Houston, classés selon les topics identifiés dans la LDA (C.Cavalière)

Dans un premier temps, si l'on observe, d'après la figure 6.37, la distribution générale des tweets de crise géolocalisés émis le 18 avril 2016 (sans prendre en compte le *topic* de rattachement), on pourra établir les constats suivants :

- le *CBD* reste le foyer permanent de l'activité virtuelle de crise, quelle que soit l'heure considérée ;
- une activité ponctuelle apparaît dans l'extrême nord de l'aire métropolitaine (entre 8h et 10h puis entre 11h et midi) ainsi qu'au sud-ouest du *CBD* (entre 7h et 8h puis entre 9h et midi) ;
- l'activité restante est dispersée dans les marges de la métropole.

Peut-on observer un sens spatial dans la distribution des *topics* ? Dans un premier temps, il est nécessaire de rappeler le peu de discrimination sémantique constatée dans l'ensemble des *topics* (et en conséquence la possibilité que deux tweets proches étiquetés en fonction de deux *topics* différents contiennent un lexique analogue). On pourra néanmoins dégager deux tendances perceptibles : d'une part, toutes les catégories de *topics* se retrouvent dans le *CBD* ; d'autre part, lorsque le sud-ouest de la métropole s'agite (et notamment entre 9h et 10h), un îlot virtuel relié au *topic 3* (*severe weather*, *flooded*, *flooding*, *closed*) s'active. Ces lieux d'agitation temporaire correspondent-ils aux pics d'activité identifiés sur les graphiques de la figure 6.34 ? Là aussi, on peut noter trois comportements distincts (cf. figure 6.37) :

- le pic d'activité virtuelle identifié entre 7h et 8h le 18 avril 2016 sur la figure 6.34, coïncide avec l'agitation observée dans le sud-ouest du *CBD* ;
- l'îlot d'agitation identifié au sud et sud-ouest du *CBD* entre 9h et 10h n'aboutit pas à la manifestation d'un pic d'activité sur le graphique correspondant de la figure 6.34 ;
- entre 11h et midi le 18 avril 2016, le graphique correspondant de la figure 6.34 indique un nouveau pic d'activité virtuelle de crise : or, cette fois, il n'y a pas d'agitation marquée dans un lieu précis ; les cellules d'activité s'avèrent éparses et diffuses dans l'ensemble de l'aire métropolitaine.

La figure 6.38 ci-après présente l'activité virtuelle des créneaux horaires identifiés pour la journée du 19 avril 2016 (à partir de cette journée, les tweets de crise géolocalisés figurent sous forme de points rouges, étant donné les faibles proportions de tweets étiquetés par la fonction associée à la LDA). On retrouve les comportements spatiaux décrits ci-dessus : une activité persistante dans le *CBD*, une agitation dispersée en divers lieux de la métropole (par exemple entre 2h et 3h le matin) ainsi que l'apparition d'une agitation (foyer temporaire de forme linéaire entre 5h et 6h le matin), suivant le lit du *Buffalo Bayou*.

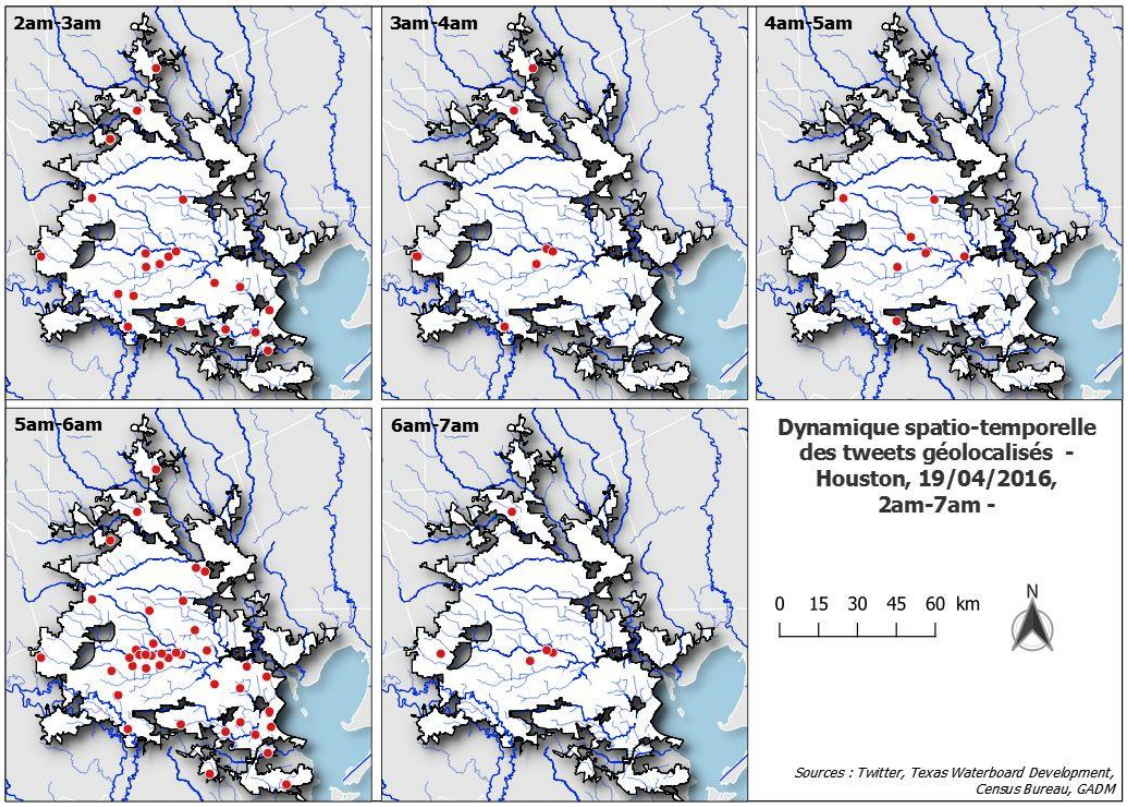


Figure 6.38 : Tweets de crise géolocalisés émis le 19/04/2016 entre 2h et 7h à Houston (C.Cavalière)

Le comportement de ce foyer semble indiquer un phénomène en cours dans le créneau horaire considéré : l'agitation est localisée en suivant une structure précise (le lit du fleuve), elle apparaît puis disparaît rapidement (l'activité virtuelle antérieure puis postérieure au créneau 5h-6h ne se solde qu'à une dizaine de tweets de crise) ; enfin, le graphique représentant les émissions de tweets de crise géolocalisés le 19 avril 2016, sur la figure 6.34, marque bien l'existence d'un pic d'activité entre 5h et 6h le matin.

La figure 6.39 présente l'activité virtuelle des créneaux horaires identifiés pour la journée du 20 avril 2016 (cf. figure 6.34) : les émissions localisées sur le *CBD* varient (en termes d'effectifs) mais persistent. Par ailleurs, l'agitation qui se manifeste pendant certains créneaux horaires reste dispersée : le pic d'activité enregitré entre 12h et 13h sur le graphique de la figure 6.34 ne correspond pas à l'apparition subite d'un îlot de tweets concentrés mais à un ensemble probable de micro-événements dispersés dans divers lieux de la métropole.

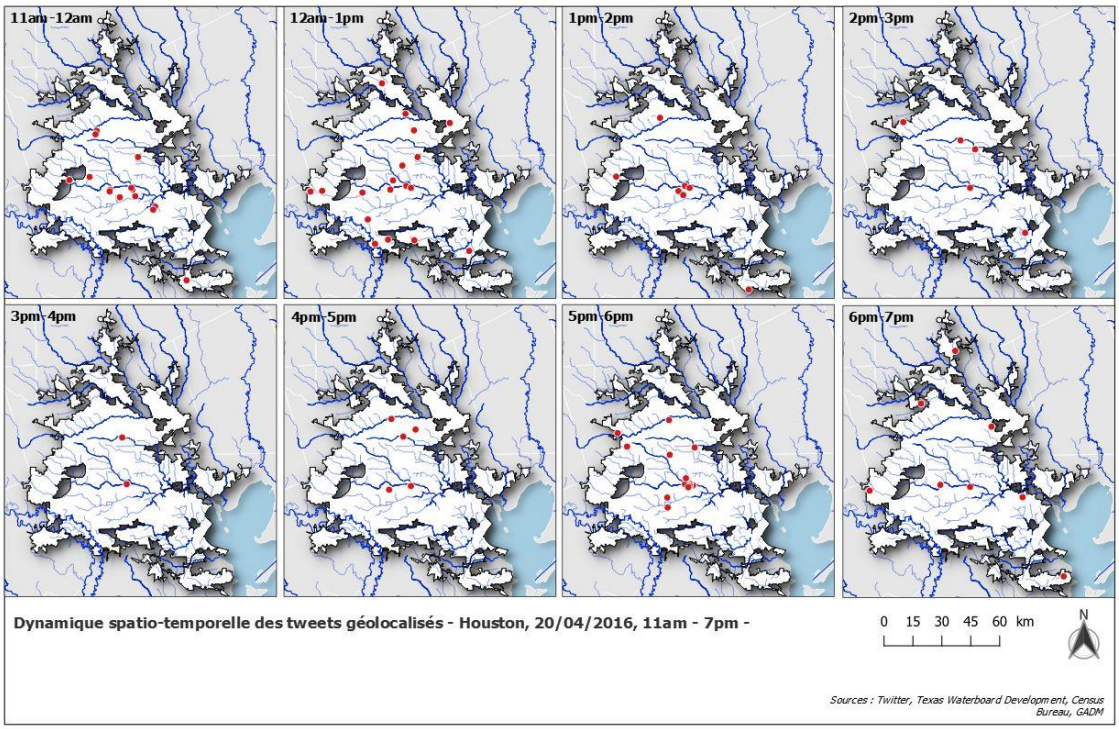


Figure 6.39 : Tweets de crise géolocalisés émis le 20/04/2016 entre 11h et 19h à Houston (C.Cavalière)

Enfin, la figure 6.40, en page suivante, représente la localisation des tweets de crise pour les créneaux horaires exploratoires identifiés en ce qui concerne la dernière journée d'événement virtuel noté, le 21 avril 2016. La période d'activité marquée, détectée sur le graphique de la figure 6.34, se manifeste sur les fenêtres cartographiques représentant les émissions de tweets de crise entre 10h et 13h. Une nouvelle fois, on retrouve les trois logiques mises en évidence précédemment :

- l'activité permanente du *CBD*, même quand tout autre lieu d'activité s'éteint brusquement (à partir de 13h) ;
- l'activation de l'îlot situé à l'ouest du CBD, le long du Buffalo Bayou ;
- une activité ponctuelle diffuse dans les marges de l'aire métropolitaine.

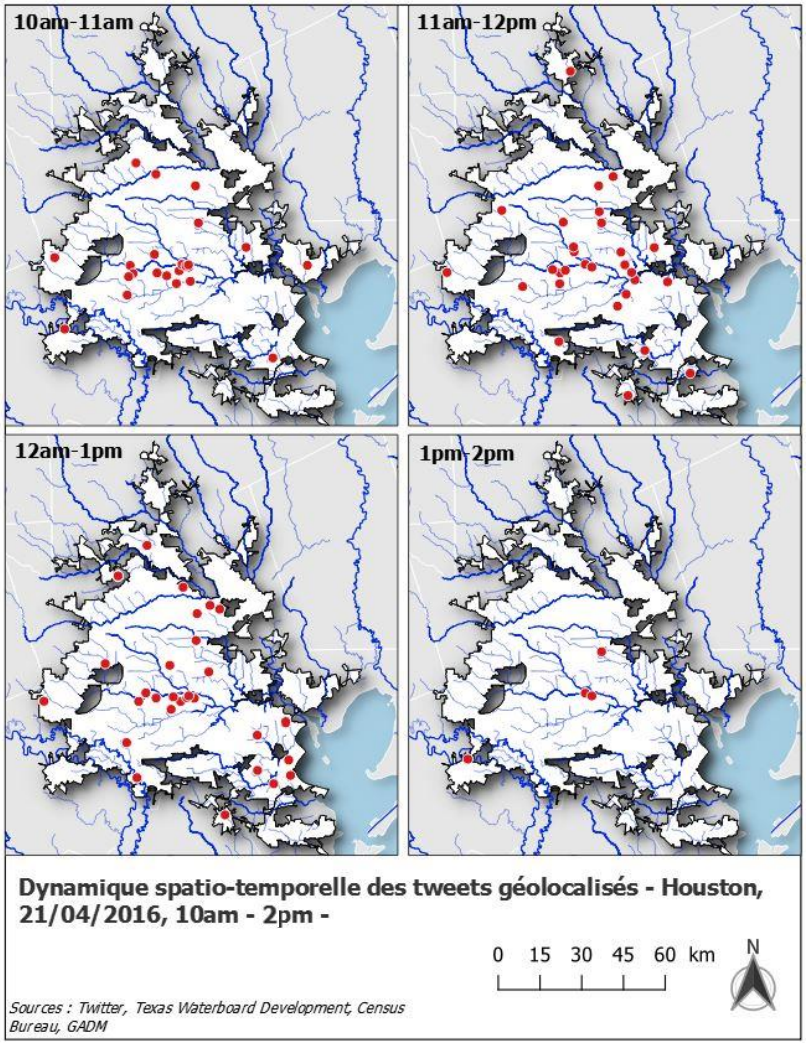


Figure 6.40 : Tweets de crise géolocalisés émis le 21/04/2016 entre 10h et 14h à Houston (C.Cavalière)

Pour conclure la première partie du test, l'agrégation temporelle des tweets de crise à l'échelle de l'heure indique l'existence vraisemblable d'une réactivité précise aux phénomènes du réel étant donné qu'on peut détecter des changements de la dynamique d'agitation/silence. Dans un second temps, la cartographie témoigne de l'existence de tendances spatiales formées par cette agitation : l'activité permanente du *CBD*, les îlots d'agitation temporaire concentrée ou diffuse le long du fleuve, ou encore l'activité éparse qui se manifeste ponctuellement dans les marges de la métropole. La suite du test est alors articulée autour de la question suivante : la sémantique des tweets peut-elle donner un sens aux variations constatées des dynamiques spatio-temporelles de l'événement virtuel ? Autrement dit, l'analyse sémantique des tweets en fonction des différentes périodes permettra-t-elle de mettre en évidence des facteurs explicatifs des dynamiques sous-jacentes aux variations spatio-temporelles constatées ?

Analyse sémantique des tweets de crise géolocalisés par période.

En ce qui concerne la journée du 18 avril, pour laquelle les tweets ont été cartographiés en fonction des *topics* de la LDA (cf. figure 6.37), il s'avèrait difficile d'identifier l'existence d'une logique de distribution spatio-temporelle de ces thèmes (qui risquait en plus d'être biaisée par le manque de précision de l'analyse effectuée). Comme signalé auparavant, on observait, sur la figure 6.37, deux comportements antagonistes : des tweets proches peuvent appartenir au même *topic* comme les tweets du *topic* 3 (*severe weaher*, *flooded*, *flooding*, *closed*) émis entre 9h et 10h ; mais la plupart du temps, des tweets proches sont rattachés à des *topics* différents.

Par l'analyse sémantique, peut-on alors obtenir des indices quant aux causes des pics et creux d'émissions de tweets de crise, identifiés précédemment ? Qu'évoquent les tweets émis sur ces lieux de veille permanente ou d'activité temporaire ? La figure 6.41 montre les variations horaires des contributions de chaque *topic* dans les tweets de crise correspondants.

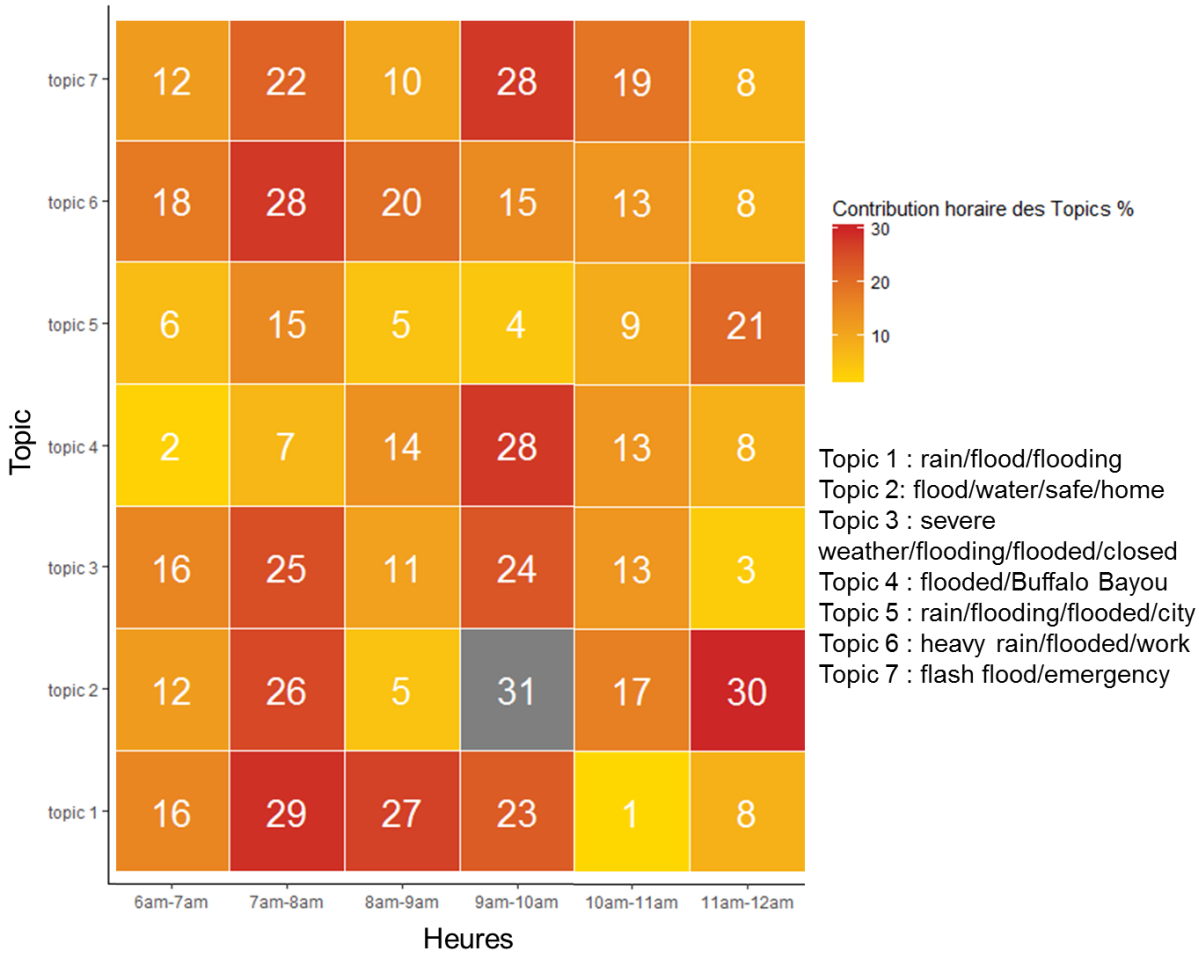


Figure 6.41 : Contribution (%) de chaque topic par tranche horaire le 18 avril 2016 entre 6h et midi

Même si les *topics* évoquent tous les inondations (*flood*/*flooded*/*flooding*), on observe une variation de leur poids en fonction du temps :

- le *topic* 2, qui prend en compte une problématique comportementale et les consignes de prudence (*home*/*safe*), est répétitif pendant certaines heures de la matinée : entre 7h et 8h, vraisemblablement avant que les habitants ne quittent leur domicile pour rejoindre leur lieu de travail ; le pic de contribution du *topic* 6 (*heavy rain*/*flooded*/*work*) est d'ailleurs concomitant à cet horaire (pour rappel, le 18 avril 2016 était un lundi). L'activité du *topic* 2 reprend entre 9h et 10h et enfin entre 11h et midi ;
- le *topic* 3 témoigne également de pics vraisemblablement concordants avec les flux pendulaires, entre 7h et 8h, puis 9h-10h (les mots clés du *topic* indiquent en effet des inondations en ville ainsi que des routes fermées à la circulation) ;
- le *topic* 4 indique la crue et les inondations du fleuve *Buffalo Bayou* : ce *topic* s'active à partir de 8h et connaît son poids maximum entre 9h et 10h (soit le créneau horaire pendant lequel l'îlot d'agitation apparaît le long du fleuve [cf. carte de la figure 6.37]). Sur ce même créneau horaire, le *topic* 7 (*flash flood*/*emergency*) affiche une même contribution à 28%.

Avec les données de la station de jaugeage de l'*USGS* localisée sur le *Buffalo Bayou*, on pourra vérifier si l'apparition temporelle de certains *topics* correspond à la période de crue du fleuve. La figure 6.42 ci-dessous indique les hauteurs d'eau du fleuve à la station *USGS 08074000* (cette station a été identifiée grâce aux tweets officiels qu'elle a émis et se trouve localisée sur la carte accompagnant le graphique, par le triangle rouge). Les tweets d'alerte en question ont été émis à 9:22 ("*#USGS08074000 Buffalo Bayou at Houston, TX is above NWS flood stage (28ft)*") pour signifier la mise en alerte consécutive au dépassement du seuil de hauteur d'eau fixé par le *NWS* à 28 pieds, soit 8,53 mètres. Le second tweet de cette station est émis le soir à 21:22 pour signifier la fin de la période de crue : "*#USGS08074000 Buffalo Bayou at Houston, TX is below NWS flood stage (28ft)*".

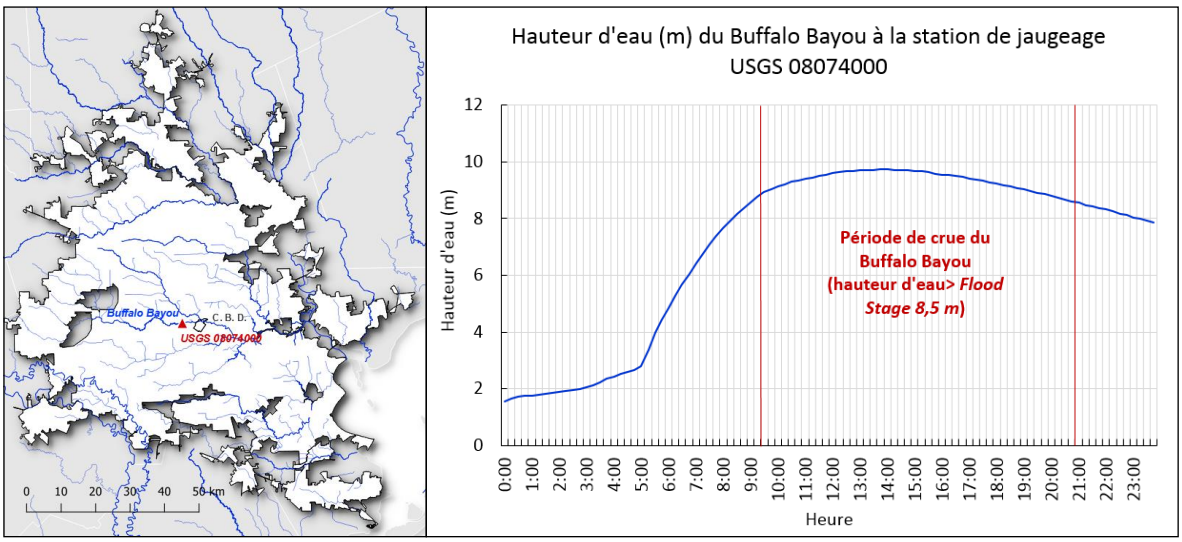


Figure 6.42 : Hauteur d'eau du Buffalo Bayou dans le centre de Houston, le 18 avril 2016 (Source : USGS-National Water Information System Web Interface)

Si jusqu'à présent, nous n'avons pu établir de relation spatiale entre intensité du phénomène physique et réponse virtuelle, les résultats s'avèrent cette fois positifs en ce qui concerne la réactivité virtuelle temporelle face au phénomène (comme le laissaient supposer les résultats de l'analyse spatio-temporelle seule) : l'élévation rapide du niveau de l'eau à la station débute à 5h le matin et le *flood stage* est atteint vers 9h. C'est effectivement dans cette tranche horaire de 9h-10h que les *topics* 4 (*Buffalo Bayou/flooded*), 7 (*flash flood/emergency*) et 2 (*flood/water/safe/home*) indiquent leurs plus fortes contributions dans la journée.

Pour nuancer les résultats du paragraphe 6.2.1.1[10], on observe ici une réactivité temporelle directe face au phénomène physique et à l'alerte lancée localement (certains tweets officiels marqueurs de phénomènes violents en cours seraient alors plus significatifs que d'autres sur le réseau). Cette réactivité temporelle est-elle organisée spatialement ? La figure 6.43 affiche les tweets géolocalisés rattachés aux *topics* 2, 4 et 7, émis entre 9h et 10h le 18 avril 2016.

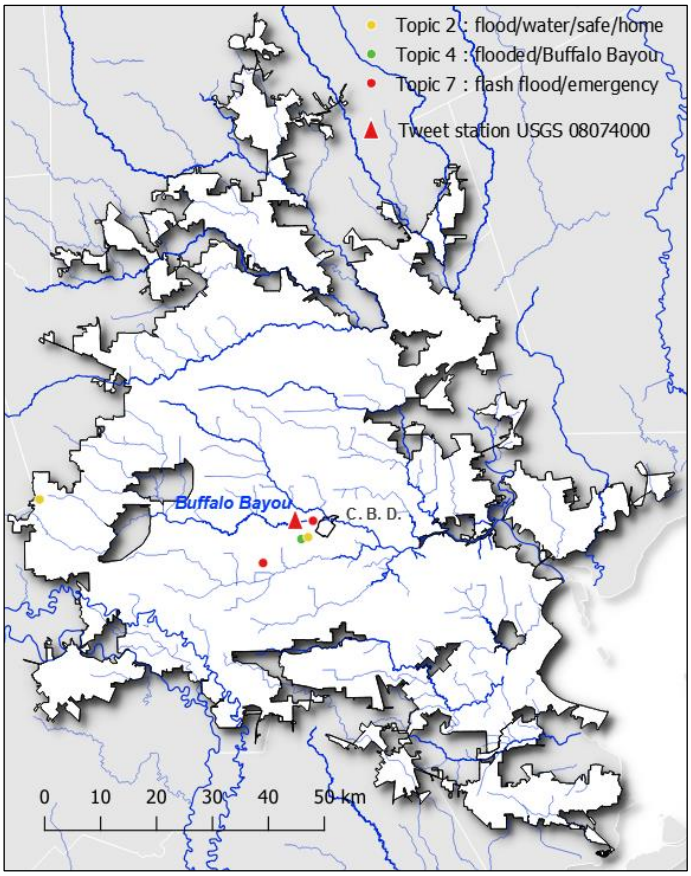


Figure 6.43 : Tweets de topics 2, 4 et 7 émis pendant la période de l'alerte crue, entre 9h et 1h le 18 avril 2016, Houston (C.Cavalière)

[10] Pour rappel, il s'agisait alors de l'étude de la réactivité virtuelle face à l'émission de tweets officiels marqueurs d'alerte ; celle-ci indiquait, pour le phénomène de pluies-inondations d'avril 2016, l'existence d'une réactivité virtuelle antérieure à l'émission de ces alertes officielles.

Le tweet le plus proche de la station de l'*USGS* se situe à 2,8 km ; le plus éloigné se trouve à 46 km. Ici encore, on observe un nouveau résultat positif : les tweets proches de la station (2,8 km ≤ station USGS ≤ 3,3 km) commentent directement le phénomène local d'inondations ("*#houston we have a problem montose* [nom d'un quartier proche du CBD et traversé par le Buffalo Bayou] *is flooded #rainydays #texasfloods*" ; "*and it flooded today, car gone… reasons why to never come to houston again*").

Les tweets plus éloignés (9,4 km ≤ station USGS ≤ 46,2 km) décrivent également des phénomènes locaux, survenus dans d'autres quartiers de la métropole : un phénomène local d'inondation qui bloque l'accès au quartier de Bellaire (sud-ouest du *CBD*) pour les *navetteurs* ; un affluent du Buffalo Bayou en crue dans l'extrême ouest de la métropole. Dans ce cas, même si la réactivité n'est pas entièrement consacrée à la crue du Buffalo Bayou, elle s'avère pour le moins concordante avec les phénomènes d'échelle très fine.

Ensuite, l'exploration sémantique des journées du 19 au 21 avril 2016 a été effectuée à partir de nuages de mots représentant les cooccurrences lexicales les plus fréquentes dans les horaires relevés précédemment. Ces nuages de mots sont représentés dans les figures 6.44 à 6.46 (les tendances qu'on peut dégager de l'analyse sémantique sont décrites à la suite des figures).

Figure 6.44 : Cooccurrences lexicales observées le 19 avril 2016 entre 2h et 6h

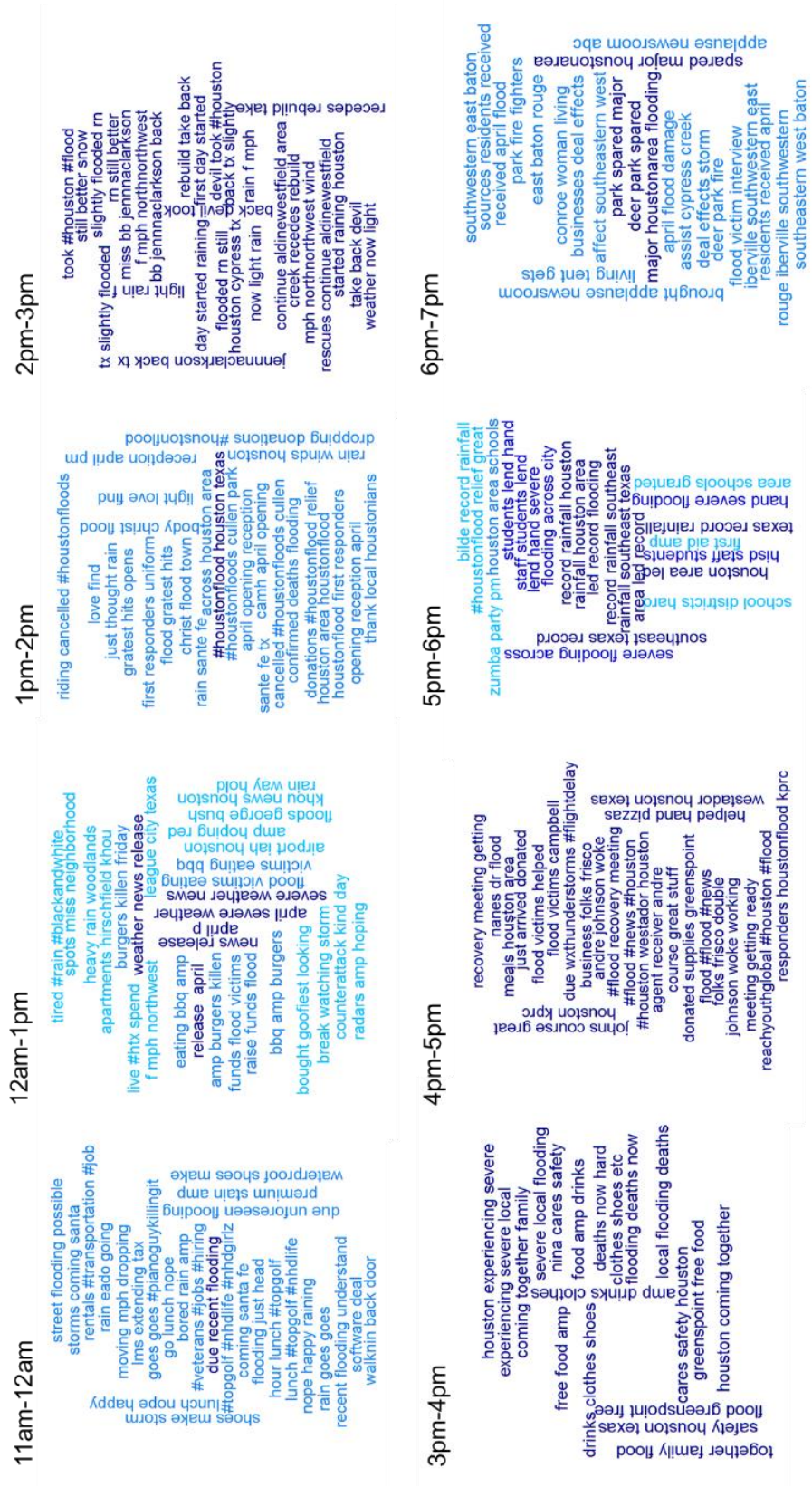


Figure 6.45 : Cooccurrences lexicales observées le 20 avril 2016 entre 11h et 19h

Figure 6.46 : Co-occurrences lexicales observées le 21 avril 2016 entre 10h et 14h

Peut-on alors interpréter, par la sémantique, les variations des dynamiques d'émission constatées par la représentation horaire des flux de tweets de crise et par leur cartographie ? Le 19 avril (figure 6.44), le pic d'activité enregistré entre 2h et 3h le matin est probablement lié à des acteurs publics et privés (d'après la tonalité du discours des tweets) : ils annoncent des phénomènes physiques toujours en cours ou attendus ("*storm conditions*", "*flood warning*"). Le discours se retrouve structuré autour de deux thèmes :

- les infrastructures ou sites ouverts/fermés : "*alert: hisd closes all schools due to weather*", "*a running list of what's open and closed during the rain in Houston*" ;
- les contraintes de circulation routière : "*commuters confront rainstrangled city*" ;

Ici, l'événement virtuel peut être la conséquence d'un effet de rediffusion d'une poignée de tweets de crise : le tweet contenant l'association *flood warning* est émis trois fois, le tweet mentionnant les difficultés de circulation se trouve deux fois et il en est de même pour les tweets mentionnant les étudiants ayant été renvoyés chez eux ainsi que le tweet fournissant un lien vers une page web inventoriant les magasins ouverts ou fermés pendant l'épisode de

pluies-inondations. Cet effet *retweet* s'est dissipé dans l'heure suivante, dont le discours se trouve davantage focalisé sur les conditions environnementales en temps réel : "*flooding strike*", "*some roads flooding*", "*face more severe weather*" ainsi que la mention d'un décès. Faut-il alors interpréter la baisse des émissions et le recentrage du contenu lexical sur un thème principal comme des indices témoignant d'une période plus critique ? Dans un second temps, si l'on compare la sémantique du pic enregistré entre 5h et 6h le matin, toujours le 19 avril, au créneau horaire précédent, c'est vraisemblablement de nouveau l'effet de rediffusion d'un unique message qui crée l'événement virtuel : "*must have wellingtons for h-town frequent 100 year flood*" (ce message représente 70% des tweets émis pendant ce créneau horaire). Ce message a été exclusivement diffusé le 19 avril 2016 entre 4h57 et 5h56 et se retrouve dans l'ensemble des territoires de la métropole. Que signifie-t-il ? Une recherche rapide sur le Web indique que Wellington correspond à une marque de bottes en caoutchouc ; sur Twitter, ces messages renvoient à une page web assurant la promotion de certains modèles pendant le phénomène. Au final, l'événement virtuel enregistré le 19 avril entre 5h et 6h provient d'un spam publicitaire multi-comptes et multi-sites.

Le 20 avril 2016, l'événement noté entre 12h et 13h (figure 6.45) se traduit par une activité éparse et diffuse (cf. carte de figure 6.39) et en effet, on retrouve plusieurs types d'informations qui s'entremêlent :

- un mince effet de rediffusion : les tweets mentionnant les conditions environnementales en cours "*weather news release*" et le soutien aux sinistrés "*raise funds flood victims*" représentent 27% de l'ensemble des tweets émis dans ce créneau ;
- la mention des inondations en cours : "flood george bush airport iah houston"
- la mention de pluies en cours : "*heavy rain woodlands*", "*break watching storm*".

Dans ce deuxième exemple, il est difficile d'avancer un comportement virtuel ou la survenue d'un phénomène ou événement dans le réel pour expliquer ce regain d'activité : en effet, les thèmes des perturbations environnementales et des mesures de soutien aux populations sont également inscrits sur le réseau dans les tranches horaires encadrant le créneau 12h-13h[11] : "*storms coming*", "*due recent flooding*" (entre 11h et 12h), "*donation houston relief*", "*just through rain*" (entre 13h et 14h). En revanche, le nouveau pic d'activité constaté le même jour entre 17h et 18h peut s'expliquer par l'apparition subite de trois informations uniquement visibles dans ce créneau (figure 6.45) : des cumuls pluviométriques et inondations records ("*record rainfall in the houston area led to record flooding*"), l'adaptation des cours pour les écoles affectées ("*school districts hard hit by flooding won't have to make up two days*"), ainsi que la mobilisation d'étudiants pour aider les sinistrés ("*students lend a hand during severe flooding across the city*").

[11] Cette tendance peut aussi tout simplement être liée au créneau horaire considéré : pendant le temps méridien, les utilisateurs sont généralement plus actifs sur le réseau.

Enfin, le 21 avril 2016, la carte de la figure 6.40 indiquait un événement virtuel à la fois concentré le long des rives du Buffalo Bayou et diffus à travers la métropole, qui se tarissait brusquement à partir de 13h. En explorant la sémantique de l'événement virtuel associé aux créneaux horaires alors retenus, on observe l'ensemble des facteurs soulignés ci-dessus (figure 6.46) :

- le matin, de 10h à 11h, on enregistre une réactivité face à un nouveau phénomène pluvieux annoncé : "*lms weather update! stay alert, ladies and gentlemen, to the weather*", "*new storms approaching: flash flooding, gusty winds*".

- Quelques messages sont émis deux fois, mais par le même compte : ils concernent les perturbations aériennes *"#iah is currently experiencing departure delays between 46mins and 1 hr due to thunderstorms #flightdelay*", un utilisateur annonçant la reprise de son travail "*happy to get back to work today; a nice break from being trapped indoors from the #houstonflood*" ainsi que le retour de la pluie "*it's currently raining hard in #houston*".

- A partir de 11h, on identifie le message suivant dans les associations lexicales majoritaires : "*unitedhealthcare and optum support people affected by floods in houston*" qui présente un comportement virtuel identique au message *Wellington* mis en évidence le 19 avril : il s'agit d'une assurance santé privée, dont les messages sont émis entre 11h49 et 12h54 le 21 avril 2016, et proviennent de différents comptes associés à différents sites, dans l'ensemble de la métropole.

Dans cette exploration sémantique, on pourra noter encore deux autres faits : le vocabulaire représenté dans les nuages de mots s'avère très impersonnel. En fait, le discours majoritaire mis en évidence ici (alertes, écoles fermées, dons, secours, assurances, localisation des lieux inondés) s'apparenterait davantage à un discours médiatique ou d'acteurs publics/privés de la société. Ainsi, lorsqu'on observe des pics d'activité dans les premières heures de la matinée du 19 avril, mentionnant les inondations, les infrastructures fermées ou encore la déclaration de l'état de catastrophe naturelle, on peut penser que cette activité n'est pas directement celle des sinistrés. En fait, il semble que les utilisateurs créent eux-mêmes l'information qu'ils diffusent lorsqu'on rencontre :

- de l'information au sens ambigu ou improbable : "*nope happy raining*", "*hour lunch #topgolf #nhdlife*" ;

- une orthographe volontairement erronée par l'émotion : "*rain rain go awayyyyyyy*" ;

- des indicateurs de messages de prières/espoirs : "*#prayforhouston*" ;

- de l'information traduisant le rôle de capteur environnemental de temps réel endossé par l'auteur du tweet : on trouve des tweets sauvegardant l'état d'un objet du territoire à l'instant ("*from the bridge over memorial from studemont and allen parkway #myhouston #houston #flood* [lien vers la photographie]") ou encore des tweets intégrant une dimension cognitive ("*last memorial day kolter elementary flooded. looks like we are few feet lower than that*" : le tweet introduit l'inondation survenue l'année précédente dans le même lieu comme

élément de comparaison avec le phénomène en cours ; l'école [*Kolter Elementary*] est utilisée comme objet de référence pour mesurer la gravité de l'inondation).

Le second fait que nous avons noté concerne les rapports entre la géolocalisation des tweets par GPS et la mention lexicale de lieux dans le tweet : même si les tweets sont précisément géolocalisés, l'information spatiale qu'ils véhiculent sous forme lexicale se rapporte souvent à une échelle géographique générale, à l'image des deux tweets suivants : "*heavy rains wreak widespread havoc throughout houston region*", "*residents rescued from floodwaters in nw* [north-west] *harris county*", qui sont tous deux émis depuis le *CBD* par des comptes restant actifs pendant toute la perturbation et cumulant quelques dizaines de tweets de crise géolocalisés. Ce type d'information, qui correspond à ce que nous avons nommé *information impersonnelle* plus haut, se rattacherait vraisemblablement à des émissions de comptes détenus par des médias. Qu'en est-il des tweets de crise qui mentionnent des lieux précis dans leur sémantique ? Dans les nuages de mots des figures 6.44 à 6.46, on pouvait identifier quelques mentions de lieux d'échelle fine : Greenspoint (figure 6.44, entre 4h et 5h ; figure 6.45, entre 15h et 16h) ou encore Deer Park (figure 6.45, entre 18h et 19h) ; or, l'information géolocalisée faisant mention d'un lieu précis n'est pas toujours localisée dans le lieu dont il est question :

- on trouve vingt tweets mentionnant des inondations et des évacuations à Greenspoint (nord du *CBD*) : huit tweets sont localisés dans le quartier en question ; le reste des tweets sont émis dans le *CBD* et arborent cette dimension impersonnelle que nous avons associée à des tweets à portée médiatique non créés par des individus directement affectés par la crise. En outre, dans le *CBD*, ils se trouvent associés aux émissions d'un unique compte et ne se positionnent pas toujours comme des relais de premiers tweets qui seraient émis depuis Greenspoint : par exemple, le 18 avril 2016, le compte du *CBD* qui émet des tweets contenant la mention spatiale Greenspoint est actif dès le matin alors que les émissions locales apparaissent dans la soirée (le 19 avril, ce comportement virtuel s'inverse).
- On trouve deux tweets mentionnant Deer Park (banlieue résidentielle et industrielle de l'est de Houston, à l'embouchure du Buffalo Bayou dans la Baie de Galveston) : l'un d'eux est émis depuis le lieu concerné ; le second est émis bien plus en amont du Buffalo Bayou, à l'ouest du *CBD.*

Bilan du premier test focalisé sur les dynamiques spatio-temporelles d'émergence de l'information géolocalisée de crise émise en réponse à un phénomène récurrent

Les variations d'émissions détectées témoignent de l'existence d'une réactivité locale face au réel, qui se manifeste par l'alternance de périodes d'agitation et de silence, ainsi que d'îlots ou de tweets épars (dans les lieux moins actifs virtuellement) sur les cartes (sauf dans le *CBD* qui reste actif en permanence) ; de plus, on peut détecter cette variabilité au moins à l'échelle de l'heure (et trouver une signification sémantique dans les structures observées). Quels facteurs déclenchent ces phases d'activité ?

Bilan du premier test focalisé sur les dynamiques spatio-temporelles d'émergence de l'information géolocalisée de crise émise en réponse à un phénomène récurrent (suite)

- les rediffusions de quelques messages qui se repèrent dans les tendances générales de la sémantique ;
- la survenue d'un nouveau phénomène perturbateur sur un territoire et des populations déjà perturbées ;
- l'activité opportuniste du spam publicitaire.

Deux questions restent en suspens :

- les pics de tweets les plus conséquents étaient associés à ces spams publicitaires et le vocabulaire majoritaire des tweets revêt cette dimension impersonnelle : que penser de la visibilité des individus tweetant à titre privé et de la représentativité des résultats vis-à-vis de l'activité de ces mêmes individus ?
- L'activité virtuelle du *CBD* est-elle le porte-voix des situations locales et des informations plus globales ? (ce qui expliquerait par ailleurs son activité virtuelle persistante constatée sur les cartes des figures 6.37 à 6.40).

6.2.2.3. Emergence et cohérence de l'événement virtuel en réponse à un phénomène extrême rare

Le paragraphe suivant présente la seconde série de tests effectués à partir des tweets de crise géolocalisés émis pendant la période du phénomène rare. La méthode vise toujours à explorer les relations entre temps, espace et sémantique. Comme dans le cas de l'ouragan, nous disposons d'émissions de tweets de crise géolocalisés bien plus conséquentes, nous pourrons affiner la résolution temporelle des analyses.

L'événement virtuel pendant la phase d'anticipation et les premières heures du phénomène.

Les tweets de crise géolocalisés de la métropole de Houston sont analysés à partir du 25 août 2017 (et jusqu'au 26 août, date à laquelle l'extrême sud de l'aire métropolitaine commence à être frappée). Les tweets sont clustérisés par l'algorithme DBSCAN, qui fonctionne selon deux paramètres sensibles (cf. chapitre 3) : la distance seuil, au-delà de laquelle deux tweets seront englobés dans deux clusters différents et le nombre de tweets minimal qui forment un cluster. Dans notre cas, ce choix s'avère complexe : comme dans les cas d'étude précédents, les tweets de crise géolocalisés constituent à la fois un maillage serré dans les quartiers du centre de la métropole (deux tweets ne peuvent ainsi être séparés que de quelques dizaines de mètres) mais également un maillage lâche dans les quartiers périphériques (la distance séparant deux tweets oscille généralement entre quelques

centaines de mètres et quelques kilomètres). Nous avons fixé ce critère de seuil à 500 mètres. Le nombre minimal de tweets attendus pour former un cluster est fixé à 5 (étant donné que, dans cette partie, nous travaillons sur la journée entière et non à une échelle temporelle fine [et donc sur des effectifs de tweets en théorie plus importants]).

La figure 6.47 présente les cartes des tweets clustérisés (en couleur) pour les deux journées des 25 et 26 août 2017 sur l'aire métropolitaine de Houston. Tout comme avec la LDA exécutée au paragraphe précédent, le partitionnement des tweets en clusters spatiaux nous confronte à des résultats mitigés : dans le cas des deux journées testées dans leur globalité par l'algorithme, un pourcentage non négligeable de tweets est considéré comme du bruit résiduel (car ces tweets sont trop éloignés les uns des autres pour former un groupe d'au moins cinq tweets). Le 25/08, 57,12% des tweets ne sont pas clustérisés ; le 28 août, ils sont 53,6% (figure 6.47) :

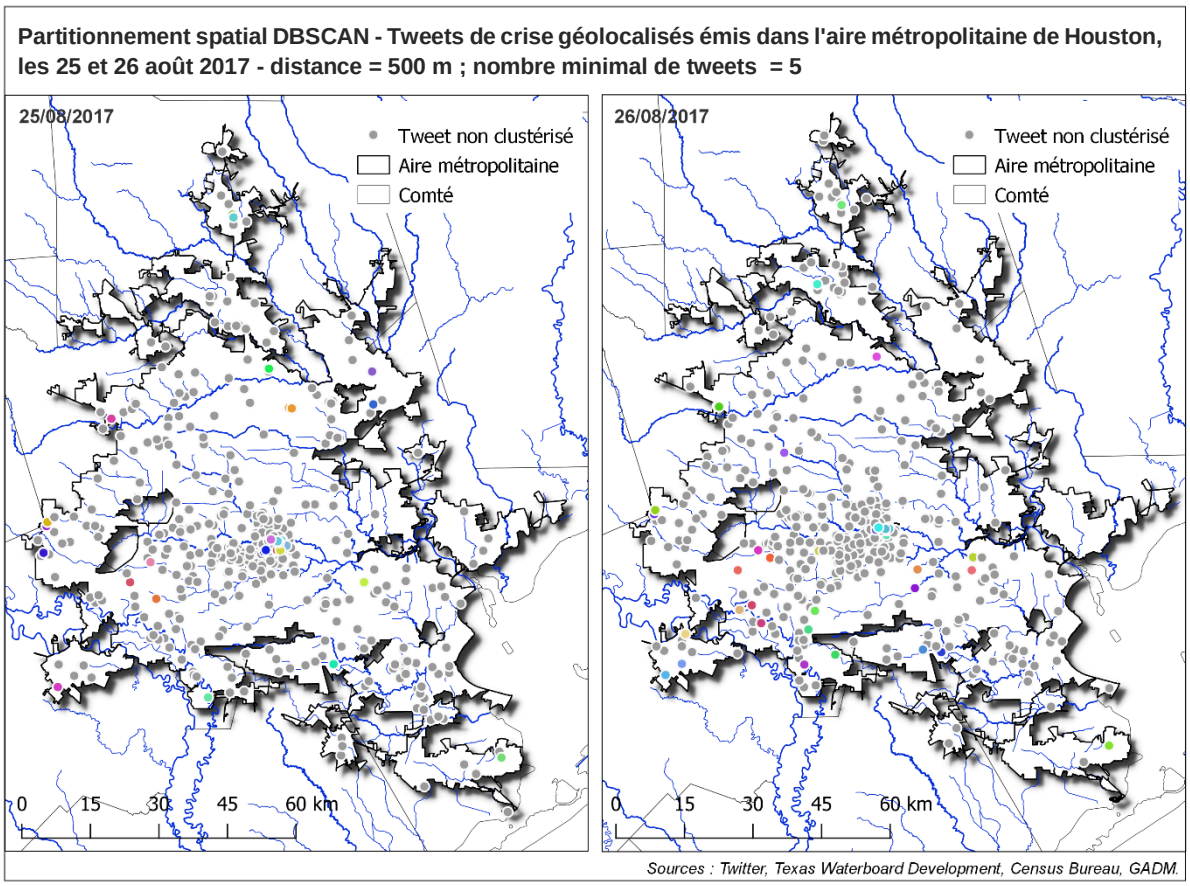


Figure 6.47 : Résultat du partitionnement spatial des tweets de crise géolocalisés par l'algorithme DBSCAN, Houston, 25 et 26 août 2017 (C.Cavalière)

Incontestablement, la tendance des tweets à se concentrer dans un espace restreint à quelques centaines de mètres s'avère être un comportement minoritaire. Pour la phase d'exploration sémantique, nous tirons au sort 50% des clusters spatiaux constitués. Que pouvons-nous faire ressortir de ces clusters et sont-ils suffisants pour donner une tendance

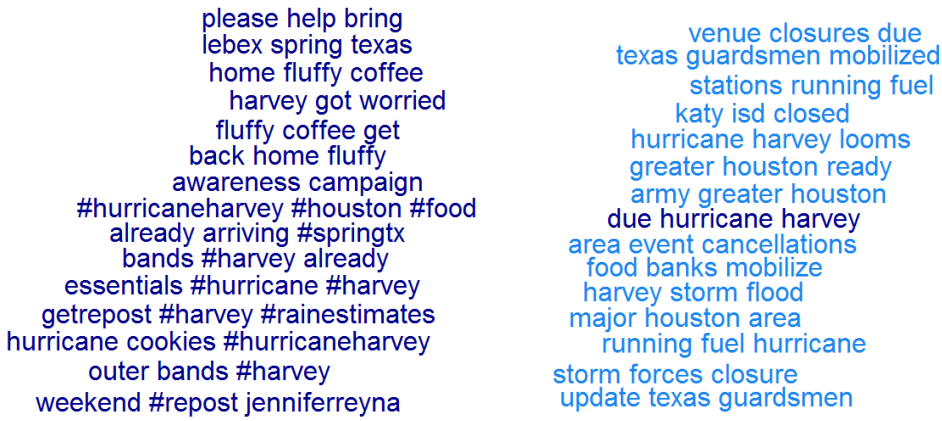
des propriétés de l'événement virtuel pendant cette première phase ? En premier lieu, les clusters tirés au sort mettent en évidence trois types de contenus générés et diffusés pour les deux journées analysées. Le tableau 6.7 présente le pourcentage de clusters agrégeant des tweets émis par des automates, par des acteurs de la société ou présentant des contenus hétérogènes qu'on peut assimiler à des informations centrées sur les individus (colonne *Tous types de contenus*) :

Tableau 6.7 : Identification du contenu lexical des clusters

% de clusters / Date	Tous types de contenus	Automates	Acteurs de la société
25/08	61,5%	30,7%	7,7%
26/08	73,3%	20%	6,7%

Dans le cas précis du phénomène extrême rare, on trouve davantage de contenu lexical qui apparaît personnel (centré sur l'individu), respectivement 61,5% et 73,3% pour le 25 et le 26 août. On peut rapidement caractériser cette information centrée sur l'individu selon deux critères principaux :

- les nuages de mots qui représentent les cooccurrences lexicales de ce type d'information apparaissent bien moins structurés que les nuages représentant des informations émises par des automates ou des informations à consonance officielle (figure 6.48).



(a). Tweets d'individus *(b). Tweets à consonance officielle*

Figure 6.48 : Comparaison d'information lexicale personnelle et à consonance officielle

Dans l'exemple ci-avant, le nuage (a) contient des tweets centrés sur les individus dont nous soumettons un échantillon ci-après : "*Hurricane Harvey got you worried? don't stress, I have a plan! #hurricane #hurricaneharvey*", "*Back home with fluffy coffee to get ready for hurricane Harvey's*", "*Haha hurricane cookies!!! #hurricaneharvey #houston #food Spring, Texas* ". En revanche, le nuage (b) contient des tweets à consonance officielle au discours moins ambivalent : "*The salvation army of greater Houston is ready to respond to hurricane Harvey*", "*Texas food banks mobilize for hurricane Harvey relief*", "*Update: more than 900 Texas guardsmen mobilized in support of hurricane Harvey*".

- l'information centrée sur l'individu est susceptible de contenir davantage de hashtags ainsi que des émoticônes. En témoignent les tweets suivants, considérés de nouveau comme des tweets relayant une information personnelle non officielle : "*#parkingofficer #houstontexas #hurricaneharvey #midtownhouston #worldstar #iwannagohome*", "*me 😂😂😂 #houston #htown #htx #mayweathermcgregor #hurricaneharvey #hurricane @ jersey village, texas"*, "*Even a hurricane won't stop me from thriving! #hurricaneharvey #prepared #morningperson #duo*".

Les deux figures suivantes (6.49 et 6.50) localisent les clusters analysés pour les deux journées consécutives et présentent, sous forme de nuages d'associations lexicales, les cooccurrences rencontrées dans les tweets agrégés en clusters (dans les nuages, plus la couleur est foncée, plus l'association lexicale est récurrente et inversement ; si le nuage ne contient qu'une seule couleur, alors les associations lexicales ne présentent qu'une seule occurrence dans le cluster). Nous avons retenu les clusters contenant de l'information associée aux contenus personnels hétérogènes (NB : les tweets clustérisés sont représentés par des couleurs attribuées au hasard : il n'y a aucune hiérarchie entre les clusters).

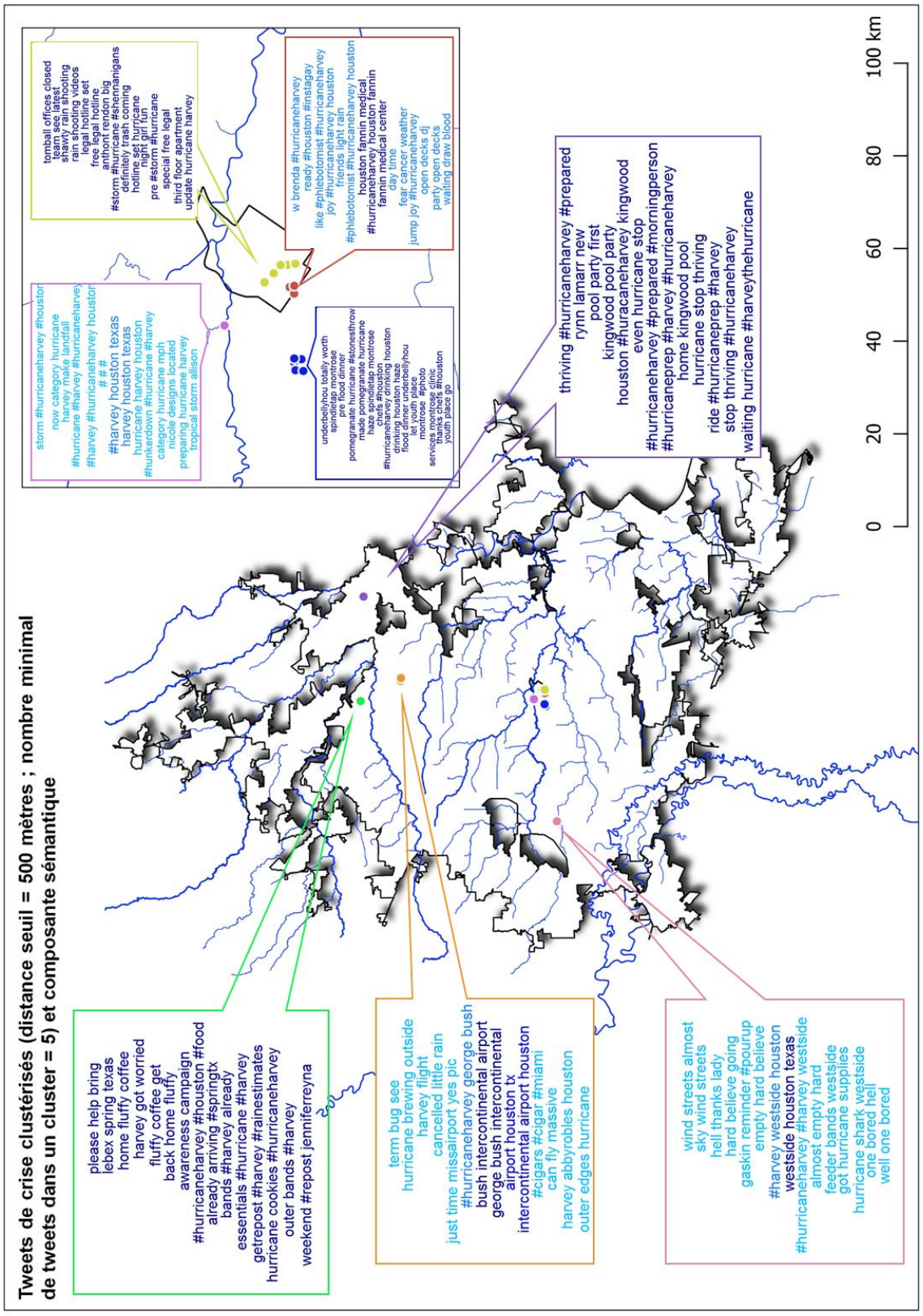


Figure 6.49 : Exploration lexicale des clusters tirés au sort contenant de l'information personnelle, le 25 août, Houston (C.Cavalière)

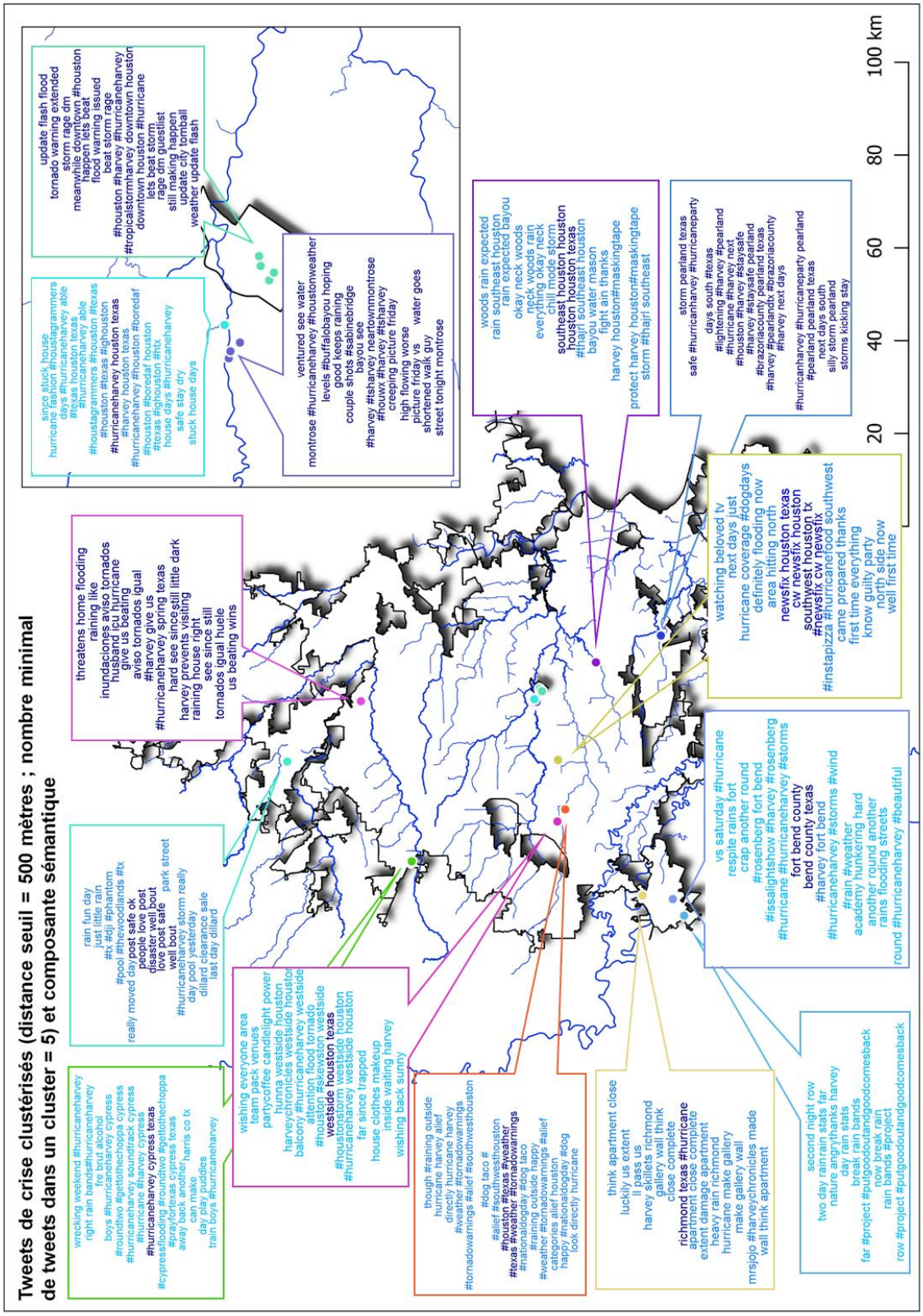


Figure 6.50 : Exploration lexicale des clusters tirés au sort contenant de l'information personnelle, le 26 août, Houston (C.Cavalière)

Sans équivoque, et malgré l'omniprésence des hashtags *#hurricaneharvey* et *#harvey*, on constate une variabilité spatiale des cooccurrences lexicales majoritaires dans chaque cluster :

- le 25 août 2017, si l'on se réfère au cluster rouge du *CBD* (cf. figure 6.49), on trouve de l'information émise près d'un établissement de santé ("*No fear about cancer or the weather #hurricaneharvey @ Houston Fannin Medical Center*") ; dans le cluster orange situé au nord du *CBD*, on trouve des tweets émis à l'aéroport intercontinental par les usagers quittant Houston avant l'arrivée de l'ouragan ("*need to get out before #harvey comes through! ✈□ @ george bush intercontinental airport in houston, tx*") ; dans le cluster rose situé à l'ouest du *CBD*, on peut identifier des tweets indicateurs des conditions environnementales locales du moment ("*Clear sky, no wind, streets almost empty. Hard to believe it's going to be chaos by tomorrow*").
- le 26 août 2017, les thèmes évoluent (cf. figure 6.50) : dans le *CBD,* les tweets du cluster bleu foncé témoignent de la préoccupation des riverains du Buffalo Bayou ("*couple of shots from #sabinebridge. #buffalobayou is high and flowing, but it's been worse*", "*a shortened walk for this guy on #nationaldogday. the bayou appears to have crested*"). Dans le cluster bleu localisé à l'extrême sud de la métropole et repérable grâce au nom de *Pearland*, on constate la récurrence des appels à la prudence ("*Storms kicking up again! stay safe out there! #houston #harvey #staysafe @ Pearland, Texas*", "*Y'all pls be safe. #hurricanharvey #hurricaneparty @ Pearland, Texas*").

Mais ce surplus d'informations personnelles (par rapport aux résultats précédents fondés sur les phénomènes récurrents) est-il utile ? Certains tweets se positionnent indéniablement comme des témoins directs des conditions environnementales, des comportements ou de l'anticipation de leur auteur face au phénomène prévu. Malgré tout, dans ces clusters de tweets spatialement proches où l'on trouve un enchevêtrement d'informations, il est parfois difficile d'identifier des structures lexicales cohérentes à l'intérieur des clusters :

- un objet particulier du territoire peut apporter cette cohérence lexicale : le 25 août à l'aéroport, on pouvait distinguer sept tweets dont les auteurs annonçaient qu'ils quittaient Houston pendant qu'ils le pouvaient encore (cf. figure 6.49, cluster orange) ;
- le début du phénomène physique marque également une cohérence spatiale des propos : le 26 août, dans le cluster vert du nord-ouest de la métropole mentionnant le quartier de *Cypress* (cf. figure 6.50), on peut identifier des tweets précis qui annoncent explicitement l'arrivée de pluies intenses et d'inondations ("*so far so good...however, torrential rains are here! @ cypress, texas*", "*the neighborhood floods at 125 ft...last time it was 127.5. hoping we can make it through*").

Le problème reste que ces tweets au discours personnel sont noyés dans d'autres messages au contenu ambigu : "*Anyone wanna guess what's on the bar menu at #dekemanor*

tonight? #hurricaneharvey @ Cypress, Texas", "*Gave myself a #manicure I won't quit my day job. #hurricaneproblems #hurricaneharvey*", "*Don't look directly at the #hurricane. #harvey @ Cypress, Texas*". Les tweets que la partition établie par DBSCAN a considéré comme du bruit résiduel obéissent-ils à ce même comportement ? Pour éclairer cette question, nous avons sélectionné des tweets écartés de l'analyse précédente dans deux milieu différents : le quartier aisé de *Greater Memorial* à l'ouest du *CBD*, longé par le *Buffalo Bayou*, et le quartier populaire de *Southeast*, au sud-est du *CBD*.

Au final, on observe cette même tendance (figure 6.51.a) : quels que soient le quartier et la journée considérés, on trouve en moyenne seulement 38% de tweets exprimant des témoignages directs et non implicites traduisant des observations environnementales ou des comportements (NB : afin de donner une idée précise des tweets que nous considérons comme témoins directs ou comme ambigus, le tableau est accompagné de quelques exemples, cf. figure 6.51.b). Notons néanmoins que la proportion de tweets véhiculant des images explicites augmente, quel que soit le quartier considéré, lorsqu'on se rapproche de la crise[12].

Quartier / Date	Greater Memorial	Southeast
25/08/2017	16,7	32
26/08/2017	53,8	50

(a). Pourcentages de tweets témoins explicites dans les deux quartiers explorés

Tweet au contenu explicite	Tweet au contenu ambigu
4th floor should be enough if it floods #hurricaneharvey (comportement)	*Current currents. Waiting out #hurricaneharvey. tsu vs famu #teamtsu #teamtxsu #teamgreen*
The bayou has risen since hurricane Harvey has become higher @ friends of MacGregor Park (observation environnementale)	*You cooking in the hurricane chef?! you buyin' i'm fryin... ... just #followthesmell*
Last year April 2016 by the crib, and this wasn't even a hurricane... yep it's time to go (appréhension et comportement)	*When you forget your favorite chocolate #harvey*

(b). Quelques exemples de tweets classés en fonction du sens de leur contenu sémantique

Figure 6.51 : Exploration du contenu lexical des tweets épars des quartiers de Southeast et de Greater Memorial, Houston, 25 et 26 août 2017

Ainsi, on ne distingue pas de structures lexicale et spatiale flagrantes : la sémantique de l'information personnelle est souvent beaucoup trop diversifiée pour faire émerger une tendance générale, à l'échelle de la journée. En dehors de l'information à consonance officielle

[12] La proximité temporelle au début de la crise doit alors vraisemblablement recadrer les sujets de conversation sur le phénomène et les événements vécus en temps réel, au détriment de ces informations ambiguës et annexes qui accompagnent le tableau dans la figure 6.51.

et de ces lieux ou objets témoins comme l'aéroport ou les rives des bayous qui fédèrent des tweets rapportés à un comportement (quitter la ville par les airs) ou à un élément concret du territoire (observer la rivière en crue), l'événement virtuel consécutif au phénomène extrême rare reste constitué d'une multitude de signaux divers, que nous nommons ici *micro-événements individuels,* définis comme suit : le micro-événement d'échelle individuelle caractérise une réactivité de l'utilisateur sans qu'on puisse identifier de logique spatiale ou environnementale directe expliquant la contribution de l'utilisateur : le discours d'un tel tweet peut être discordant par rapport au discours des tweets voisins, ou discordant vis-à-vis des conditions environnementales. Dans les cartes des figures 6.49 et 6.50, on pouvait relever des indices de la présence de tels tweets :

- le cluster violet de la figure 6.49 (nord-est du *CBD*) est uniquement constitué de ces tweets à dimension égocentrée : "*one last bike ride #hurricaneprep #harvey #hurricaneharvey #kingwood tx*", "*even a hurricane won't stop me from thriving! #hurricaneharvey #prepared #morningperson*" ;
- le cluster vert de la figure 6.50 (extrême nord du *CBD*) contient des tweets témoins de la préparation antérieure à l'arrivée du phénomène : "*we just need a half dozen sandbags to complete our flood barrier*", "*a halfdozen of #flashflood warnings and all the sudden no one parks in the street*", aux côtés desquels on trouve un tweet égocentré sur les activités pratiquées la veille : "*despite the rain, had a fun day at the pool yesterday #pool #thewoodlands #harvey*".

L'activité virtuelle pendant la phase la plus intense du phénomène.

Le 27 août 2017 correspond à la journée pendant laquelle la métropole de Houston est frappée de plein fouet ; c'est également la journée pendant laquelle le plus fort pic d'émissions de tweets de crise géolocalisés a été atteint (soit 2 011 tweets géolocalisés non rattachés à des comptes officiels). Nous relançons les analyses consistant à allier spatial, temporel et sémantique à une échelle temporelle plus fine (jusqu'alors, nous avons testé les résolutions de la journée et de l'heure). En testant la finesse de la résolution temporelle des tweets de crise géolocalisés, nous essayons, à travers l'observation de l'émergence des tweets sur un intervalle temporel court, de détecter l'éventuelle existence de cette cohérence entre distance et sémantique.

Dans un premier temps, nous avons simplement souhaité vérifier si la réactivité temporelle qu'on avait observée lorsque le Buffalo Bayou dépassait le seuil du *NWS flood stage* le 18 avril 2016, se répètait ici. Si l'on observe des pics de tweets concordant avec l'élévation du niveau des eaux, alors ceux-ci seront explorés (dans le cas contraire, nous nous séparerons des données physiques officielles pour nous concentrer définitivement sur le virtuel). Dans ce nouveau test, nous conservons la station *USGS 08074000* du Buffalo Bayou comme témoin : localisée dans le centre de Houston, elle se trouve sur le territoire qui capte le plus de tweets en temps normal et qui s'avère également des plus actifs en temps perturbé.

Le graphique suivant (figure 6.52) affiche la courbe indiquant le niveau des eaux du Buffalo Bayou dans la journée du 27 août, à une résolution temporelle de 15 minutes. Précisons que les données fournies par le portail *NWIS Waterdata* étaient tronquées à partir de 15h donc nous les avons complétées avec les données contenues dans les tweets émis par la station entre 16h30 et 23h30 et envoyées d'heure en heure (d'où la discontinuité de la courbe au-delà de 15h). La courbe du niveau des eaux est complétée par la courbe des émissions de tweets de crise géolocalisés agrégés au même pas de temps de 15 minutes (l'horodatage des tweets sera toujours arrondi au quart d'heure inférieur : un tweet émis à 9h38 est donc tronqué à 9h30).

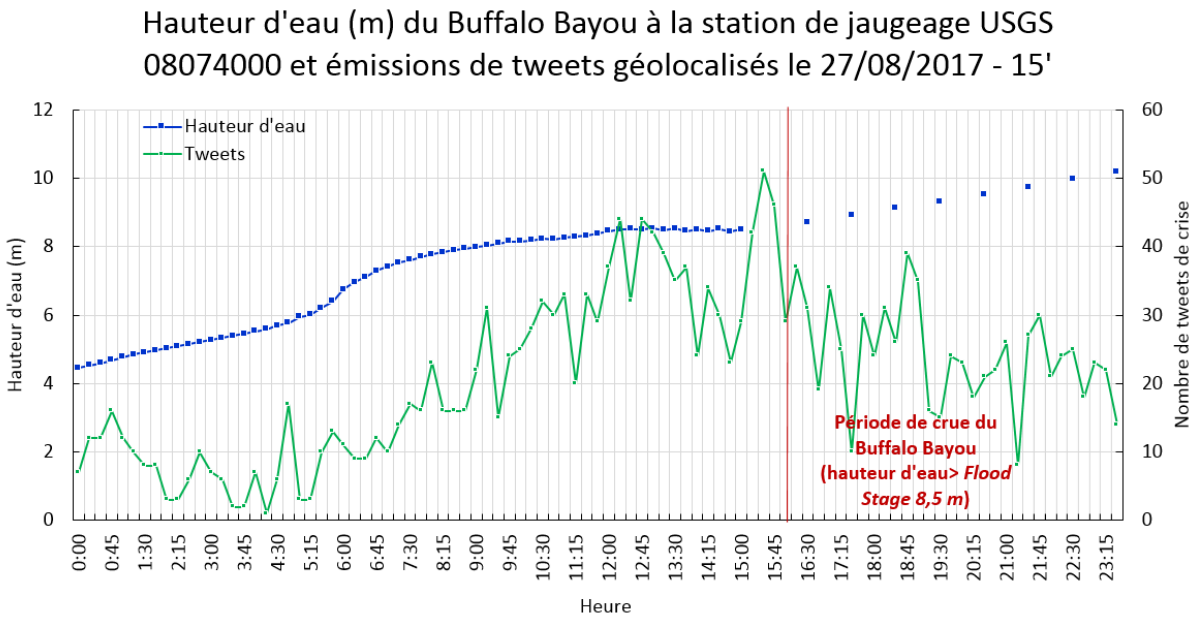


Figure 6.52 : Hauteur d'eau du Buffalo Bayou dans le centre de Houston et émissions de tweets de crise géolocalisés le 27 août 2017

Le graphique de la figure 6.52 indique que le niveau des eaux atteint le seuil d'alerte du *NWS* (NB : 8,5m) entre 16h et 16h30 (cf. tweet émis par la station à 16h30 : *"#USGS08074000 Buffalo Bayou at Houston, TX Height: 28.53ft (28ft)*"). Cette fois-ci, il n'y a pas de réponse manifeste après le dépassement de ce seuil : en fait, le plus fort pic de tweets de la journée est enregistré dans l'heure précédente, entre 15h30 et 15h45. De même, si l'on observe quelques pics de tweets dans les premières heures de la matinée (4h45-5h) à l'échelle du quart d'heure, la tendance de l'activité tweeting se positionne comme écho à la montée des eaux : malgré les oscillations de la courbe des tweets, l'activité virtuelle de crise est à la hausse dès que le niveau des eaux s'élève. En revanche, le pic de tweets du milieu de l'après-midi (15h30-15h45 marque une rupture de l'événement virtuel : l'activité décroît alors que le niveau des eaux continue d'augmenter (si cette période correspond au plus fort de la crise dans la ville, alors l'activité de l'événement virtuel décroît pendant cette phase critique). A partir de la figure 6.52, nous avons relevé quatre périodes à explorer (les quarts d'heure précédant et

suivant la temporalité d'intérêt sont conservés dans l'éventualité d'un changement spatial et/ou thématique dans les tweets) :

- 4h30-4h45-5h00 : pic matinal ;
- 9h00-9h15-9h30 : pic pendant la montée des eaux ;
- 15h00-15h15-15h30-15h45 : plus haut pic dans la journée ;
- 21h00-21h30 : deux pics séparés par un creux.

Pour faciliter la représentation cartographique de l'information lexicale, nous agrégeons les tweets par maille de dix kilomètres (afin de prendre en compte la variabilité des densités de contenus numériques en fonction des lieux de la métropole). Sur les cartes, nous faisons disparaître les mailles actives (qui sont remplacées par un point représentant leur barycentre), puisque le tout premier objectif du test ne consiste pas à se concentrer sur les quantités ou localisations précises (étant donné le maillage lâche que peut former l'événement virtuel) des tweets mais sur leur discours. Un nuage de mots représentant le contenu lexical de chaque maille est centré sur chaque point. La sémantique est catégorisée en fonction de différents thèmes identifiés à la lecture des tweets de crise, et représentée par les couleurs suivantes :

- sentiments/émotions/prières : violet
- intempéries en cours : rouge
- inondations : bleu
- dommages plus ou moins intenses aux infrastructures : noir
- observations environnementales locales : vert
- action individuelle, collective, invitation à l'engagement ou information utile à l'ensemble du réseau : orange
- appels à l'aide, intervention des secours : fuschia

Pour terminer, la couleur attribuée à chaque point correspond à la couleur du thème majoritaire identifié dans le nuage de mots (si les thèmes sont en proportions équivalentes, alors le point figure en gris). Les figures 6.53 à 6.56 présentent les cartographies résultantes (NB : pour distinguer le lexique du centre de la métropole [*CBD* et territoires compris dans un rayon de moins de dix kilomètres du *CBD*, donc les mailles limitrophes], la police adoptée est la suivante : Bradley hand).

Les figures 5.53 et 5.54, en page suivante, présentent les deux premières périodes d'augmentation puis de diminution soudaines du volume de tweets émis pour la journée du 27 août 2017 à l'échelle du quart d'heure, identifiées sur le graphique de la figure 6.52. Il s'agit des créneaux 4h30-5h15 et 9h-9h45.

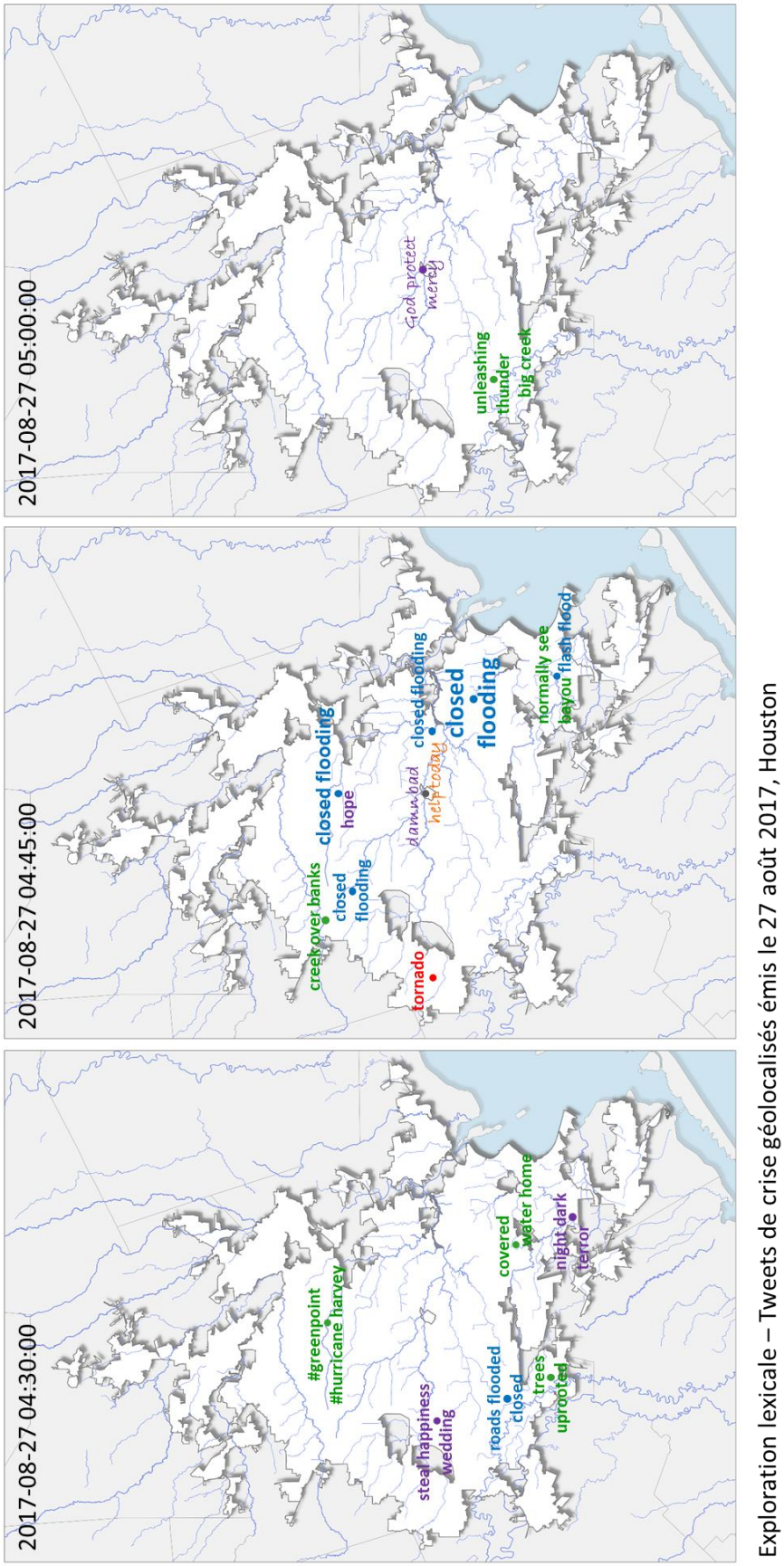


Figure 6.53 : Sémantique des tweets de crise géolocalisés émis entre 4h30 et 5h15 le 27 août 2017 – Résolution temporelle de 15' (C.Cavalière)

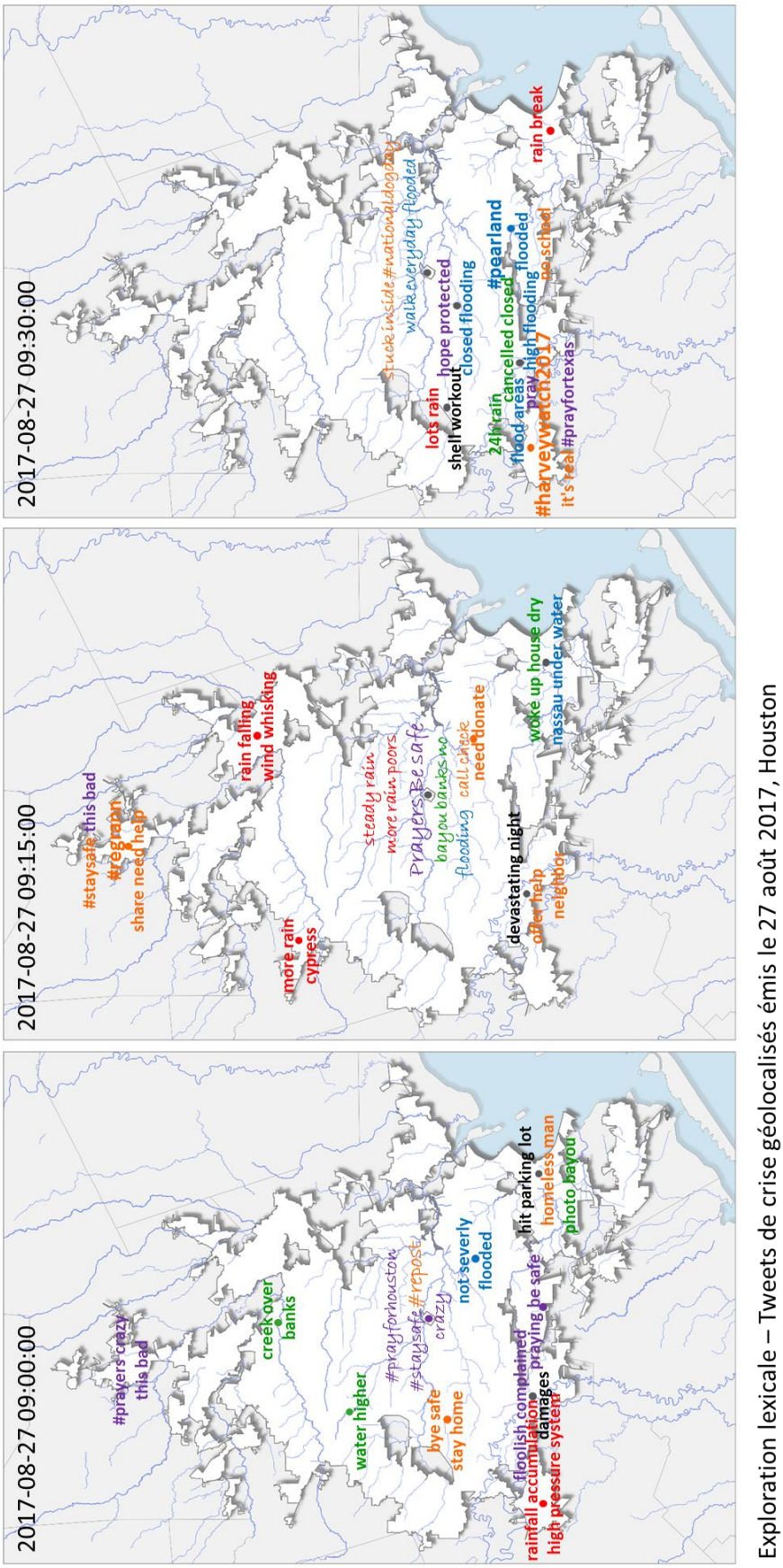
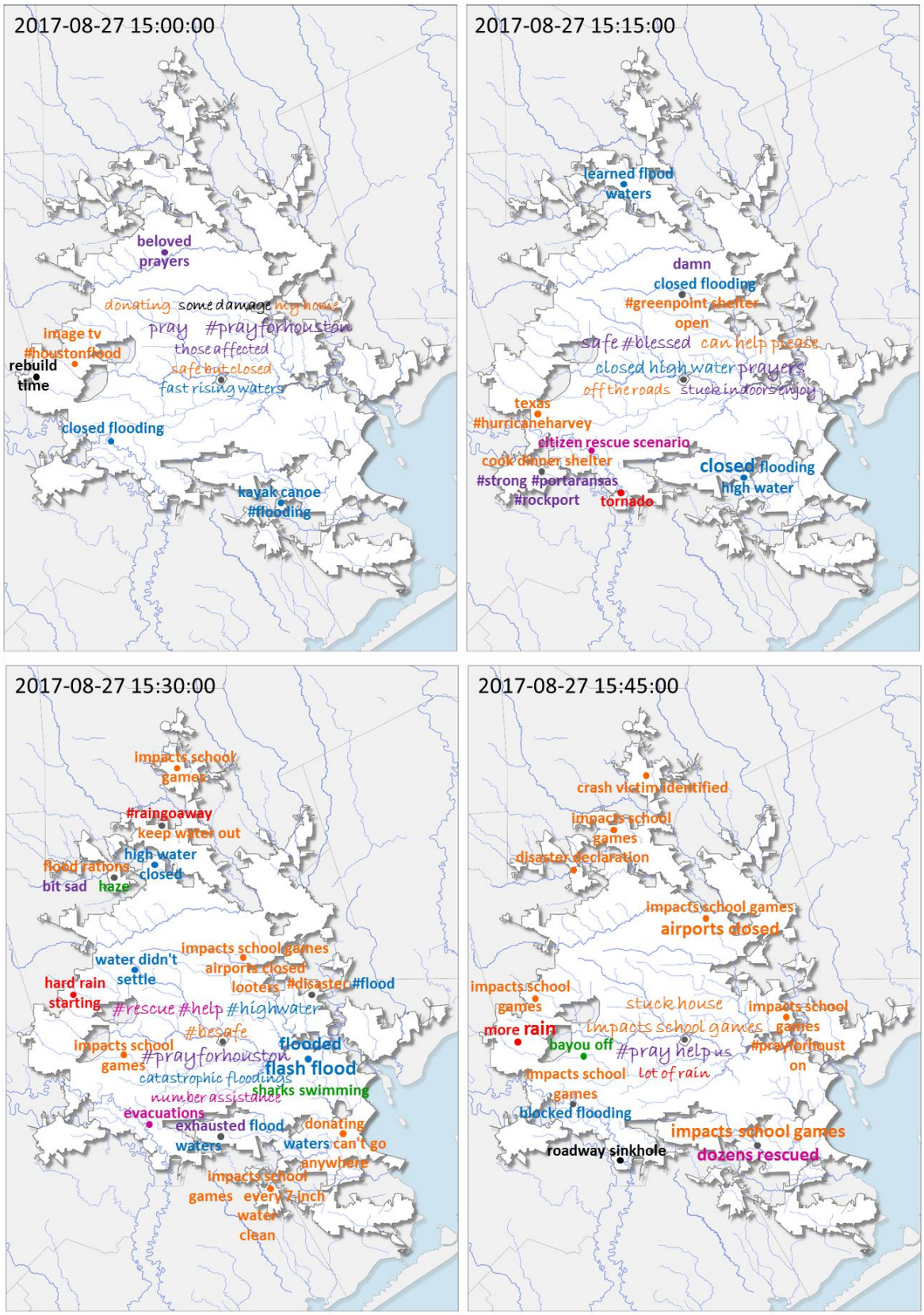


Figure 6.54 : Sémantique des tweets de crise géolocalisés émis entre 9h et 9h45 le 27 août 2017 – Résolution temporelle de 15' (C.Cavalière)

Dans le cas de l'ouragan, à quoi sont dus les pics de tweets de crise géolocalisés ? A partir des figures 6.53 et 6.54, on peut distinguer l'apparition de deux facteurs générateurs de pics de tweets de crise géolocalisés, qu'on avait déjà identifiés dans le cas des phénomènes récurrents. Le premier facteur correspond à la diffusion rapide, à un moment donné, d'une information à consonance officielle, focalisée sur un thème unique : ici, entre 4h45 et 5h, il s'agit de routes inondées fermées à la circulation (cf. figure 6.53 : "*closed due to flooding in #jerseyvillage on hwy 290 nw fwy frontage rd outbound after w rd and fm 1960 #traffic*"). Ce facteur explicatif se complète par l'activation des cellules d'activité virtuelle du centre de la métropole, qui cumulent, pour tous les quarts d'heure explorés (à partir de 9h), entre 49% et 72% des tweets émis (cf. figure 6.54). En ce qui concerne le pic enregistré entre 9h15 et 9h30, on distingue également une multiplication des informations relatives aux précipitations en cours (mots et associations de couleur rouge sur la figure 6.54) dans le centre et dans le nord de la métropole (ou il pourrait encore tout aussi bien s'agir de l'effet de l'heure, puisque l'horaire d'agitation correspond à une horaire normale d'activité virtuelle).

Distingue-t-on une logique spatiale et lexicale dans les tweets par les nuages de mots colorés par thèmes ? Une logique de spatialisation est manifeste : les lieux d'activité virtuelle indiquant des inondations sont bien voisins des cours d'eau, qu'il s'agisse de l'information à consonance officielle ou de l'information personnelle, à l'image de ce tweet, émis entre 9h et 9h15 (cf. figure 6.54), le long du Berry Bayou, au sud-est du CBD : "*We are blessed to not be severely flooded*". Et comme ce tweet l'indique, on trouve de multiples marqueurs d'intensité des phénomènes mais qui s'avèrent émis de manière irrégulière : par exemple, le long du Berry Bayou, il n'y a que cette trace mentionnée qui témoigne de l'intensité de l'inondation (on ne sait donc pas comment le phénomène évolue au cours du temps). De même, les marqueurs d'intensité de précipitations, identifiés dans le centre entre 9h15 et 9h30 (*more rain poors*, *steady rain*) disparaissent dans le quart d'heure suivant. En revanche, à l'intérieur d'une maille, les informations sémantiques ont tendance à s'entremêler, ce qui est particulièrement palpable dans le centre : entre 9h15 et 9h30, on trouve ainsi des témoins des intempéries et indondations, mais également des messages de prières ou diffusant des consignes de prudence (figure 6.54).

Les figures 6.55 et 6.56, dans les deux pages suivantes, affichent la sémantique de l'activité virtuelle enregistrée pendant le pic le plus important de la journée, entre 15h et 16h, puis pendant l'alternance pic/creux, détectée dans la soirée entre 21h et 21h45 (la résolution temporelle d'agrégation des tweets reste le quart d'heure).



Exploration lexicale – Tweets de crise géolocalisés émis le 27 août 2017, Houston

Figure 6.55 : Sémantique des tweets de crise géolocalisés émis entre 15h et 16h le 27 août 2017 - Résolution temporelle de 15' (C.Cavalière)

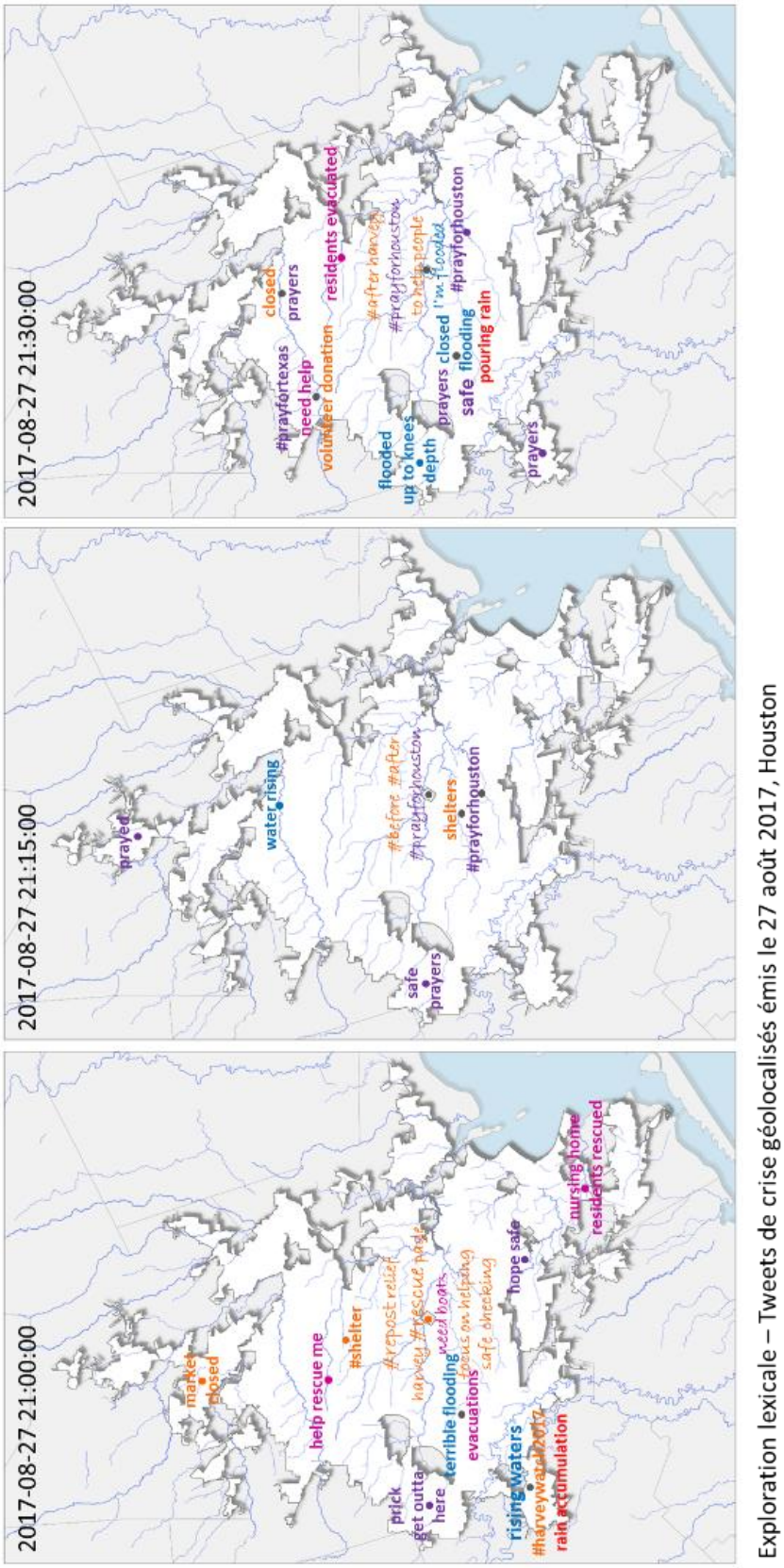


Figure 6.56 : Sémantique des tweets de crise géolocalisés émis entre 21h et 21h45 le 27 août 2017 - Résolution temporelle de 15' (C.Cavalière)

Une nouvelle fois, l'activité virtuelle la plus conséquente se concentre dans le centre (et quelle que soit la période considérée, cf. figures 6.55 et 6.56), qui revêt toujours une sémantique variée (messages de prières, état des inondations, précipitations, informations relatives à la sécurité ou au partage de liens pour collecter des dons ou faciliter l'organisation des secours : *harvey #rescue page* [figure 6.56, entre 21h et 21h15]), alors que les autres territoires de la métropole s'inscrivent davantage dans le registre des mentions d'inondations ou encore de secours : c'est notamment le cas à partir de 15h30, créneau pendant lequel on détecte la première mention explicite de secours et d'évacuation (figure 6.55 : *"#rescue #help #flood*"). De la même manière que pendant les heures matinales, les marqueurs locaux d'intensité des précipitations et des inondations se montrent spatialement épars et temporellement irréguliers : par exemple, l'extrême-ouest de la métropole compte deux témoins de précipitations, respectivement émis à 15h42 puis à 15h58 (cf. figure 6.55 : "*hard rain starting again in cypress/katy*" et "*rain rain rain and more rain here. #floods*"). En revanche, le soir, ce territoire n'enregistre plus de témoin direct des conditions météorologiques locale : on a sans doute un tweet pouvant indiquer implicitement la poursuite des intempéries "*ok #harvey ya miserable prick ya...get the f*** outta here already*!" (figure 6.56 entre 21h et 21h15) mais complété par la première et unique mention contenant un indice d'intensité de l'inondation locale (cf. figure 6.56 à 21h40 : "*update on storm harvey, up to knee length depth in most places my neighborhood*").

Que dire des facteurs à l'origine des pics d'activité virtuelle respectivement détectés entre 15h30 et 15h45 puis du creux soudain détecté entre deux pics de 21h à 21h45 ? Le pic identifié entre 15h30 et 15h45 peut s'expliquer par le facteur suivant, qu'on avait également identifié lors des événements récurrents : la rediffusion d'un unique message en divers lieux de la métropole : "*hurricane harvey impacts school openings, football games* [Lien vers une page web]". Ce message adopte le comportement identique aux messages ambigus qu'on avait mis en exergue lors du phénomène de pluies inondations d'avril 2016 : il n'est diffusé rapidement que le 27 août 2017 entre 15h42 et 15h59, puis il disparaît définitivement de l'événement virtuel ultérieur. A cette rediffusion d'un unique message s'ajoutent des mentions de la survenue de crues éclair, par un unique compte, dans le quartier de Pasadena au sud-est du *CBD*. Pendant la soirée, l'alternance pics/creux résulte vraisemblablement de la variabilité quantitative de la cellule d'activité du centre mais également de nouvelles mentions d'inondations survenues dans le sud-ouest de la métropole, visibles sur les cartes des tweets émis entre 21h et 21h15 "*rising waters*", "*terrible flooding*", puis entre 21h30 et 21h45 "*I'm flooded*", "*up to knee depth*").

Les tweets émis dans le centre de la métropole, c'est-à-dire le lieu qui concentre toujours le plus de tweets de crise géolocalisés, quelles que soient les conditions, sont-ils significatifs ? Quelle place occupent-ils dans l'événement virtuel ? On peut estimer à 77% les tweets émis dans le *CBD* et dans un rayon de dix kilomètres autour du *CBD* qui ne sont pas directement rapportés à des phénomènes locaux vécus par les auteurs des tweets. En fait, on peut distinguer deux comportements antagonistes entre le centre et les périphéries de la

métropole : la journée, les messages émis depuis le centre concernent essentiellement l'émotion par les prières (cf. figure 6.54 et 6.55 : "*reminder. we gon be alright. #prayforhouston #hurricaneharvey @ houston, texas*", "*my heart hurts. this is my city! now it's under water. surreal! please #prayforhouston. #harvey*") ; cette même thématique reste marginale dans les périphéries de la ville, qui affichent davantage de tweets incluant des observations de l'environnement direct, des signalements d'inondations ou des actions (cf. figure 6.53 et 6.56 : "*the visit of hurricane harvey in my area roads all flooded, trees uprooted sitting on roads*", "*creek just went over its banks...and so it starts...please god, let the rain stop! @ cypress, texas*", "*nursing home residents rescued after waiting in several feet of water*"). Le soir, on constate le comportement inverse : les messages de prières apparaissent dans les périphéries de la ville alors qu'un témoin d'inondation "*I'm flooded*" se manifeste dans le centre. Faut-il alors considérer que les lieux enregistrant des concentrations de tweets à tonalité empathique constituent des marqueurs de moindre intensité ou de moindre vulnérabilité des populations concernées ? Pour autant, on pourra noter que c'est le centre de la métropole qui inclut le lieu d'émission des tweets destinés à assurer l'entraide et à recruter des volontaires.

Détecte-t-on des évolutions sémantiques au cours de la journée et dans des territoires identiques ? On peut en effet distinguer une catégorie de tweets qui était active le matin et dont l'importance s'efface dans la journée : il s'agit de la catégorie représentée en vert (observations locales de l'environnement direct). En fait, dans l'ensemble des territoires affectés, on constate un glissement thématique progressif dans la journée : l'ensemble des thèmes, dont les observations locales, sont ancrés dans les discours matinaux. Ce thème vert des observations de l'environenment local disparaît entre 15h et 15h30 puis réapparaît à partir de 15h30. Le soir, il se révèle marginal et cède la place aux sentiments/émotions ainsi qu'aux actions individuelles ou collectives (cf. figure 6.56 : "*if you need help, want to volunteer or have a donation to be made for relief from harvey please* [tweet tronqué]"). Si nous nous interrogeons également sur les thèmes minoritaires, il s'avère que la mention de dégâts occasionnés (en dehors des routes fermées pour inondations) n'apparaît que six fois. De la même manière, même si la thématique des secours se trouve inscrite dans le réseau à partir de 15h30, on n'identifie qu'une seule personne demandant directement du secours (et pour elle) par l'émission d'un tweet géolocalisé.

Enfin, observe-t-on une variabilité spatiale et temporelle locale des émissions de tweets ? Elle n'est pas flagrante : si des tweets s'activent à plusieurs kilomètres de distance d'un quart d'heure à l'autre (ou d'une heure à l'autre), les territoires actifs, à plus petite échelle, restent identiques. La figure 6.57, en page suivante, soumet ainsi une cartographie double des mailles de 10 km en fonction des effectifs de tweets recensés pour la journée du 27 août 2017, et de ces mêmes mailles en fonction que ledit effectif soit inférieur ou supérieur ou égal à la valeur médiane (qui est de 9 tweets par maille).

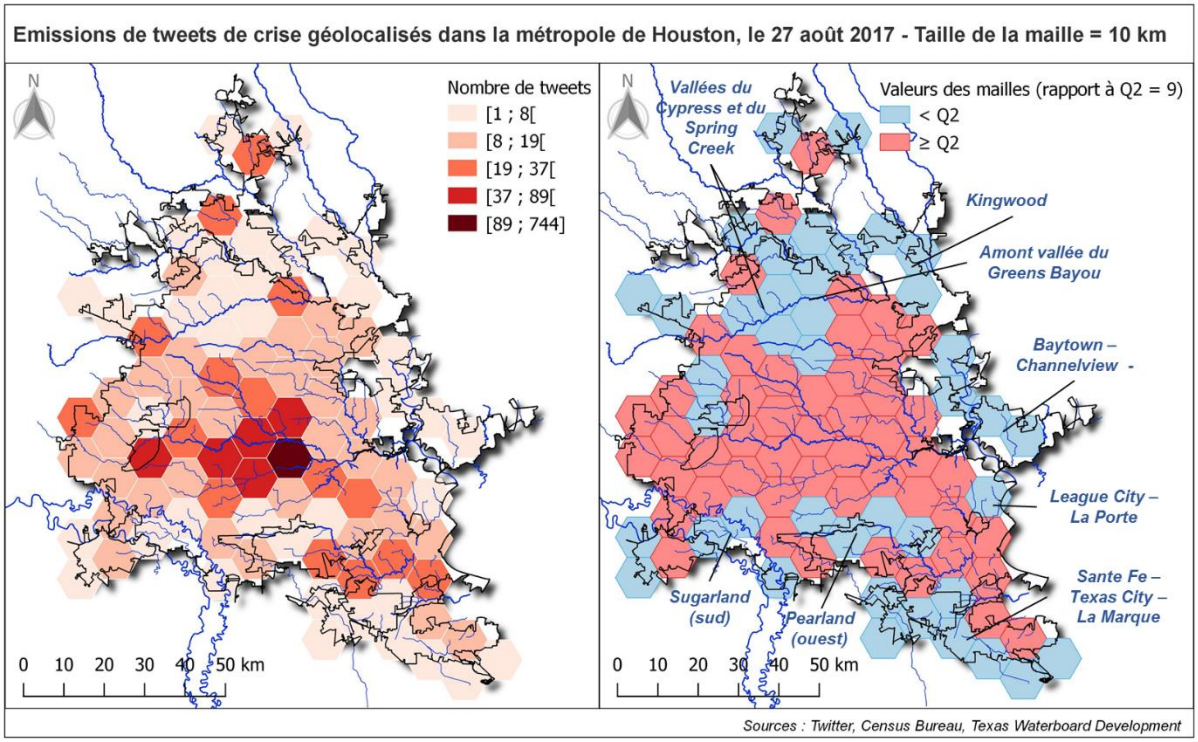


Figure 6.57 : Les territoires de la métropole de Houston en fonction de leur inscription dans l'événement virtuel du 27 août 2017 (C.Cavalière)

Le haut-lieu de l'activité virtuelle du 27 août 2017 reste donc le *CBD* ainsi que les cellules limitrophes au nord et à l'ouest ; par ailleurs, si l'on regarde la carte des effectifs de tweets, il s'avère qu'un certain nombre de foyers d'activité (dont le grand foyer du centre, le foyer de l'extrême ouest, le sud-est ainsi que les trois mailles les plus actives dans le nord de la métropole) correspondent aux structures de l'activité normale qu'on avait explorées, en début de chapitre, par les tweets bruts émis en avril 2016 (cf. figure 6.5). Aux côtés de ces territoires à l'empreinte virtuelle marquée, d'autres lieux restent temporairement silencieux voire quasiment invisibles, et ont tendance à se concentrer dans les territoires périphériques nord (vallées du Cypress Creek et du Spring Creek), est (Kingwood, Baytown, Channelview, League City) et sud (La Marque, Santa Fe, Texas City, Pearland et Sugarland) de la métropole.

De la même manière, les rythmes d'émissions de tweets s'avèrent temporellement discontinus (hormis dans le centre de la métropole). On peut citer l'exemple du quartier de Sugarland-Rosenberg au sud-ouest, qui semble réactif à la survenue d'un phénomène local ou à la survenue d'un événement social, mais dont l'activité virtuelle se dissipe rapidement : l'exemple ci-dessous (figure 6.58) reprend les cartes représentées dans la figure 6.54 : une première phase de l'événement virtuel local est enregistrée de 9h à 9h15 pour annoncer de forts cumuls de pluies. L'événement virtuel se dissipe de 9h15 à 9h30, puis réapparaît dans une deuxième phase pour annoncer des inondations "*flood areas*" et partager des photographies (envoyées avec le hashtag *#harveywatch2017*) ; d'où une nouvelle fois cette question : faut-il considérer le silence de 9h15 à 9h30 comme une phase critique ?

Figure 6.58 : Réactivité et dissipation temporelle rapide des phases de l'événement virtuel dans une périphérie active de la métropole, le quartier Sugarland-Rosenberg, le 27 août 2017

Pour terminer, que peut-on conclure de ce test ? Pendant une journée de crise intense sur la métropole, l'événement virtuel s'avère quantitativement suffisant pour analyser les traces à une résolution temporelle fine (et contrairement à ce qu'on avait constaté les 25 et 26 août 2017, il n'y a ici pas de tweet de crise au contenu personnel ambivalent). Mais dans l'événement virtuel lié à un phénomène extrême rare, la dimension quantitative du tweet reste un leurre : une concentration importante de tweets en un lieu donné n'est pas le gage d'apports sémantiques pertinents pour identifier les phénomènes réels qui affectent le lieu en temps réel. Dans les faits, on trouve davantage de témoins des phénomènes réels dans les territoires éloignés du centre de la métropole. Dans une journée de crise intense, le signal émis par un unique tweet en un lieu donné peut s'avérer plus porteur de sens qu'une concentration de tweets en un autre lieu à l'activité habituelle ; pour autant, le problème reste que ces signaux se montrent irréguliers. En conséquence, s'il paraît difficile d'identifier des changements et phases précises dans le centre de la métropole, qui tweete en continu, l'événement virtuel des marges semble animé par une réactivité directe au réel, qui se traduit par des rythmes d'émission discontinus ainsi qu'une variabilité de la sémantique.

Structures de l'activité après le passage de l'ouragan.

Nous choisissons d'étudier l'après ouragan de la métropole de Houston pour la journée du 30 août 2017[13] : si l'on conserve comme même point de référence, la station de jaugeage du Buffalo Bayou, celle-ci indique que le niveau de l'eau est en-dessous du *NWS flood stage*, le 29 août 2017 à partir de 22h36 ("*#USGS08074000 - Buffalo Bayou at Houston, TX is below NWS flood stage (28ft)*"). Dans un premier temps, nous vérifions l'adéquation de la résolution temporelle définie précédemment, le quart d'heure, avec une période post-phénomène hydrométéorologique (figure 6.59) :

[13] Même si nous avons bien conscience du fait que si nous cherchions à mettre en évidence les dynamiques virtuelles d'organisation et de résilience en situation post-crise, il faudrait suivre l'évolution de l'événement virtuel de cette phase pendant plusieurs semaines, voire plusieurs mois après le passage de l'ouragan.

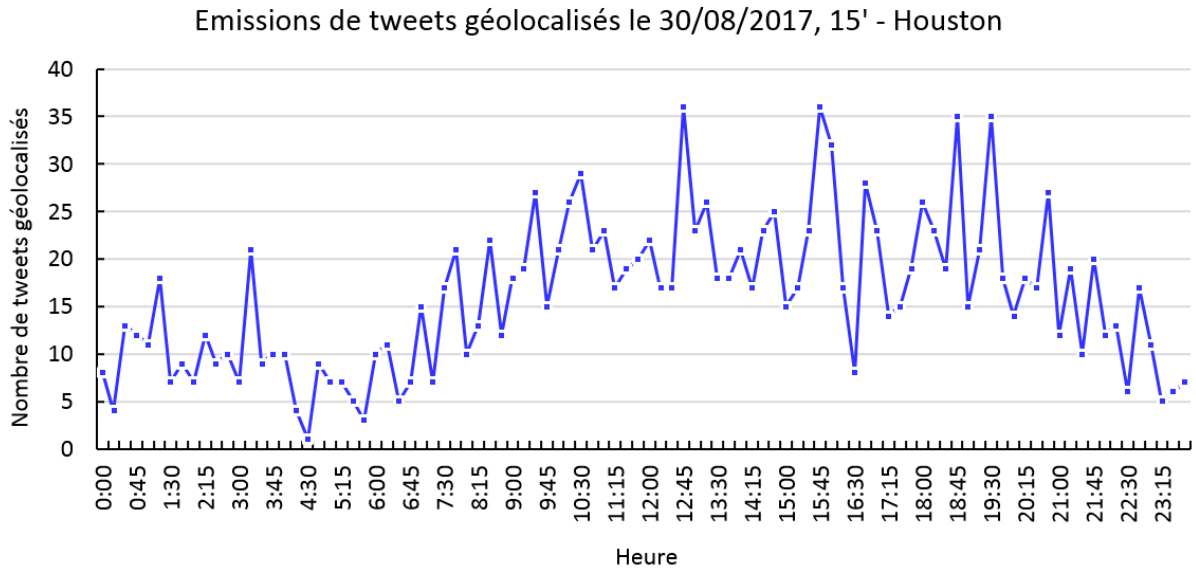


Figure 6.59 : Emissions de tweets de crise géolocalisés le 30 août 2017, Houston – Résolution temporelle de 15 minutes

En dépit du fait que cette journée enregistre un cumul de tweets plus faible que le 27 août (1 532 tweets le 30 août et, pour rappel, 2 011 tweets le 27 août), cette résolution reste significative en période post-phénomène ; on peut alors distinguer un comportement temporel précis qui s'avère finalement général dans l'événement virtuel lié au phénomène rare, quelle que soit sa phase. On peut lui attribuer deux caractéristiques pour Houston :

- une tendance générale marquée par la hausse de l'activité tweeting dès que les individus s'activent le matin, puis sa stabilisation et enfin sa décroissance progressive dans la soirée ;
- les pics sont perceptibles sur des créneaux réguliers : entre 3h et 4h du matin, à la mi-journée (12h30-13h30), au milieu de l'après-midi (15h30-16h30) ainsi qu'en début de soirée (18h30-19h30).

Au final, même si nous avons pu observer une réactivité directe face à l'élévation des eaux, la distribution temporelle des tweets de crise géolocalisés, considérée dans sa globalité, semble suivre une logique non perturbée. Les individus tweetant pendant le phénomène ne changeraient ainsi pas leurs habitudes temporelles de participation à la création/diffusion de contenus géolocalisés[14].

A partir de la figure 6.59, nous avons relevé trois périodes temporelles d'analyse et appliquons une représentation cartographique identique à celle qui a été adoptée pour les cartes du 27 août 2017, des figures 6.53 à 6.56. Ces périodes sont les suivantes :

[14] On peut en effet envisager cette hypothèse, étant donné que nous avons vu qu'une majorité des utilisateurs actifs pendant et après le phénomène sont localisés dans le centre, et à la lecture de leurs discours, ne semblent pas être les plus affectés par les intempéries et leurs conséquences.

- 3h00-3h15-3h30 : pic matinal
- 12h30-12h45-13h00 : pic de la mi-journée
- 18h30-18h45-19h00 : pic de début de soirée

Comme précédemment, les nuages de mots associés aux mailles sont représentés par différentes couleurs, en fonction du thème de rattachement associé au mot ou à l'association lexicale :

- sentiments/prières/remerciements : violet
- actions d'entraide ou de secours : rouge
- collecte de dons ou recrutement de volontaires : orange
- état des lieux des inondations : bleu
- état des personnes ou des biens : noir
- observations environnementales locales : vert
- signalements de pillages ou d'arnaques : rouge brique

Le 30 août est également le jour où l'ouragan, alors rétrogradé en tempête tropicale, touche terre une seconde fois à l'est de Houston (et du Texas), sur la ville de Port Arthur. Un certain nombre de tweets émis depuis Houston concernent cette ville : le cas échéant, la sémantique figurant sur la carte est colorée en fonction des thèmes ci-dessus mais elle apparaît en italique. Les figures 6.60 à 6.62 présentent les cartographies résultantes (NB : pour distinguer le lexique du centre de la métropole [*CBD* et territoires compris dans un rayon de moins de dix kilomètres du *CBD*], la police adoptée reste la suivante : Bradley hand) ; elles sont commentées dans les pages qui les suivent.

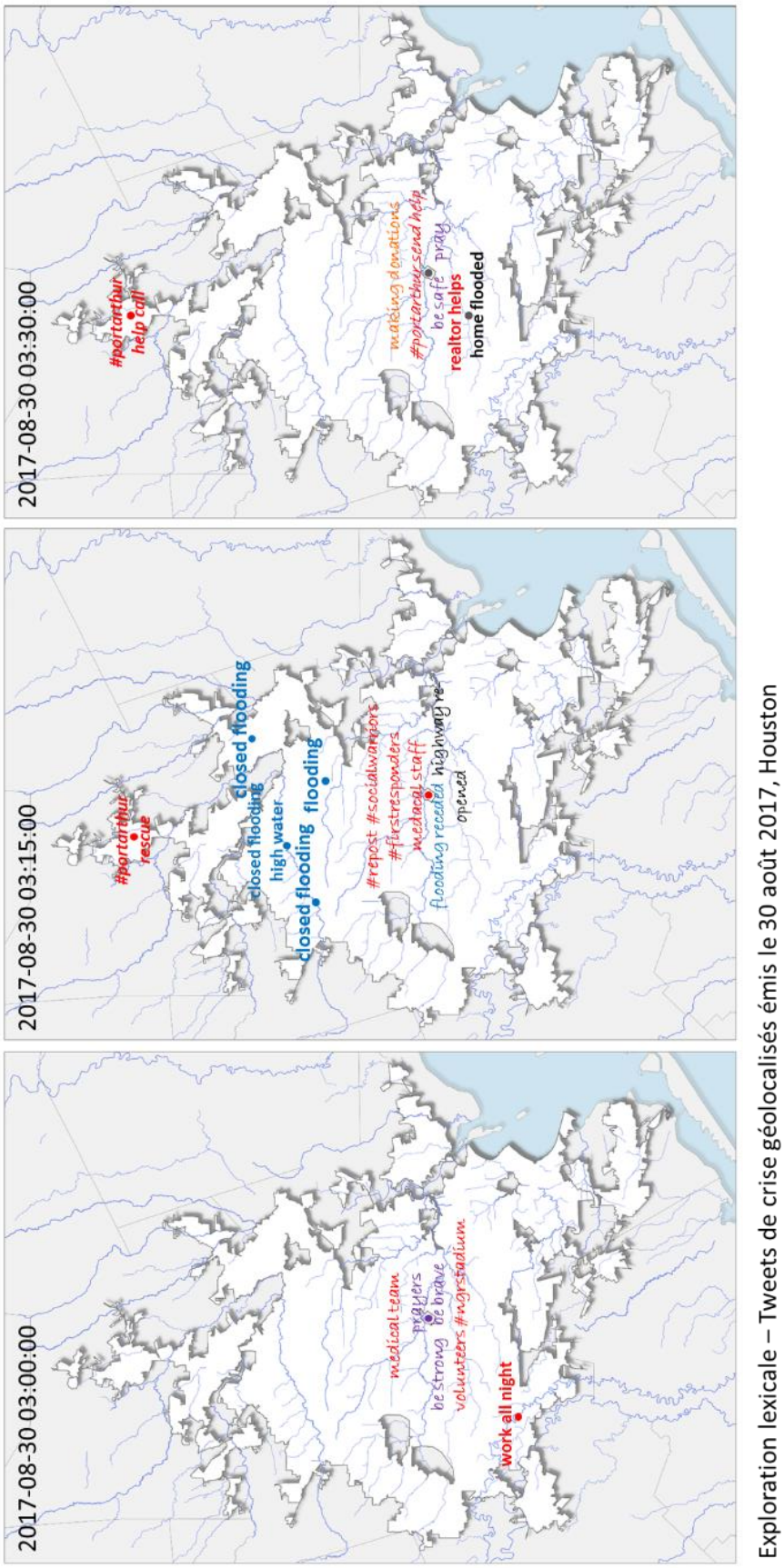


Figure 6.60 : Sémantique des tweets de crise géolocalisés émis entre 3h et 3h45 le 30 août 2017 - Résolution temporelle de 15' (C.Cavalière)

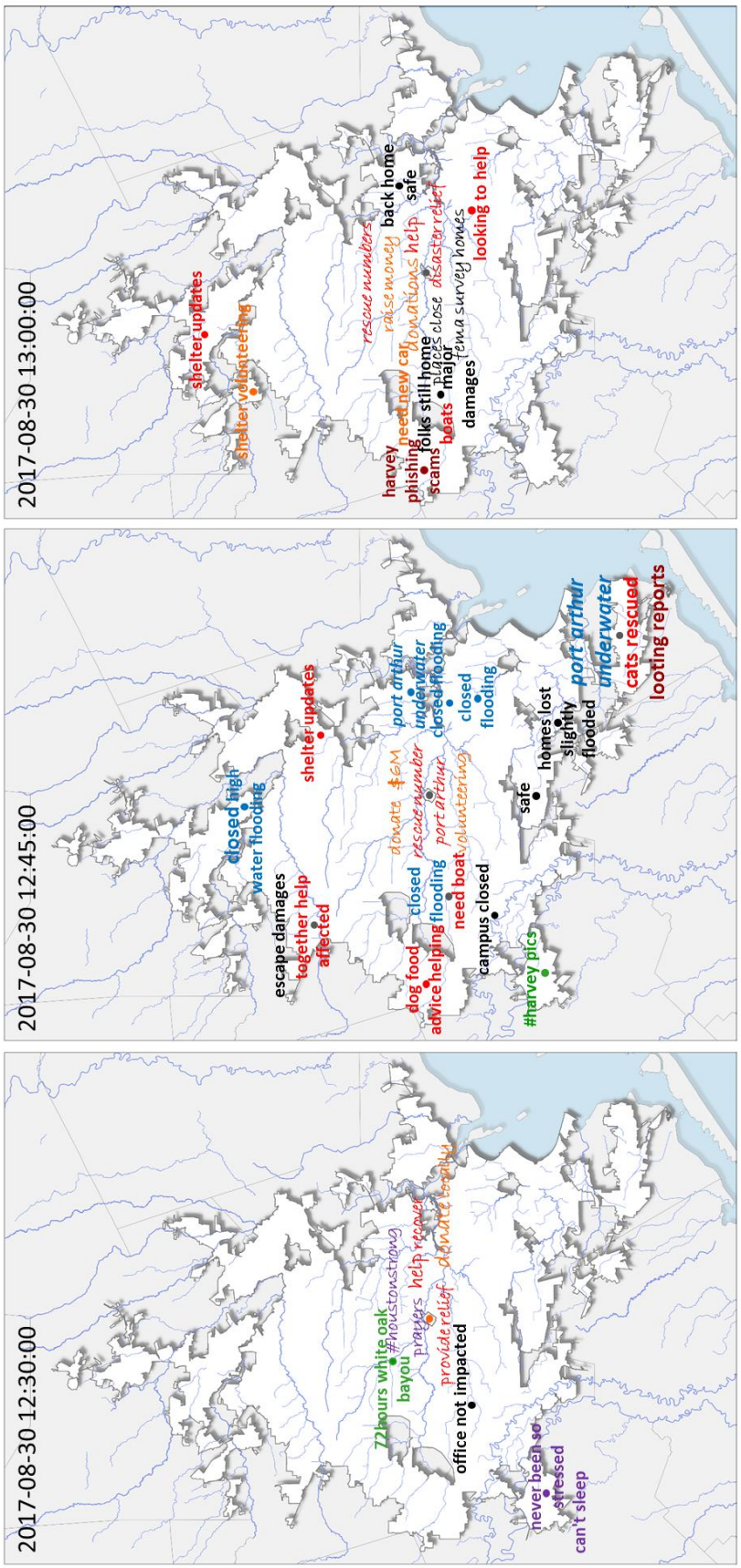


Figure 6.61 : Sémantique des tweets de crise géolocalisés émis entre 12h30 et 13h15 le 30 août 2017 - Résolution temporelle de 15' (C.Cavalière)

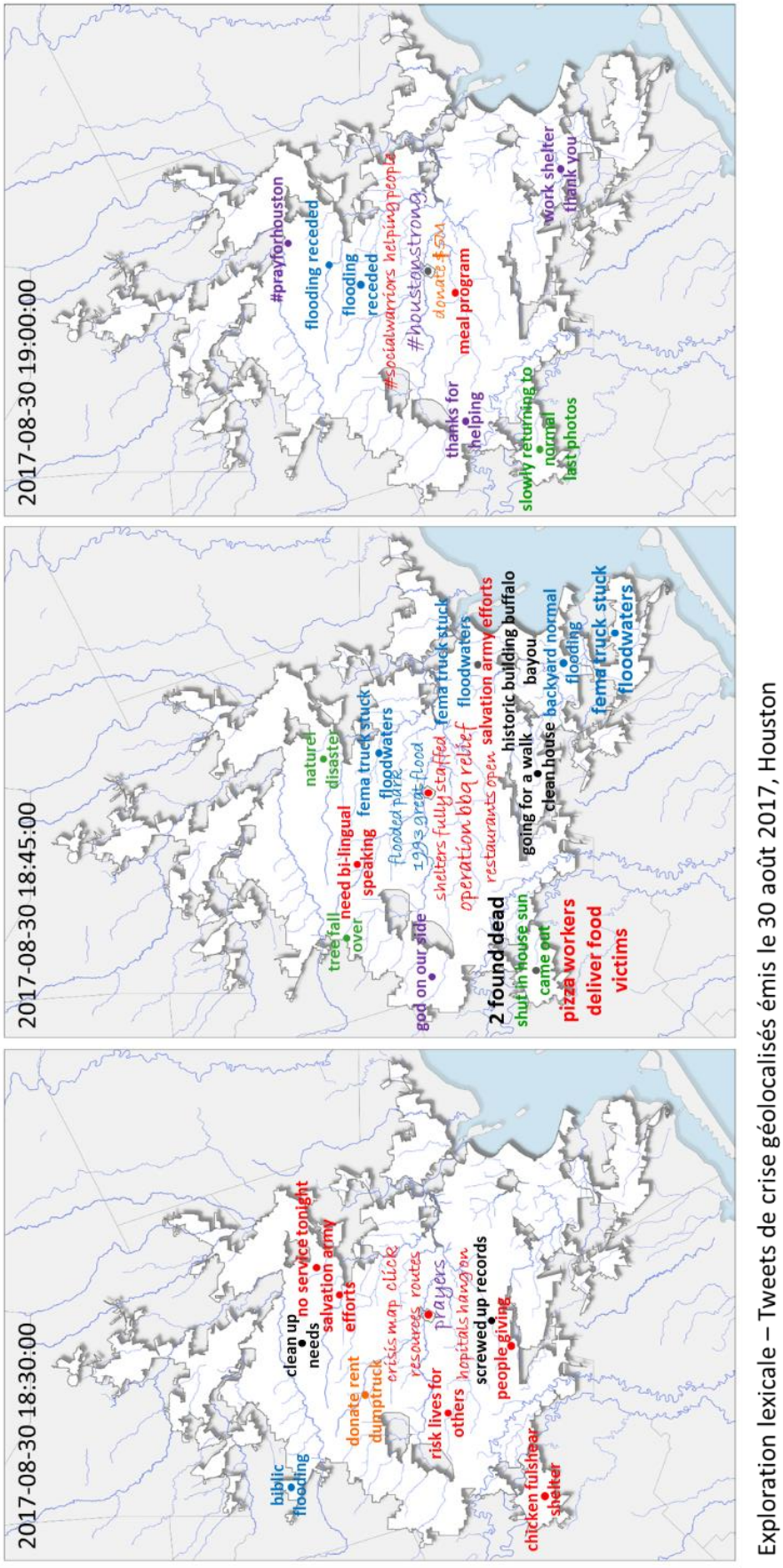


Figure 6.62 : Sémantique des tweets de crise géolocalisés émis entre 18h30 et 19h15 le 30 août 2017 - Résolution temporelle de 15' (C.Cavalière)

Le premier constat qu'on peut établir à partir de ces cartes reste que les territoires actifs en période post-phénomène correspondent aux territoires qui s'illustraient déjà par leur présence virtuelle avant et pendant la période de crise : le centre de la métropole, les banlieues de l'extrême-ouest (Katy, Cypress) et du sud-ouest (Sugarland). L'extrême nord de la métropole (Conroe) se montre actif dans l'organisation de l'aide pour Port Arthur dans les premières heures de la matinée (cf. figure 6.60) puis disparaît définitivement (pour rappel, dans la journée du 27 août, on trouvait déjà, à Conroe, des messages diffusant des informations relatives à l'aide et également des messages de soutien). En revanche, on distingue une activité plus marquée dans l'est de la métropole en période post-phénomène, notamment de 12h45 à 13h (cf. figure 6.61), puis de 18h45 à 19h (cf. figure 6.62). On pourra identifier, comme causes de l'agitation virtuelle, trois séries d'une poignée de retweets concernant : les inondations de Port Arthur ("*port arthur underwater*"), les fermetures d'axes routiers inondés ("*closed flooding*") ainsi que le signalement de pilleurs ("*looting reports*") dans l'extrême sud-est (figure 6.61). Le soir, on retrouve ce comportement, au travers du message concernant la FEMA : "*photo shows fema truck stuck in harvey's floodwaters*", auquel s'ajoute néanmoins un marqueur local relatif à l'intensité de l'inondation : "*it's still in my neighbors backyard, but that's normal flooding*" (figure 6.62). Le second fait qui n'a pas évolué entre la période où Houston est frappée par l'ouragan et la période post-phénomène, est le suivant : quelle que soit la temporalité considérée, le centre reste le lieu d'émission des tweets géolocalisés concernant la problématique des dons et de l'organisation générale de l'aide, notamment par les annonces de collecte et de recrutement de volontaires : "*volunteers !!* □ *#nrgstadium #hurricaneharvey #houston @ houston, texas*" (cf. figure 6.60, entre 3h et 3h15), "*here's rescue numbers for port author and beaumont #texasstrong* 🙏🏻 *@ third ward, houston*" (cf. figure 6.61, entre 12h45 et 13h), "*share this number so we can get all the shelters fully staffed asap #houstonstrong @ houston, texas*" (cf. figure 6.62, entre 18h45 et 19h). Ce thème ne reste néanmoins pas l'apanage du centre et se diffuse dans les périphéries de la métropole (cf. figure 6.62, entre 18h45 et 19h : "*needs bi-lingual speaking peeps! please help if you're able* 🙏🏻🙏🏻*#houstonstrong*[15]").

On pourra également distinguer quelques évolutions dans le contenu global de l'événement virtuel du 30 août 2017 :

- il mentionne des phénomènes en téléprésence : les 25 et 26 août, le réseau de Houston restait quasi muet face au phénomène en cours à Rockport (pour rappel, première ville littorale directement affectée par l'ouragan alors de catégorie 4) ; le 30 août, alors que la tempête touche terre une seconde fois, l'aide aux sinistrés de Port Arthur s'organise également par le réseau virtuel de Houston (cf. figures 6.60 et 6.61 : "*if you're in port arthur and you need help call this number asap! #hurricaneharvey2017*", "*please share this with port arthur residents waiting to be evacuated*").

[15] Faut-il considérer ce tweet comme un indice de la présence locale d'une communauté vulnérable et particulièrement affectée ?

- les problématiques (ou centres d'intérêt) véhiculées par les tweets : la dimension émotionnelle est moins inscrite dans la sémantique des tweets émis le 30 août dans le centre de la métropole (contrairement au 27 août où le hashtag *#pray* et ses variantes se déclinaient dans l'ensemble des temporalités explorées) au profit de l'action : on passe d'un événement virtuel où des utilisateurs tweetent leur empathie vis-à-vis des autres (laissant supposer qu'ils sont moins directement affectés par le phénomène), à un événement virtuel où ces utilisateurs restent actifs mais manifestent leur intérêt à s'engager pour les autres ("*we will be mailing suits, shirts, ties, male accessories down to #houston to help those in* [tweet tronqué]*", "we are here to help. if your home was damaged by #hurricaneharvey please dm me so we can come"*).

Quelles informations font défaut dans les tweets de crise géolocalisés du 30 août 2017 ? On pourra citer trois éléments :

(1) Les dégâts et dommages : finalement, on identifie peu de contenus faisant état des dommages aux habitations, infrastructures ou aux personnes. Côté humain, le réseau mentionne deux décès dans le comté de Fort Bend (sud-ouest de la métropole), on peut encore distinguer un individu annonçant qu'il est sauf ainsi qu'un autre indiquant qu'il rentre chez lui sauf (cf. figure 6.61, respectivement entre 12h45 et 13h puis entre 13h et 13h15). Mais en ce qui concerne les dégâts matériels, la plupart des tweets ne sont pas précisément ciblés sur un objet local : à Cypress, dans le nord-ouest, on peut distinguer un seul utilisateur indiquant que sa maison a échappé aux dégâts (cf. figure 6.61, entre 12h45 et 13h) des inondations (mais est-ce un cas isolé ou y a-t-il eu finalement peu de dégâts dans ce quartier ? pour rappel, Cypress fait partie des lieux ayant subi les plus forts cumuls pluviométriques entre le 24 et le 31 août 2017, soit entre 740 et 1 119 mm d'eau [cf. figure 6.3]). Dans le sud-est, on nous dit que certaines maisons sont légèrement inondées (cf. figure 6.61, entre 12h45 et 13h) alors que d'autres sont *perdues* (faut-il entendre complètement détruites ?). Le problème principal reste que ces tweets ne sont pas présents :

- dans l'ensemble des lieux affectés par la crise (nous avons déjà mis en évidence les territoires de la métropole qui étaient peu voire pas inscrits dans l'événement virtuel du 27 août, cf. figure 6.57).
- dans l'ensemble des lieux affectés par la crise et qui sont visibles virtuellement : à l'extrême sud-ouest, dans le quartier Sugarland-Rosenberg, on avait détecté de l'activité virtuelle le 27 août ; celle-ci perdure le 30 août, mais sans mention de dommages aux habitations ou infrastructures (cf. figures 6.60 à 6.62). De la même manière, on avait également enregistré de l'activité sur Cypress au nord-ouest les 27 et 30 août, dont l'utilisateur indiquant que son domicile n'avait pas été touché. A l'inverse, on a enregistré de l'activité dans le quartier de Santa Fe – Texas City – La Marque le 30 août (alors que ce territoire était quasi invisible le 27 août) et bien qu'il fasse également partie des territoires ayant subi les plus forts cumuls de pluie, cette activité virtuelle du 30 août n'est pas focalisée

sur les dégâts (mais sur Port Arthur, le signalement de pillages ou encore de secours aux chats).

Qu'observe-t-on en comparant ces observations avec les données témoins de dégâts de la FEMA, au 31 août 2017 ? La figure 6.63 représente la localisation des dégâts aux habitations, inventoriés par la FEMA à la date mentionnée, ainsi que la localisation des tweets de crise pour trois territoires périphériques de la métropole (Sugarland, Cypress et Santa Fe/Texas City/La Marque). Le tweet n'est pas un capteur systématique des dommages et de leur intensité :

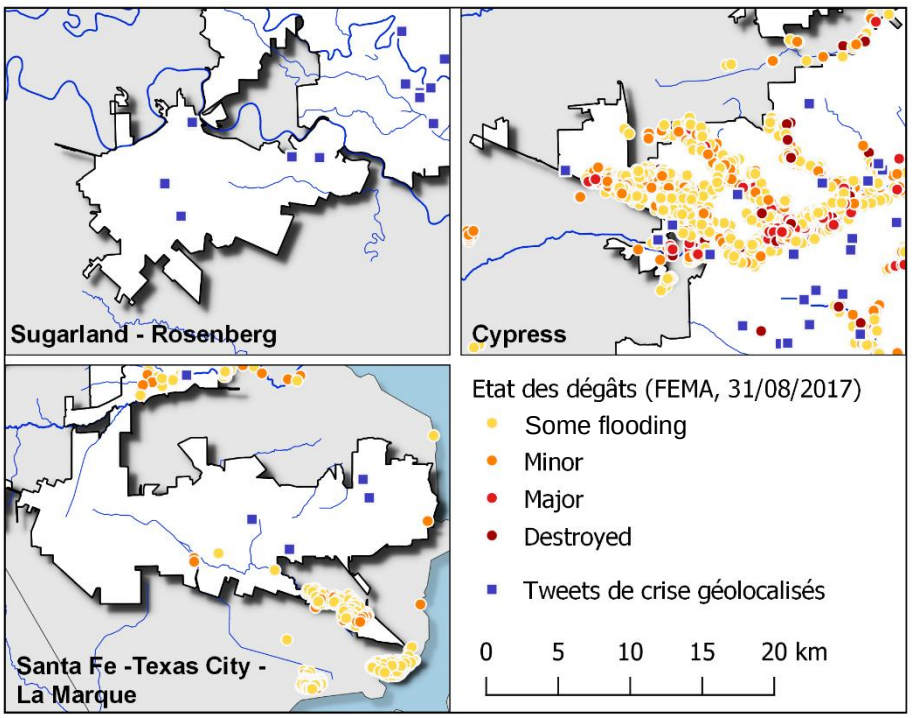


Figure 6.63 : Etat des dégâts recensés par la FEMA au 31/08/2017 dans trois quartiers de la métropole (C.Cavalière)

Dans l'extrême sud-est, observe peu de dégâts majeurs (La Marque) et aucune émission de tweets géolocalisés n'est enregistrée dans la proximité directe des maisons affectées au sud. A Sugarland, aucun dommage n'est encore recensé mais on trouve davantage de tweets géolocalisés (et sans doute les enquêteurs n'étaient pas encore passés : on savait, au 30 août, par le tweet indiquant les difficultés de circulation "*photo shows fema truck stuck in harvey's floodwaters*", que ces enquêteurs se trouvaient plutôt dans l'est de la métropole). Au contraire, le tweet mentionnant le domicile de l'utilisateur épargné correspondrait en effet à un cas isolé à Cypress (quartier dans lequel on trouve des maisons ayant subi des dégâts majeurs et complètement détruites à proximité des bayous, mais dont les tweets géolocalisés ne se répartissent pas dans les lieux affichant les dégâts les plus importants).

La sémantique du tweet géolocalisé apparaît comme insuffisante pour aider à appréhender le niveau des dégâts sans contact direct avec le terrain. En effet, si la sémantique d'un seul tweet suffit à lancer un signal de perturbation en cours, l'évaluation des dégâts requerrait des tweets en quantité, étant donné l'hétérogénéité constatée des dommages dans un périmètre restreint : à Cypress, on peut identifier, espacées par quelques dizaines de mètres, des maisons détruites comme des maisons ayant subi des dégâts mineurs. Le tweet géolocalisé n'apparaîtrait alors pas comme un primo-indicateur de terrain fiable et remet en cause l'un des principes du *crowdsourcing* appliqué aux tweets de crise géolocalisés : si l'on suppose que le contenu généré par la foule a la capacité de créer de l'information plus précise et plus complète que des experts dépêchés sur place après le phénomène, le tweet géolocalisé ne s'ancre, dans les conditions actuelles de sa production, guère dans cette posture.

(2) Dans cet extrait d'analyse, peu de tweets nous permettraient lexicalement d'entrevoir comment les habitants/sinistrés utilisateurs ont appréhendé ou vécu la crise.

(3) Le 30 août, l'ensemble des territoires sur lesquels on détecte de l'activité virtuelle s'avèrent dans des phases différentes : dans le nord, on détecte de nouvelles routes fermées en raison des inondations (cf. figure 6.60) alors que dans le sud, on trouve déjà des messages témoignant de mesures de résilience, par le retour et le nettoyage des maisons (cf. figure 6.62 : "*friends are everything. took hours of work and a number of people to clear my friends house*"). A ce titre, on pourra distinguer un autre tweet intéressant, qui illustre bien le potentiel de la trace (cf. figure 6.62) : "*it's still in my neighbors backyard, but that's* ***normal flooding***." En effet, si l'on veut progresser vers la cartographie du ressenti face à la crise et de son évolution, ce tweet justifie l'importance de la connaissance de terrain que l'individu acquiert dans son environnement familier (ici, le jardin du voisin est inondé, mais à un niveau jugé habituel donc il n'y a pas d'emportement de la part de l'utilisateur et sans doute la phase de décrue est engagée). Malheureusement, parmi les tweets de crise géolocalisés par GPS, les tweets présentant ce profil sont minoritaires.

Bilan relatif à l'apport des dynamiques spatio-temporelles et à la cohérence lexicale des tweets de crise géolocalisés

Force est de constater que la localisation des principaux foyers de l'activité virtuelle de crise, considérée à l'échelle globale de la métropole, ne dépend pas de la localisation, ni de l'intensité d'un phénomène, mais des structures de l'activité normale : lors de l'étude du phénomène récurrent en avril 2016, en nous référant aux tweets marqueurs d'alerte officielle, nous avons vu que les territoires proches de ces marqueurs n'enregistraient que peu voire pas de réponse virtuelle. Dans le cas de l'ouragan, ce comportement spatial est encore plus flagrant : les structures inventoriant les effectifs de tweets de crise géolocalisés les plus volumineux correspondent aux structures de l'activité virtuelle régulière, de ses vides et de ses pleins.

Bilan relatif à l'apport des dynamiques spatio-temporelles et à la cohérence lexicale des tweets de crise géolocalisés (suite)

En revanche, il y a une réactivité temporelle fine et globale aux phénomènes et événements du réel, marquant l'engagement d'une dynamique virtuelle temporelle conjointe à la dynamique temporelle réelle : en effet, si les cumuls de précipitations ne se positionnent pas comme des facteurs explicatifs de la distribution spatiale des tweets de crise géolocalisés (cf. chapitre 5), l'élévation du niveau des eaux exerce une influence sur l'agitation temporelle du réseau. Sur le plan spatial, cette réactivité se manifeste comme suit : une minorité d'îlots (qui comptent néanmoins plus de 50% des émissions) concentrent de fortes densités de tweets de crise géolocalisés (c'est notamment le cas des quartiers du centre de la métropole et de quelques lieux situés en périphérie), et ces îlots contrastent avec l'activité "résiduelle" qui s'avère éparse et diffuse dans l'ensemble de la métropole (comportement spatial qui a été mis en évidence par l'utilisation de l'algorithme DBSCAN).

Quel que soit le type de phénomène considéré, les pics d'agitation maximale enregistrés correspondent aux multiples *retweets* d'un unique message ; nous n'avons en revanche pas approfondi les analyses concernant les modalités de diffusion de ces messages en raison du caractère ambigu de leur contenu ou de leur origine. Le second facteur d'agitation correspond à la survenue de phénomènes locaux qui sont généralement à l'origine d'une poignée de messages types comme les tweets annonçant la fermeture de routes inondées, mais également de messages émanant des individus réagissant face à la pluie ou à l'inondation en cours.

A l'échelle locale, le tweet de crise géolocalisé obéit-il à une logique spatiale ? On peut distinguer une certaine cohérence des tweets peu distants, émis depuis le *CBD* et les territoires proches du centre (soit les tweets de crise qui témoignent des dynamiques d'organisation des communautés ou qui véhiculent des sentiments). On identifie également cette cohérence spatiale dans les tweets à consonance officielle indiquant les fermetures d'axes routiers inondés. Mais dans les autres territoires explorés, la sémantique des tweets émis restant très hétérogène, nous considérons que la création de tweets de crise s'affiche comme une contribution individuelle ponctuelle sans structure générale apparente : ce contenu est en effet très irrégulièrement mis à jour (d'où les alternances entre périodes d'activité et de silence, et l'impossibilité d'assurer le suivi régulier de l'intensité d'un phénomène en un lieu précis car les utilisateurs semblent tweeter à un moment choisi). D'autre part, il est difficile de savoir si l'activité virtuelle d'un lieu est proportionnelle aux dégâts et aux besoins des populations. Si l'on détecte un signal individuel unique indiquant que tout va bien chez l'utilisateur en question, doit-on alors considérer qu'il n'y pas non plus de perturbation majeure chez ses voisins, ou alors que le tweet est motivé par le fait que seule sa propriété ait été épargnée dans sa rue ?

6.3. Identification des lieux d'intérêt dans la métropole : approches statistiques et sémantiques

Dans ce dernier point, et maintenant que nous avons exploré les dynamiques et lieux de réactivité de l'événement virtuel, nous posons, en écho au tout premier chapitre du manuscrit, la question de la valorisation cartographique du matériau tweet géolocalisé, sous deux angles distincts : en fonction des phénomènes, nous avons vu que la distribution spatio-temporelle du tweet de crise géolocalisé ne répond pas, dans tous les cas, ni à la logique spatiale de l'intensité du phénomène, ni au lancement de l'alerte virtuelle. Peut-on alors envisager de détecter des structures de réactivité par les tweets seuls, sans recours à des données officielles de réalité-terrain ? Dans un second temps, peut-on interpréter la sémantique des tweets géolocalisés afin de caractériser la vulnérabilité de populations locales ?

6.3.1. Explorer les tweets de crise géolocalisés par eux-mêmes

6.3.1.1. Détection d'anomalies dans l'activité virtuelle

Dans le chapitre et les paragraphes précédents, nous avons constaté que les tweets de crise géolocalisés ne pouvaient pas être considérés comme des marqueurs quantitatifs d'intensité étant donné que la distribution spatiale des tweets de crise semble influencée par les biais d'utilisation socio-spatiaux de temps normal. En revanche, l'émission d'un tweet suffit à lancer un signal local de perturbation en cours : ce signal prend son sens non pas par la quantité importante de tweets émis en un lieu, mais par sa sémantique.

Peut-on alors rapidement détecter le signal de ces lieux intéressants, localisés en dehors du centre de la métropole qui tweete en temps de crise comme en temps normal, en ayant comme unique recours les tweets (donc sans comparaison préalable avec toute autre donnée) ? Nous effectuons un test pour les deux mêmes journées du 27 et du 30 août 2017 ; ce test consistera à détecter une activité tweeting anormale, en se focalisant sur le rapport entre l'activité de crise et l'activité normale en un lieu précis du territoire :

- le territoire métropolitain est découpé en mailles de dix kilomètres de côté (pour rappel, le réseau formé par la distribution spatiale des tweets géolocalisés est dense dans le centre de la métropole mais lâche dès qu'on observe les quartiers périphériques).
- on collecte l'ensemble des tweets géolocalisés émis entre avril et juin 2017 dans l'aire métropolitaine de Houston (on ne collecte que les tweets du printemps 2017 car l'activité normale peut être perturbée pendant la saison estivale, et très certainement perturbée par la survenue de l'ouragan en août).
- dans chaque maille, afin de gommer l'effet hors-normes des lieux qui concentrent l'activité virtuelle normale, on calcule le rapport entre le nombre de tweets de crise

géolocalisés émis pour une journée perturbée et le nombre moyen de tweets bruts émis par jour sur les trois mois de collecte.

La figure 6.64 indique le résultat pour les deux journées étudiées prédécemment, le 27 août 2017 (jour le plus violent) et à J+3, le 30 août. Les mailles inactives en temps normal et n'enregistrant que des tweets de crise sont représentées en noir ; les mailles dont l'activité quotidienne de crise est supérieure à l'activité moyenne quotidienne figurent en rouge brique (ratio > 1) ; les mailles qui présentent une activité de crise équivalente à l'activité moyenne quotidienne sont représentées en rouge (ratio = 1) ; les mailles dont l'activité virtuelle de crise est inférieure à l'activité normale quotidienne apparaissent en orange (0 < ratio < 1) ; pour finir, les mailles qui n'enregistrent aucune activité de crise sont en jaune (ratio = 0).

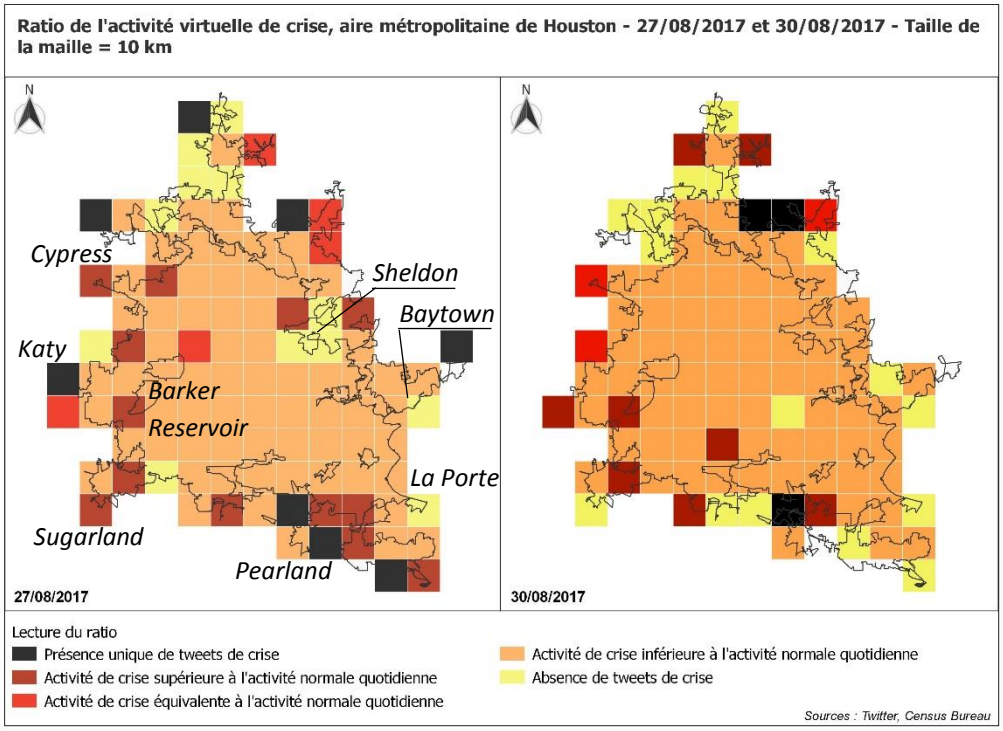


Figure 6.64 : Ratio entre l'activité tweeting de crise quotidienne et l'activité moyenne quotidienne, aire métropolitaine de Houston (C.Cavalière)

Sur les deux journées représentées, 68% des mailles présentent une activité de crise moins importante que l'activité normale ou quasi nulle ; ces mailles s'étendent du centre de la métropole vers ses marges. En revanche, la minorité de mailles qui enregistrent une activité de crise équivalente ou supérieure à la normale, ou uniquement détectée en période de crise, se localisent dans les quartiers périphériques de la métropole. On retrouve notamment les marges dans lesquelles on avait identifié une sémantique particulière, et plus porteuse de sens vis-à-vis du phénomène que celle du centre, ainsi que des périodes alternant activité et silence : Cypress, Katy, Sugarland (du nord-ouest au sud-ouest), Barker Reservoir à l'ouest, Pearland au sud-est. Il reste néanmoins des lieux ayant subi des dégâts qui ne transparaissent pas, quelle que soit la méthode employée, dans l'événement virtuel : ici, on pourra noter la

vallée du Cypress Creek (qui traverse le nord de la métropole d'ouest en est depuis Cypress) et du Spring Creek (plus au nord), Sheldon et Baytown à l'est, ainsi que La Porte, qui borde la baie de Galveston au sud-est. La figure 6.65 représente ainsi les deux cas de Baytown et des vallées du Cypress Creek et du Spring Creek, en comparant l'activité des mailles au 27 août 2017 ainsi que les dégâts aux habitations, inventoriés au 31 août 2017 : la concentration des dégâts laisse supposer des crises locales intenses pendant la survenue du phénomène au 27 août mais l'empreinte spatiale de l'événement virtuel résultant reste inférieure, voire nulle, par rapport à l'activité virtuelle moyenne quotidienne enregistrée en période normale.

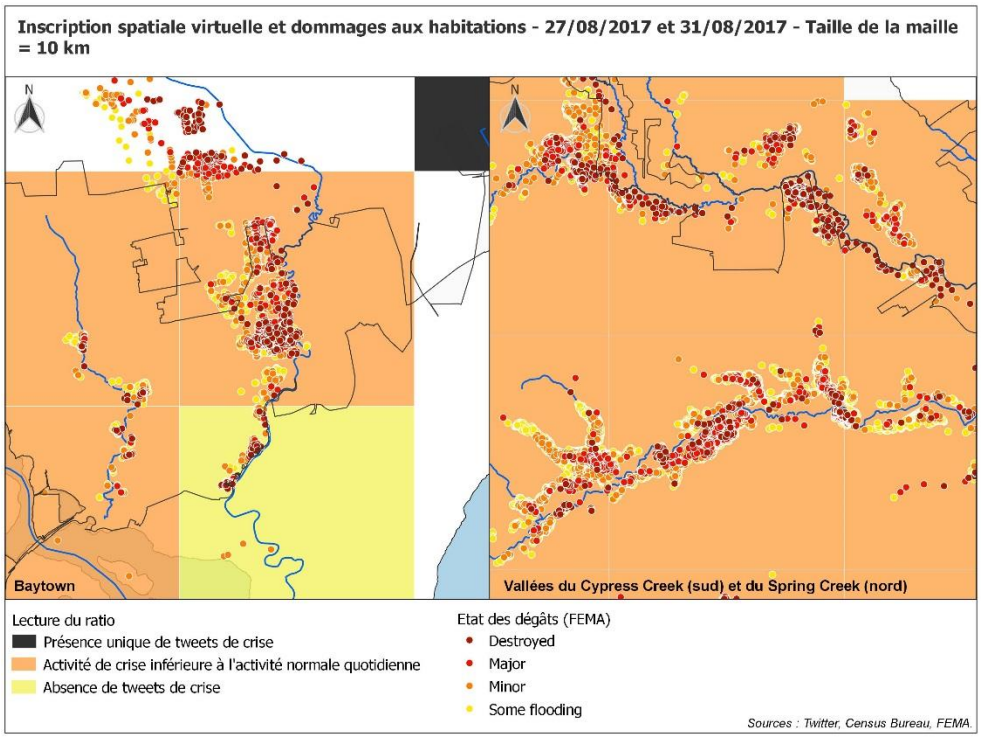


Figure 6.65 : Empreinte spatiale virtuelle des tweets de crise et dommages aux habitations dans les confins de la métropole, 27/08/2017 et 31/08/2017 (C.Cavalière)

Peut-on alors observer l'existence d'un lien entre le niveau d'activité de la maille au 27 août et l'intensité des dégâts inventoriés au 31 août (en supposant que cette variable soit un indicateur de l'intensité locale de la crise) ? Un test de Khi² est effectué entre les modalités des deux variables qualitatives *états des dégâts* et *degré d'activité de la maille en crise* (ici, les modalités correspondent à la légende des figures 6.64 et 6.65 : absence d'activité de crise, présence unique d'activité de crise, activité de crise équivalente à la normale, activité de crise supérieure à la normale, activité de crise inférieure à la normale). Le résultat observé rejette le modèle d'indépendance entre les deux variables (khi² observé de 1 636 pour une valeur théorique de 26,22 associée à un risque de 1% et 12 degrés de liberté[16]). En revanche, le

[16] A considérer avec un certain recul étant donné l'effectif total important du tableau (109 684 habitations inventoriées).

graphique des écarts entre les valeurs observées et le modèle d'indépendance ne met pas en évidence des comportements d'activité virtuelle opposés en fonction de l'intensité des dégâts (figure 6.66) :

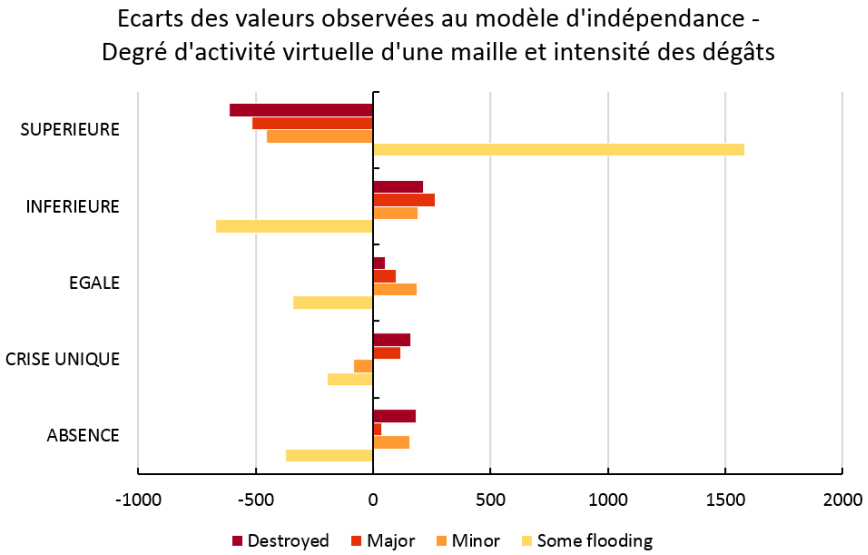


Figure 6.66 : Ecarts au modèle d'indépendance - Intensité des dégâts aux habitations et degré d'activité de la maille

Les quelques mailles enregistrant une activité supérieure à la normale (cf. figure 6.64) témoignent d'une sur-représentation d'inondations mineures (*some flooding*) alors que l'ensemble des modalités indiquant des dégâts plus importants sont sous-représentées ; pour cette catégorie, le degré d'activité de crise n'est probablement pas un indicateur fiable. En revanche, l'ensemble des mailles dans lesquelles on enregistre soit un niveau d'activité de crise inférieure ou égale à la normale, soit un silence total ou encore une activité enregistrée uniquement pendant la crise, présentent une certaine homogénéité dans l'intensité des dégâts : sous représentation des inondations mineures et sur-représentation des dégâts mineurs (à l'exception des mailles actives uniquement pendant la crise), majeurs et des habitations complètement détruites.

Une nouvelle fois, on ne peut pas standardiser un comportement virtuel : certaines mailles témoignent en effet d'un comportement particulier lorsqu'elles sont frappées (silence, activation unique) mais l'indice statistique s'avère incomplet et spatialement discriminant : on retrouve une sur-représentation des dommages les plus violents (*major*, *destroyed*) dans les mailles où l'activité de crise se montre inférieure à la normale. Or, dans les lieux en question, on trouve aussi bien les quartiers du centre, moins affectés que les territoires mentionnés dans la figure 6.65 : ainsi, sur une maille de 100 km² et si l'on considère exclusivement les dégâts les plus importants (*major* et *destroyed*), on trouve, dans le centre de la métropole (*CBD* et quartiers limitrophes), un total de 357 habitations affectées ; à Baytown et dans la vallée du Cypress Creek, on en trouve respectivement 854 et 840.

6.3.1.2. Exploration des poches d'activité virtuelle exceptionnelle en situation de crise

Si les lieux d'activité exceptionnelle en situation de crise détectés par cette méthode ont tendance à correspondre aux lieux dans lesquels on a constaté l'émergence d'une information de crise pertinente à un pas de temps fin, le quart d'heure, peut-on ici retrouver cette pertinence lexicale en analysant des tweets géolocalisés à très grande échelle spatiale mais sur une journée entière ? Dans un premier temps, on explore les mailles du 27 août 2017 présentant une activité exceptionnelle (figure 6.67) :

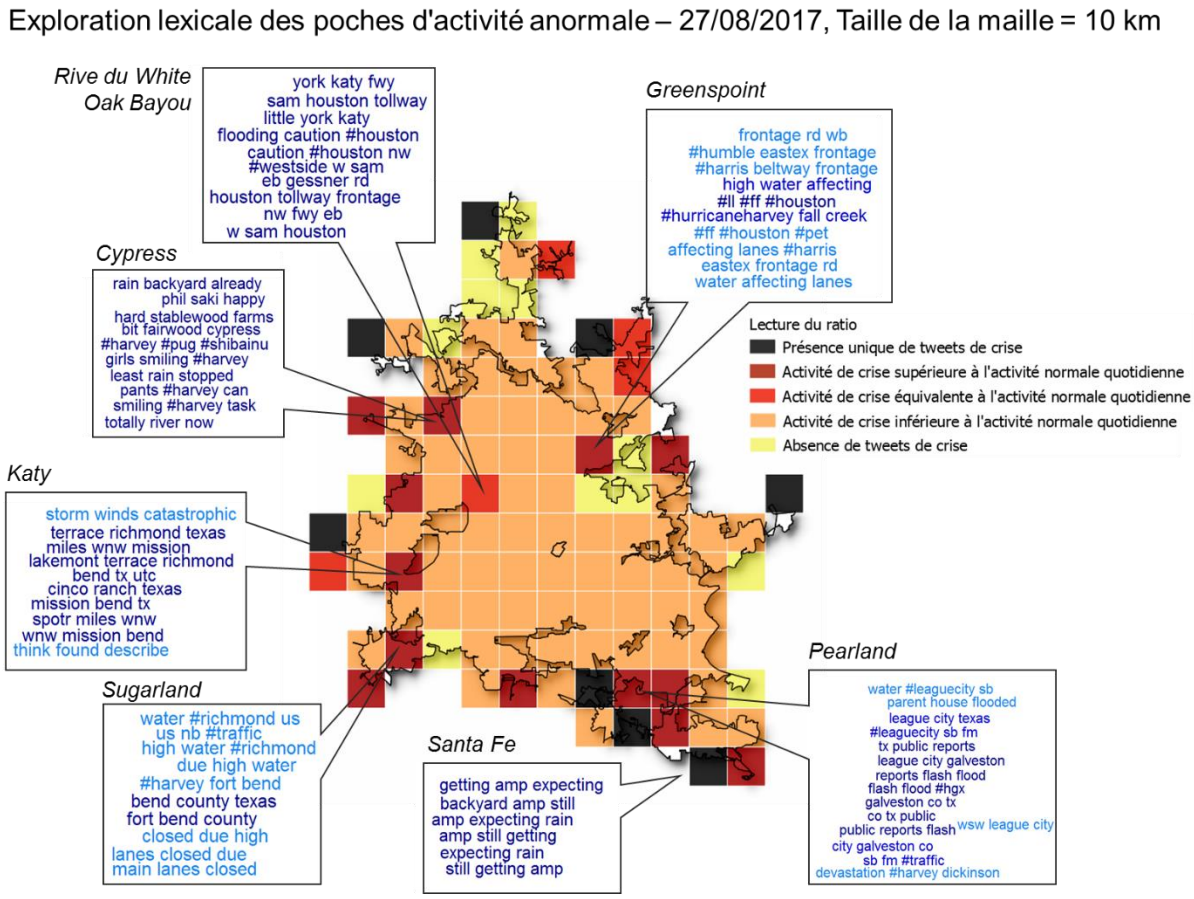


Figure 6.67 : Exploration de la sémantique des poches d'activité exceptionnelle de crise, 27/08/2017 (C.Cavalière)

En fait, on peut établir le même constat que précédemment, lorsqu'on avait étudié la sémantique des journées du 25 et du 26 août 2017 (cf. paragraphe 6.2.2.3) : l'analyse sémantique menée à l'échelle de la journée ne met pas en évidence l'existence d'informations personnelles centrées sur l'individu dans son environnement. Ce résultat provient très vraisemblablement du fait qu'on retrouve à plusieurs reprises, et quel que soit le lieu de la métropole, ces messages à caractère officiel qui apparaissent plusieurs fois et supplantent la minorité de messages personnels dans l'affichage des nuages de mots. Ainsi, comme on peut le constater à plusieurs reprises le 27 août (cf. figure 6.67), les associations lexicales les plus

récurrentes et par conséquent visibles dans les nuages de mots concernent les perturbations de voiries inondées : "*flooding. in #westside on w sam houston tollway frontage rd sb between little york and the i-10 katy fwy #traffic*" (maille représentant une rive du White Oak Bayou), "*high water affecting all lanes in #harris on beltway 8 n frontage rd wb between hwy 59 and hardy toll rd #traffic*" (maille de Greenspoint).

Dans un second temps, on souhaite savoir comment l'activité virtuelle exceptionnelle du 27 août 2017 a évolué dans ces mêmes mailles, le 30 août. La figure 6.68 représente donc les résultats de l'analyse sémantique effectuée dans les mailles identiques à la figure 6.67 :

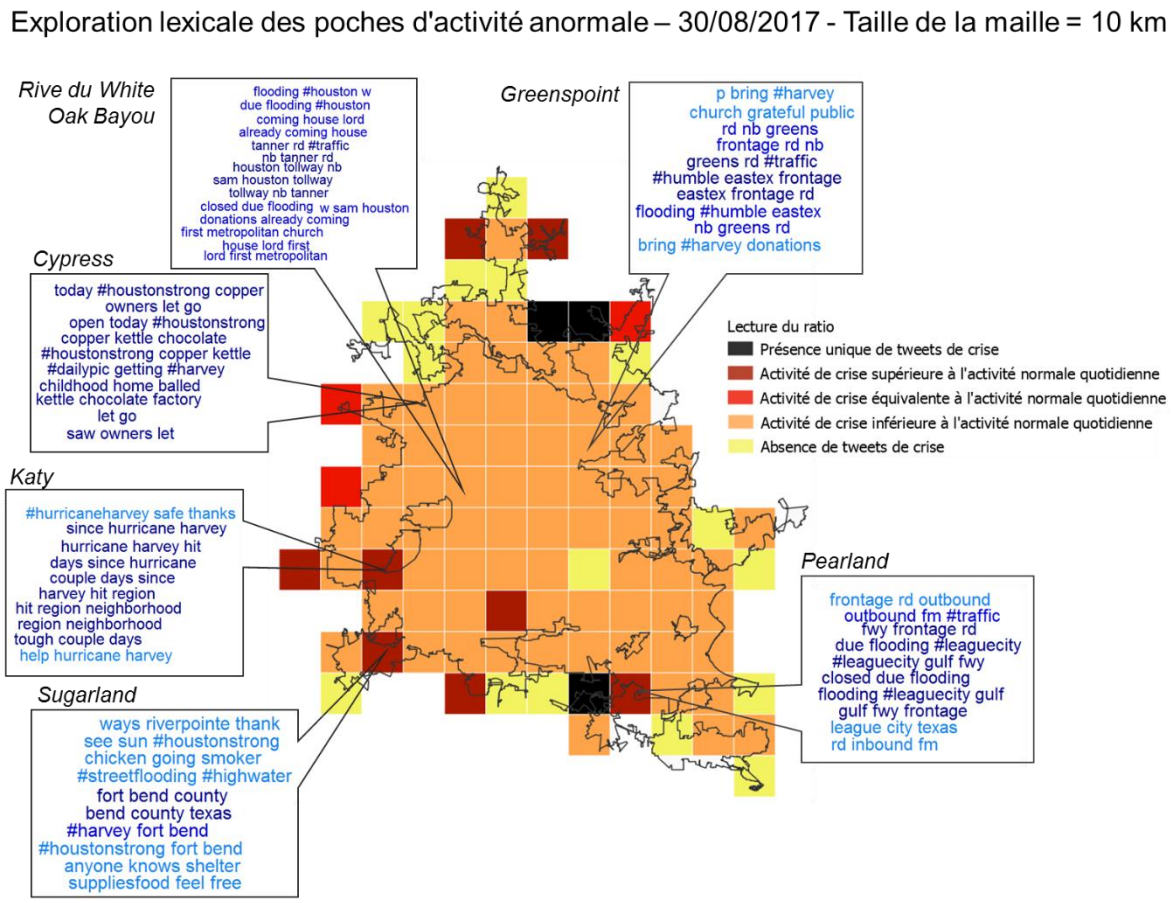


Figure 6.68 : Suivi de l'évolution des mailles au 31 août 2017 (C.Cavalière)

Au 30 août 2017, seules trois mailles affichent encore une activité virtuelle supérieure à la normale : Katy, Sugarland et Pearland (cf. figure 6.68) :

- à Katy, l'activité virtuelle reste principalement liée à un seul utilisateur, non mobile, qui fait l'état des inondations (9 tweets émis en 30 minutes) : "*this is what 2.5 feet of rain looks like in our neighborhood*", "*kashi house now underwater, we are safe*", "*no words, just water #grandmissionisflooded*", *etc.* ;
- à Sugarland, on enregistre davantage d'utilisateurs mais le discours se montre très hétérogène. Certains tweets sont des témoins directs des événements : "*these are the last of*

hurricane harvey photos, life is slowly returning to normal here" qui dénote avec les nombreux tweets enregistrant le besoin d'aide ("*we are here in this community to help out.... come on in, we are also accepting donations*", "*drop a comment, repost, share.... call me in my store @2812327443 I'm here to help*", "*looking to help! I have tons of clothes and towels and blankets to donate*").

- à Pearland, et en dehors des tweets à consonance officielle, le discours est davantage orienté sur le retour des habitants : "*friends are everything. took hours of work and a number of people to clear my friends house*", "*u.s. border patrol helping the flooding victims get back to check on flooded homes*".

Pour résumer, on peut trouver un sens spatial (mais non représentatif de la diversité des situations locales) à la distribution des tweets de crise géolocalisés en créant un indice sous forme de ratio, qui a l'avantage d'atténuer les effets de présence de tweets de crise dans les territoires qui sont actifs de manière habituelle. Mais pour trouver un sens qualitatif à la sémantique, il faut analyser les tweets dans une résolution temporelle fine, dans la même idée de gommer l'effet de la diffusion de ces messages à consonance officielle qui masquent l'information focalisée sur les personnes. En effet, dès qu'on collecte quelques dizaines de tweets de crise géolocalisés sur des unités spatiales de dix kilomètres de côté, on semble perdre du sens sémantique : en raison de la forte hétérogénéité des contenus, un tweet contenant des informations personnelles pertinentes peut être noyé dans des tweets énonçant un contenu répétitif (comme les voiries fermées en raison des inondations).

En outre, il est nécessaire de mettre en évidence ces tweets aux informations personnelles car ce sont eux qui offrent une porte d'entrée pour l'appréhension du vécu et des centres d'intérêts des individus pendant et après la crise. On peut prendre comme exemple ces deux tweets, postés le 27 août 2017 dans le quartier de Sugarland : ils ne sont séparés que de 1,5 kilomètre et de deux heures mais révèlent des sensibilités virtuelles radicalement différentes dans la motivation à tweeter : "*floated down our street (that never ever floods!!) 🐊 in all seriousness i hope everyone stays safe*", *"home safe!!! cats are fine, except they puked *everywhere*. good thing we're getting a new floor*." Dans un tel contexte, on ne peut toujours pas valider la pertinence de la première loi de Tobler quant aux tweets de crise géolocalisés.

Cette variabilité sémantique, qui marque une hétérogénéité certaine dans les préoccupations des individus en période de crise, et sans doute également une hétérogénéité dans les impacts de la crise sur les individus, peut-elle être recoupée à une variable statistique de terrain ? Nous testons alors l'activité virtuelle de crise avec un dernier paramètre de terrain,

qui sera cette fois social : il s'agit de l'indice de vulnérabilité sociale (*Social Vulnerability Index*, *SVI*, cf. chapitre 4) des individus recensés dans une entité géographique[17].

6.3.2. Appréhender la diversité des profils de territoires par la sémantique de l'activité virtuelle

6.3.2.1. L'activité virtuelle constitue-t-elle un marqueur de la vulnérabilité des territoires et de leurs populations ?

Dans un premier temps, nous souhaitons savoir si l'activité virtuelle de crise peut être considérée comme un marqueur de la vulnérabilité des territoires. La figure 6.69 ci-après représente les tweets clustérisés émis entre le 27 et le 31 août 2017 sur la métropole de Houston, ainsi que l'indice de vulnérabilité sociale par *census tract* (NB : la cartographie ayant été réalisée sous RStudio avec le package *leaflet*, la sémiologie graphique s'avère différente de celle des géographes : le nombre de tweets agrégés varie en fonction du niveau de zoom [si un tweet n'est pas agrégé, il apparaît alors sous la forme d'une punaise bleue] ; le nombre figurant au centre du cluster correspond au nombre de tweets agrégés. Enfin, la gradation des couleurs en fonction du nombre d'entités agrégées est la suivante : vert [valeurs les plus faibles] > jaune > orange [valeurs les plus fortes]).

[17] Pour rappel, l'indice de vulnérabilité sociale correspond à l'attribution d'un rang à une entité : une valeur faible signifie que l'entité considérée est classée parmi les territoires les moins vulnérables du terrain d'étude et inversement.

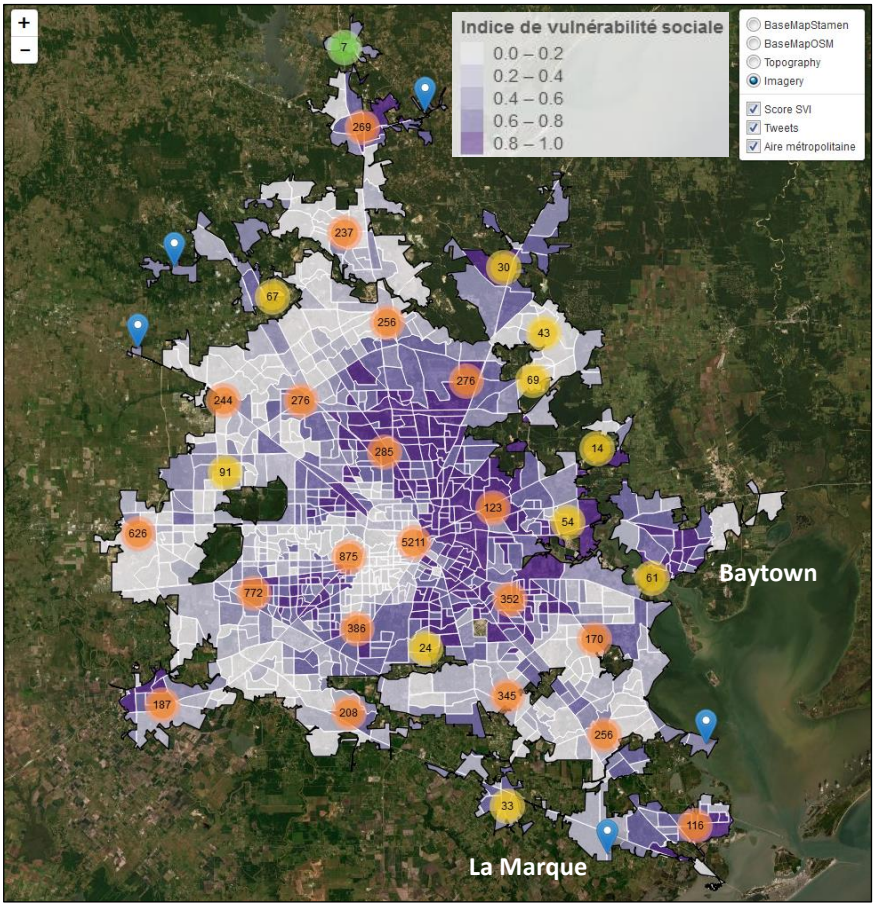


Figure 6.69 : Indice de vulnérabilité sociale et agrégats de tweets de crise, Houston, 27-31/08/2017 (C.Cavalière)

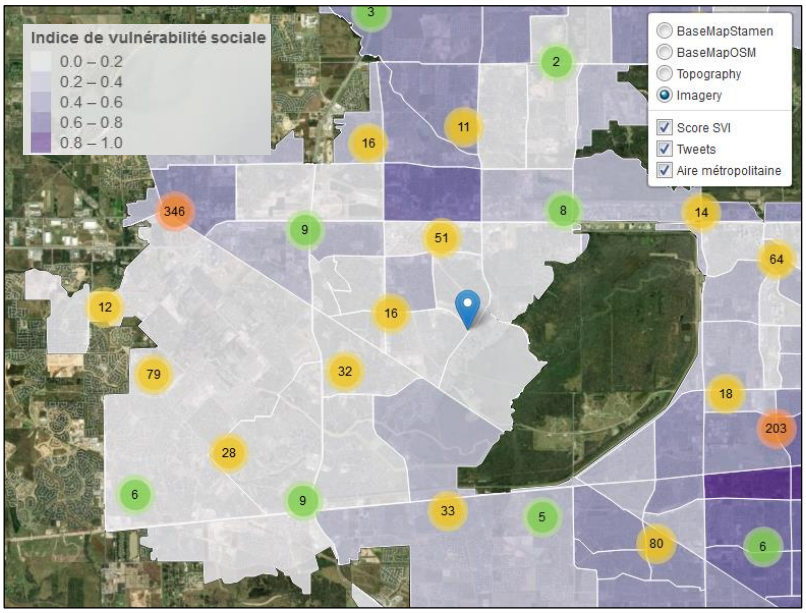
La cartographie générale ci-dessus ne dégage pas de tendance spatiale perceptible entre les deux variables : le *CBD* et les *census tracts* voisins qui agrègent le volume de tweets de crise le plus conséquent font partie des territoires les moins vulnérables. Pour autant, on identifie des poches d'activité dans les territoires périphériques de la métropole, qu'ils soient classés parmi les plus vulnérables (comme la couronne nord-est-sud autour des quartiers du centre ainsi que les quartiers des périphéries extrêmes est et sud, dont les noms sont indiqués sur la carte : Baytown et La Marque) ou le contraire. L'examen du nombre de tweets de crise inclus dans les *census tracts* regroupés en fonction des classes d'indices de vulnérabilité sociale fournit quelques précisions sur ces tendances (tableau 6.8) :

Tableau 6.8 : Vulnérabilité sociale et activité virtuelle des census tracts

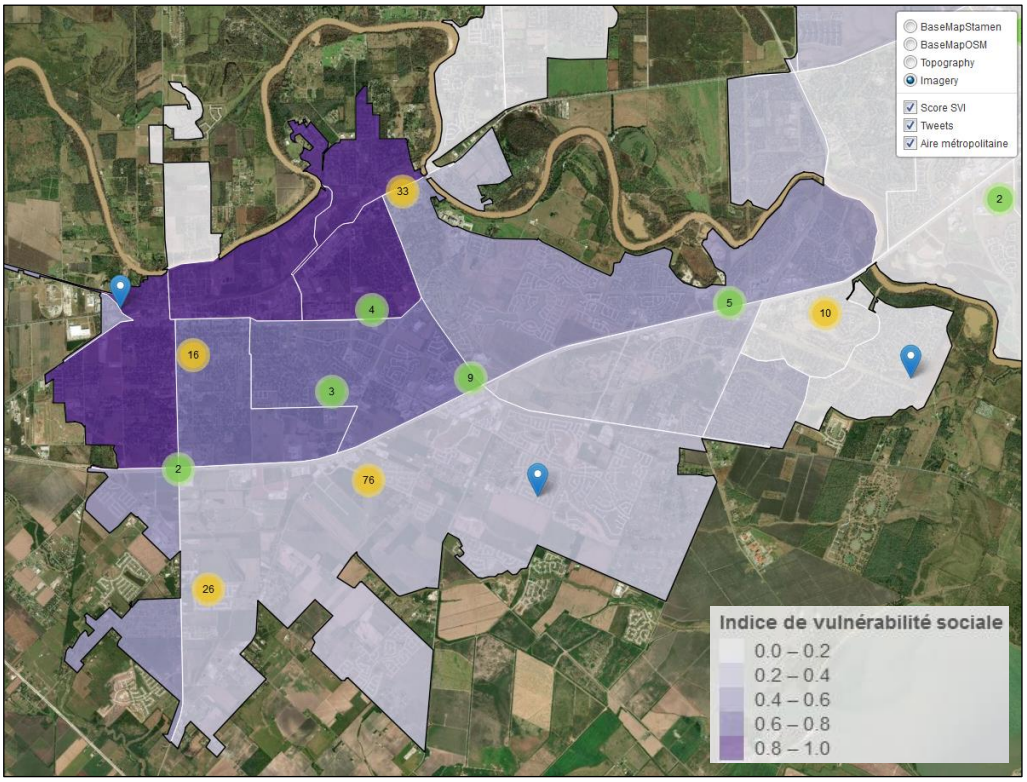
Classe	Census Tracts	Fréquence Census Tracts	Tweets de crise géolocalisés	Fréquence Tweets de crise géolocalisés
[0 ; 0,2[	240	23,98	2 389	20
[0,2 ; 0,4[	189	18,88	5 553	46,48
[0,4 ; 0,6[	149	14,89	1 788	14,97
[0,6; 0,8[	198	19,78	1 136	9,51
[0,8 ; 1]	225	22,48	1 081	9,05

Dans la métropole, 42,86% des *census tracts* affichent un score de vulnérabilité sociale strictement inférieur à 0,4 (ils font donc partie des 40% d'entités les moins vulnérables) ; pourtant, ces unités cumulent à elles-seules 66,48% des émissions de tweets de crise géolocalisés pendant la période étudiée (et notons que c'est la deuxième classe [0,2 ; 0,4[qui cumule les plus fortes quantités de tweets émis : 46,48% de l'ensemble des émissions regroupées sur 18,88% du territoire métropolitain). Les unités présentant les scores de vulnérabilité les plus élevés de la métropole (22,48% des *census tracts* affichent un score de *SVI* compris entre 0,8 et 1 et font donc partie des unités les plus vulnérables) ne sont virtuellement pas invisibles mais ne capturent que 9% des émissions totales de tweets de crise géolocalisés.

Dans un deuxième temps, si l'on zoome sur des territoires d'échelle plus fine, et parmi ceux qu'on a mis en évidence précédemment, on pourra alors dégager une tendance spatiale plus précise : dans les marges, l'activité vituelle ne semble pas s'inscrire dans les lieux aux populations les plus vulnérables. La figure 6.70 présente deux nouvelles fenêtres cartographiques de tweets agrégés à un niveau plus fin sur les quartiers de Katy (a) et de Sugarland (b) dans l'extrême ouest de la métropole : à Katy, alors qu'on avait précédemment détecté une suractivité les 27 et 31 août, les lieux depuis lesquels on enregistre les plus fortes émissions de tweets ne sont pas les plus vulnérables (NB : le cluster enregistrant 346 tweets correspond en fait à un spam) ; c'est en effet dans les *census tracts* de la partie sud du quartier qu'on enregistre le plus de tweets de crise (les scores de vulnérabilité oscillent entre 0,01 et 0,03). En revanche, à Sugarland, la fracture est nettement plus perceptible : la partie nord contient les territoires les plus vulnérables mais c'est la partie sud-sud-est (dont les scores de vulnérabilité sont compris entre 0,03 et 0,28) qui concentre les émissions de tweets les plus importantes.



(a). Katy



(b). Sugarland

Figure 6.70 : Indice de vulnérabilité sociale et agrégats de tweets de crise, sud-ouest de Houston, 27-31/08/2017 (C.Cavalière)

Nous répétons alors le test statistique afin d'observer l'éventuelle relation entre les valeurs de l'indice de vulnérabilité sociale et la mesure de l'activité virtuelle de crise par rapport à l'activité virtuelle normale. Dans ce nouveau test de khi², les entités de référence seront donc les *census tracts* de la métropole ; l'indice de vulnérabilité est classé d'après la

partition indiquée dans la légende cartographique des figures 6.69 et 6.70. Nous avons recalculé, pour chaque *census tract*, un indice fondé sur ratio entre le nombre moyen de tweets émis par jour du 27 au 31 août 2017 d'une part, et le nombre moyen de tweets émis par jour pendant la période de référence (avril-juin 2017). Le ratio est ensuite transformé en variable qualitative selon les modalités suivantes :

- si le census tract ne présente aucune activité (période de référence et période de crise), la modalité associée est *TOUJOURS INVISIBLE* ;
- si l'activité de crise est supérieure à l'activité normale (ratio > 1), la modalité associée est *ACTIVITE SUPERIEURE* ;
- si l'activité de crise est inférieure à l'activité normale (ratio < 1), la modalité associée est *ACTIVITE INFERIEURE* ;
- si l'activité de crise est nulle mais existante en temps normal, la modalité associée est *INVISIBLE CRISE*.

Si l'on regarde le graphique des écarts entre les effectifs observés et les valeurs du modèle d'indépendance (figure 6.71), on pourra noter :

- parmi les entités dont l'activité virtuelle de crise s'avère inférieure à la normale, une sur-représentation des territoires les moins vulnérables, ce qui rejoint la tendance qu'on avait précédemment mise en évidence par les mailles : une majorité des territoires affichent une activité inférieure à la normale en période de crise, même dans les hauts-lieux de l'activité virtuelle normale, comme le *CBD* ;
- mais surtout, si l'on observe les écarts liés aux territoires dont l'activité de crise est nulle, on constate une sur-représentation des lieux présentant les scores de vulnérabilité les plus élevés (classes [0,6 ; 0,8[et [0,8 ; 1]) alors que les lieux les moins vulnérables sont sous-représentés.

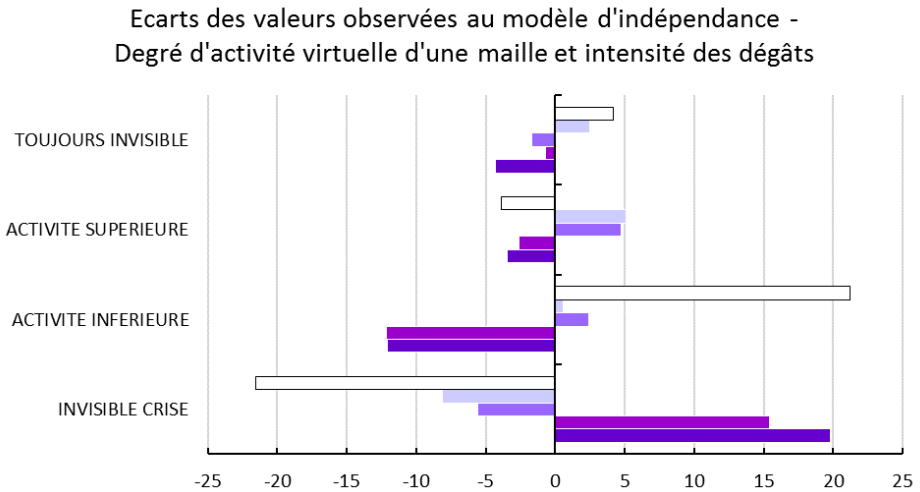


Figure 6.71 : Ecarts au modèle d'indépendance - Indice de vulnérabilité sociale et degré d'activité de la maille

Le test de Khi² indique une valeur de 40,31 (pour une valeur théorique de 26,22 avec 12 degrés de liberté et un risque d'erreur de 1%) ; la relation entre degré d'activité numérique en période de crise et indice de vulnérabilité sociale se confirmerait (mais elle reste en revanche de faible intensité, le V de Cramer calculé étant de 0,12). Ainsi, il y aurait une tendance à l'invisibilité des territoires les plus vulnérables en période de crise mais celle-ci ne peut pas être généralisée à leur ensemble.

6.3.2.2. Variations spatiales de la sémantique en fonction de la vulnérabilité

Si le degré de l'activité virtuelle ne peut pas être considéré comme un marqueur fort de la variabilité spatiale de la vulnérabilité sociale des populations, pouvons-nous au moins considérer la sémantique du tweet comme une variable significative par elle-même ? Autrement dit, les utilisateurs emploient-ils un vocabulaire particulier en fonction du profil de vulnérabilité sociale du territoire local, le *census tract* ?

Pour effectuer ce test sémantique, nous nous appuyons sur les constat suivant des résultats précédents : l'observation de discontinuités dans les valeurs de l'indice de vulnérabilité sociale entre des unités de recensement voisines (cf. figure 6.70.b). Ces discontinuités de vulnérabilité sociale marquent-elles également des ruptures au niveau du degré d'activité virtuelle des mailles ? Le cas échéant, ces ruptures d'activité sont-elles sémantiquement significatives ? Dans un premier temps, nous avons exploré le cas du quartier de Sugarland-Rosenberg, présenté dans la figure 6.70.b. La figure 6.72 ci-dessous présente la carte des indices de vulnérabilité sociale des *census tracts* du quartier (à laquelle on a ajouté un encart représentant les différents niveaux d'activité virtuelle de crise par rapport à la normale). Sur cette fenêtre cartographique sont représentés des nuages de mots construits à partir des tweets de crise géolocalisés émis entre le 27 et le 31 août 2017, et selon les règles suivantes :

- les tweets sont agrégés sur un pas de distance de 1 km ;
- dans les nuages de mots ne figurent que les adjectifs (en bleu) et les verbes (en rouge) : en effet, dans le chapitre précédent, nous avons vu que ces deux catégories grammaticales mettaient en évidence une plus grande variabilité sémantique qu'une représentation incluant l'ensemble des catégories de mots ;
- la police utilisée pour représenter les nuages varie toujours de la même manière : la taille est proportionnelle à la fréquence du mot mais son apparence varie en fonction de la classe de vulnérabilité : pour les classes de vulnérabilité sociale 1 et 2 (soit un *SVI* inférieur à 0,4), la police d'affichage est la suivante : Bradley Hand ITC ; pour la classe 3 (*SVI* compris entre 0,4 et 0,6), la police est la suivante : Calibri Light ; enfin, pour les classes 4 et 5, soit un *SVI* supérieur à 0,6 , la police employée est : Arial Narrow.

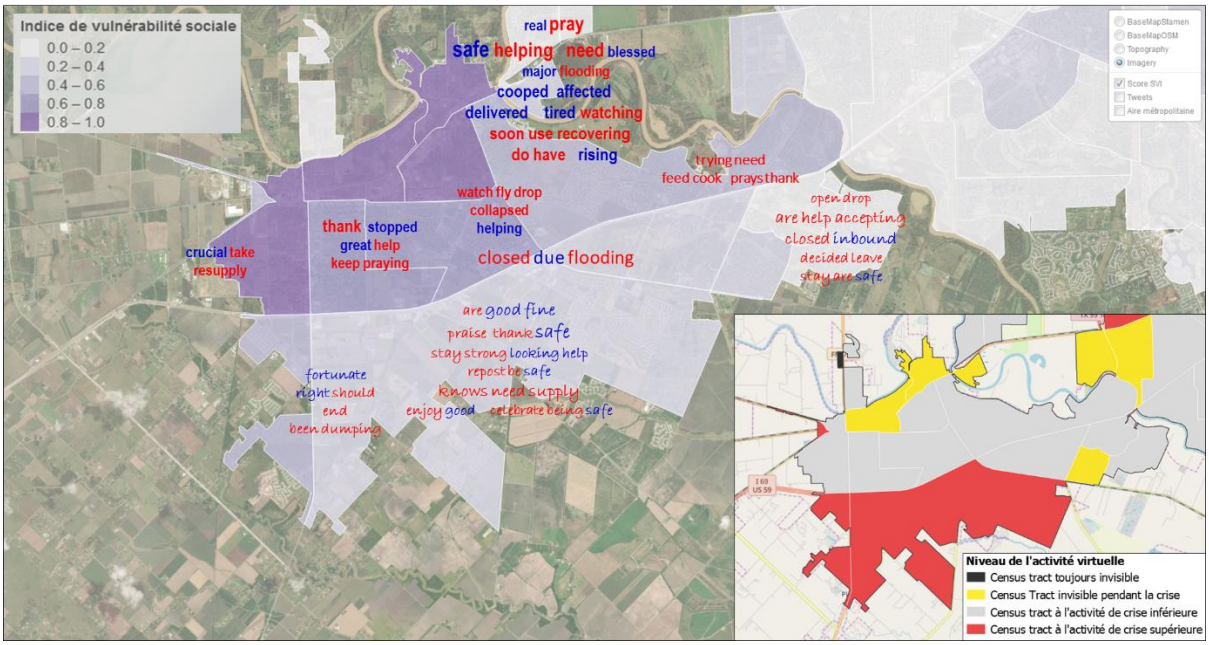


Figure 6.72 : Exploration lexicale (verbes et adjectifs) en fonction des profils de vulnérabilité et d'activité, Sugarland-Rosenberg (C.Cavalière)

A Sugarland, la sémantique de la réponse virtuelle se trouve spatialement structurée. En effet, bien qu'on retrouve certains adjectifs ou verbes communs (*safe*, *pray*, *help*) aux *census tracts* du quartier présentant des niveaux de vulnérabilité sociale différents, leur emploi marque des contrastes socio-spatiaux :

- le nord correspond aux lieux les plus vulnérables (*SVI* compris entre 0,6971 et 0,9893) et dans la proximité directe de la Brazos River. C'est dans ces entités spatiales du nord qu'émergent la majorité de *need* et *helping*, mais également les témoins d'inondation (*major flooding*), de dégâts aux infrastructures ("*right by my house, the road has collapsed*") et le besoin direct de vivres pour les sinistrés "*so cool to watch 3 helicopters fly in and drop off supplies for #hurricaneharvey flood victims*", "*take heart #rosenberg! crucial resupply after #hurricaneharvey is here! @ the ice house*". On peut également noter l'absence d'activité dans deux de ces *census tracts*, représentés en jaune dans l'encart (et qui contiennent cependant des lieux d'habitations).

- A l'est de ces territoires, on trouve des individus actifs mais qui restent dans des entités moins vulnérables (*SVI* compris entre 0,0339 et 0,4108) et dont le discours ne témoigne pas de la situation d'urgence : un utilisateur témoigne de son engagement dans les refuges "*getting set up to cook dinner for 75 folks in a rosenberg shelter. being hotdog and hamburger*" ; un autre marque son hésitation, du 28 au 31 août, à quitter son domicile (mais sans mentionner d'inondation ou de dégâts).

- Au sud, on trouve également des populations moins vulnérables (*SVI* compris entre 0,2023 et 0,2886) mais qui s'avèrent les plus actives pendant la situation de crise : c'est en effet au sud de Sugarland que se trouve l'anomalie de l'activité virtuelle de crise (en rouge

dans l'encadré de la figure 6.72). Ici, on identifie un discours radicalement différent : on retrouve certes des marqueurs d'engagement à l'entraide "*anyone that knows of a shelter in need of supplies/food feel free to contact me*". En revanche, les adjectifs *good*, *safe*, *fine* apparaissent à plusieurs reprises et on trouve également des témoins indiquant que les résidents du sud ont vraisemblablement été épargnés : *"now dry; about to enjoy good food... #hurricane #harvey #update"* ; "*we pop bottles for everything else so why not celebrate being safe with my family #hurricane*".

Pour vérifier si les résultats pertinents observés ici (le silence des lieux vulnérables, le vocabulaire révélateur des situations d'urgence dans le voisinage de ces lieux invisibles, et une activité supérieure à la normale qui n'est pas forcément la conséquence d'une crise grave), peuvent se répéter, nous avons exploré un second lieu test, plus au nord de la métropole. Ce second ensemble de *census tracts* est localisé dans le quartier de *North*, à 5km au nord du *CBD* et qui affiche également un contexte de rupture en termes de vulnérabilité et d'activité virtuelle. La cartographie résultante est représentée dans la figure 6.73 (les règles de construction sont identiques à la figure 6.72). Dans les faits, ce territoire exploré présente une enclave à l'indice de vulnérabilité faible (0,332) mais qui se trouve entourée de *census tracts* parmi les plus vulnérables de la métropole (sur une couronne nord-ouest-sud). C'est en revanche dans cette enclave qu'on identifie une activité virtuelle de crise supérieure à la normale (alors que les *census tracts* voisins sont soit invisibles, soit moins actifs).

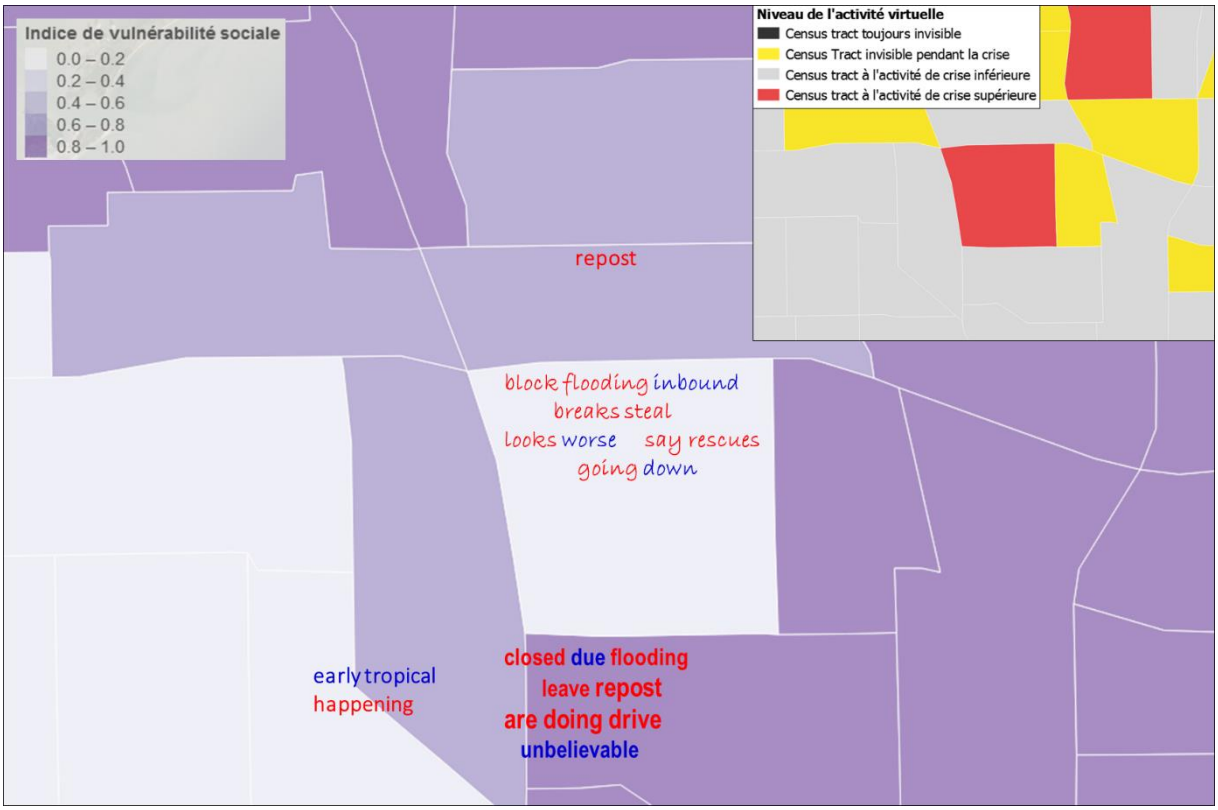


Figure 6.73 : Exploration lexicale (verbes et adjectifs) en fonction des profils de vulnérabilité et d'activité, North (C.Cavalière)

Dans le quartier de North, on trouve finalement peu d'information personnelle pour avoir la possibilité de comparer la sémantique employée à la variabilité spatiale de la vulnérabilité et du niveau de l'activité virtuelle. Dans les entités vulnérables et moins actives (notamment au sud de l'enclave), on retrouve en effet ce discours à consonance officielle annonçant la fermeture d'axes routiers inondés. La mention de secours s'identifie exclusivement dans l'enclave "*officials say as many as 1000 water rescues in the houston area so far*" mais ne concerne pas l'environnement direct de l'utilisateur ; par ailleurs, on ne peut distinguer que deux groupes de mots qui traduisent l'intensité ressentie du phénomène observé dans l'environnement direct, inclus dans des tweets émis dans deux *census tracts* à la vulnérabilité différente : "*Many intersections look like this or worse*" (à l'intérieur de l'enclave SVI de 0,332) et "*Still unbelievable #HurricaneHarvey 45 North and N Main*" (SVI de 0,943, et l'objet du tweet est de nouveau être une intersection routière).

Bilan de l'exploration des variations spatiales de l'activité virtuelle et de la vulnérabilité sociale des territoires

Etant donné que nous avons présenté les effectifs de tweets de crise comme un paramètre non significatif du vécu de la crise dans un territoire donné, nous avons alors créé des indices afin de gommer l'effet de l'activité habituelle. La variabilité spatiale des niveaux d'activité de crise, caractérisés à partir de ces indices, met en évidence deux comportements virtuels :

- l'activité de crise s'avère finalement inférieure dans la majorité des entités, y compris dans celles qui correspondent aux foyers habituels de l'activité virtuelle normale ; en revanche, la détection d'une activité moins prononcée en période de crise ne correspond pas à un signal unique : certaines entités qui présentent ce profil enregistrent de fortes concentrations de dégâts (ce qui laisse supposer une crise plus intense) alors que d'autres restent peu affectées.
- D'autres entités restent silencieuses ou au contraire, s'activent.

Le niveau de l'activité virtuelle de crise est-il un bon indicateur pour renseigner la gravité d'une situation locale ? En comparant ces différents degrés d'activité aux valeurs de l'indice de vulnérabilité sociale, il apparaît rapidement que les entités invisibles dans l'événement virtuel ont tendance à correspondre à des territoires aux populations vulnérables alors que les entités témoignant d'une activité supérieure s'observent plutôt sur des territoires moins vulnérables. En outre, à l'échelle locale d'un quartier, on peut repérer des discontinuités dans la vulnérabilité, qui se reflètent dans le degré d'activité virtuelle des entités. La sémantique peut-elle alors traduire ces discontinuités qu'on a observées en croisant les degrés d'activité virtuelle aux indices de vulnérabilité sociale ?

Bilan de l'exploration des variations spatiales de l'activité virtuelle et de la vulnérabilité sociale des territoires (suite)

Dans le cas du quartier de Sugarland-Rosenberg, le résultat s'est avéré positif : on constate bien une hiérarchisation spatiale de la sémantique en fonction de la vulnérabilité et de l'activité virtuelle :

- les territoires les plus vulnérables sont invisibles ;
- leurs voisins, moins vulnérables, témoignent d'une activité virtuelle de crise inférieure à la normale mais focalisée sur les secours et le besoin d'aide ;
- dans les entités du sud, les moins vulnérables, on constate une suractivité ainsi qu'un vocabulaire centré sur les individus indiquant qu'ils sont saufs.

Après cette première exploration, on pourrait avancer l'hypothèse selon laquelle, d'une part, les discontinuités spatiales de l'indice de vulnérabilité sociale, qui se recoupent aux discontinuités du degré de l'activité virtuelle, peuvent marquer la survenue d'une crise locale intense, et d'autre part, que la sémantique (résumée aux adjectifs et aux verbes) peut à elle seule témoigner de la survenue de cette crise. Le problème reste que le comportement virtuel local détecté à Sugarland semble propre à ce lieu mais n'est pas généralisé à d'autres territoires : dans le quatier de North, même si l'on observait ces ruptures spatiales, la sémantique des tweets se montrait trop pauvre pour estimer la gravité de la crise en fonction des profils des entités de recensement.

Conclusion du chapitre 6

Dans ce dernier chapitre, nous nous sommes attachés à la description du comportement spatio-temporel de l'activité virtuelle dans le milieu le plus propice au numérique, l'aire métropolitaine. L'objectif principal consistait à estimer la pertinence du tweet de crise géolocalisé comme marqueur des dynamiques et des paramètres du phénomène réel.

En réponse à cette piste de recherche, le premier constat effectué est le suivant : à l'échelle de l'aire métropolitaine, la survenue d'un phénomène extrême d'origine hydrométéorologique ne modifie pas les structures fondamentales de l'activité virtuelle géolocalisée : en effet, qu'il s'agisse de Houston ou de San Antonio (les deux villes test sélectionnées à l'échelle globale de l'aire métropolitaine), les hauts-lieux de l'activité virtuelle restent identiques, et en particulier dans le cas du *CBD* qui cumule le volume de tweets le plus conséquent, en période normale comme en période de crise. En revanche, la cartographie des tweets géolocalisés reste, à cette échelle plus locale, une carte des *vides* et des *pleins* ; en conséquence, même si la métropole constitue le lieu des pratiques numériques, tous ses territoires ne sont pas équitablement représentés. De même, si l'on cherche à établir des modèles explicatifs de la spatialisation de l'activité virtuelle géolocalisée, il s'avère difficile d'identifier des facteurs généralisables à tout lieu et dans tout contexte : la logique centre-périphérie autour d'objets, à San Antonio, ne se présente pas comme telle à Houston ; par ailleurs, dans cette métropole côtière, l'activité virtuelle de crise apparaît même comme une variable non-autocorrélée (témoignant ainsi de l'hétérogénéité de l'activité virtuelle entre des entités voisines).

Dans un deuxième temps, nous cherchions à mesurer la dynamique de l'événement virtuel et son rapport au réel, tout en nous référant au critère de distance des tweets entre eux, mais aussi par rapport à des marqueurs du réel (en l'occurrence, les tweets associés aux stations de l'USGS). En premier lieu, la réactivité locale face à des marqueurs de phénomènes réels n'est pas systématique : dans le cas des phénomènes récurrents, l'amorce de l'augmentation des flux de tweets de crise s'effectuait avant l'émission des tweets marqueurs de phénomènes en cours. Mais dans le cas de l'ouragan, on observait une réactivité directe à ces marqueurs ainsi qu'à la montée des eaux : notamment, l'effet de réactivité décrit par (Bossu *et al*., 2018) en réponse à un séisme, était ici perceptible dans les trente premières secondes après émission d'un tweet marqueur de l'USGS.

L'étude de la réactivité temporelle a ainsi été effectuée sur divers pas de temps et indique une résolution fine (le quart d'heure) permettant de détecter des variations spatiales et sémantiques (et indépendamment des effectifs de tweets émis). Cette réactivité temporelle se traduit, sur le plan spatial et sémantique, par les comportements suivants :

- le *CBD* reste un lieu d'activité permanente ;

- un pic identifié sur le graphique représentant les flux temporels de tweets peut se manifester par une agitation locale sous forme d'un îlot de tweets ou par une agitation dispersée dans les différents territoires de la métropole. Pour autant, cette agitation soudaine n'est pas forcément significative. En effet, la cohérence sémantique dans un îlot de tweets ou lors d'une agitation subite s'identifie principalement dans deux cas : lors de la diffusion de messages à consonance officielle (comme la fermeture de routes) ou encore lors d'une phase rapide de retweets d'un unique message.

Dans les autres cas, nous avons constaté que l'information émise par les individus au cœur de la crise, renseignant leur environnement local, était très irrégulière. En revanche, l'information que nous avons qualifiée de micro-événement individuel (soit un tweet de crise qui n'indique pas un vécu direct ou une situation locale) s'avérait très inscrite sur le réseau, et notamment dans le *CBD*. En conséquence, une trop forte diversité locale de la sémantique ou la présence de ces tweets de micro-événement pourrait être interprétée comme un indice de moindre gravité locale de la crise.

Dans cette perspective, nous nous sommes finalement focalisés sur la question de l'alternance entre périodes d'activité et de silence, ainsi que de l'invisibilité numérique locale, à travers la définition d'un indice mesurant le niveau d'activité de crise d'une entité spatiale, que nous avons recoupé à l'indice de vulnérabilité sociale défini par le *CDC*. En premier lieu, le degré de l'activité virtuelle de crise se positionne comme un paramètre plus significatif que l'effectif de tweets de crise en un lieu donné, dans la mesure où nous avons pu déterminer l'existence de relations statistiques entre le degré d'activité de crise d'une entité spatiale et l'intensité des dégâts ou encore l'indice de vulnérabilité sociale. Les derniers résultats continuent ainsi de confirmer les propos tweetés par Anthony Robinson en ce qui concerne l'invisibilité des populations vulnérables : en effet, dans les *census tracts* figurant parmi les individus les moins bien classés au regard de la vulnérabilité des populations, on constate généralement une activité virtuelle de crise inférieure à la normale ou quasi nulle. Il est cependant possible de mettre en évidence l'existence de ces situations locales vraisemblablement critiques par l'activité sémantique des entités voisines : ce comportement était manifeste à Sugarland-Rosenberg, quartier dans lequel les entités les plus vunérables étaient inactives mais dont les entités voisines étaient présentes sur le réseau. Leur sémantique était ainsi hiérarchisée et pouvait être appréhendée comme un indicateur à part entière, de par la présence d'un vocabulaire indiquant l'urgence à proximité des territoires les plus vulnérables, alors que les utilisateurs des territoires moins vulnérables du sud émettaient des messages à consonance positive et égocentrés. Néanmoins, la répétitivité d'une telle configuration n'est pas encore constatée.

Partie II : Contributions à l'extraction et au traitement des tweets de crise géolocalisés - Conclusion

Dans ce travail, nous proposons une contribution méthodologique articulée autour de deux axes : l'extraction de tweets utiles à l'exploration d'un thème précis d'une part et les pistes d'analyse envisagées de ces tweets utiles, recoupés à des données témoins, d'autre part. Ces démarches d'extraction et d'analyse de données sont documentées par des schémas afin d'assurer leur reproductibilité, soit pour appliquer l'extraction de tweets utiles à d'autres types de phénomènes naturels ou événements sociaux, soit pour appliquer les analyses sur d'autres phénomènes de nature identique à ceux qui ont été explorés ici afin d'estimer la généralisation des résultats observés, dans des territoires différents ou dans des temporalités autres.

Quelles structures ou tendances les analyses ont-elles mises en évidence ? En premier lieu, il existe un déséquilibre indubitable entre les territoires en ce qui concerne leur visibilité numérique, et ce déséquilibre est perceptible à toutes les échelles géographiques, quel que soit le contexte considéré : tout territoire correspondant à un foyer d'activité virtuelle géolocalisée, à l'échelle de l'Etat comme à l'échelle de la métropole, est avant tout un individu statistique hors-normes (ce qui correspond aux métropoles à l'échelle de l'Etat et principalement au *CBD* dans chaque métropole). En conséquence, toute analyse géographique qui est fondée sur la sélection d'un lieu en fonction du nombre de traces qu'il contient arbore une dimension spatialement discriminante.

Dans un second temps, il s'avère difficile de mettre en évidence des logiques spatiales systématiques qui seraient explicatives de l'activité virtuelle géolocalisée de crise : l'existence d'un phénomène pluvieux intense ne garantit pas une réactivité (dans le chapitre 5, nous avons vu qu'un phénomène de pluies-inondations qui avait frappé l'extrême est du Texas était quasiment passé inaperçu sur le réseau ; de même, à Austin, un phénomène naturel s'était retrouvé éclipsé par un événement culturel). Par ailleurs, la quantité de tweets émis en un lieu n'est pas le miroir de l'intensité des précipitations cumulées. En fait, quand on s'intéresse à ce paramètre de réponse virtuelle à l'échelle de la métropole, il apparaît que les structures mobilisées en temps de crise restent celles qui sont déjà activées en temps normal ; de la même manière, les principaux foyers d'activité virtuelle géolocalisée restent identiques, quel que soit le contexte environnemental. A ce constat, nous avons répondu que les effectifs et la localisation des foyers de l'activité virtuelle géolocalisée de crise ne constituaient pas des paramètres fiables pour renseigner les effets locaux de la crise.

La réactivité temporelle s'est en revanche révélée comme un paramètre sensible et fin dans le cas de l'ouragan (réaction enregistrée à l'échelle des dizaines de secondes, puis analysée à la résolution du quart d'heure) ; se traduit-elle alors par un sens spatial et

sémantique particulier ? En fait, à l'échelle globale de l'aire métropolitaine, les apports demeurent encore limités car les tweets de crise se montrent trop bruités :

- on a de nombreux messages à consonance officielle qui certes, apportent une certaine cohérence dans l'événement virtuel (quand par exemple on peut détecter une agitation soudaine liée à l'émission de tweets indiquant la fermeture de routes inondées), mais n'ont pas pour objet l'individu dans son environnement perturbé ;
- ce type de tweet (l'individu dans son environnement perturbé) se détecte pendant des phases d'agitation mais reste minoritaire et émis de manière trop irrégulière pour assurer un suivi précis des phénomènes et événements locaux ;
- à côté de ces tweets, on trouve encore deux autres comportements : les micro-événements individuels (soit, en général, les tweets à tonalité empathique qui ne nous donnent guère d'information sur le vécu de l'utilisateur) et l'agitation soudaine due à la diffusion d'un message unique qui disparaît dès le quart d'heure suivant.

En conséquence, et après les résultats des analyses sémantiques, nous avons émis l'hypothèse selon laquelle l'hétérogénéité des thèmes identifiés dans un agrégat de tweets, ainsi que la forte présence de vocabulaire témoignant de micro-événements, signalent probablement des territoires moins affectés que ceux qui affichent des contenus homogènes (c'était notamment le cas de Port Arthur pendant le passage d'Harvey). En fait, si l'on recherche des paramètres significatifs de l'activité virtuelle de crise, les résultats se montrent plus probants si l'on se détache du contexte environnemental de la crise : c'est ce que nous avons tenté dans la dernière expérience, en recoupant l'indice de vulnérabilité sociale des populations dans une unité de recensement fine et un indice permettant de mesurer le degré d'activité virtuelle de crise. La cohérence spatiale et sémantique qu'on peut difficilement mettre en évidence à l'échelle globale de la métropole est ici perceptible : dans le quartier de Sugarland, on a ainsi pu observer une variabilité sémantique qui reflétait la vulnérabilité des territoires ainsi que le degré de leur activité virtuelle de crise (les territoires les plus vulnérables s'avèrent invisibles ; les territoires limitrophes, moins vulnérables, sont moins actifs qu'en temps normal mais contiennent un vocaculaire de gestion de crise ; les territoires les moins vulnérables sont plus actifs mais ne contiennent pas de témoin d'aide d'urgence ou de crise locale violente).

Conclusion générale

Conclusion générale

Rappel des problématiques et des questionnements de la recherche

Dans les sociétés contemporaines où tout aspect pratique du quoditien tend à s'enraciner dans des systèmes et outils numériques fonctionnant en réseau, les plateformes du Web social sont devenues le nouvel espace virtuel de gestion et d'expression des individus connectés, et quelles que soient les régions du monde considérées. Pour les générations et individus connectés, ce sont désormais les lieux dématérialisés du Web social qui sont à l'écoute de leurs préoccupations du moment. Il devient alors quasiment logique de s'interroger sur l'utilité, autre que commerciale, des diverses formes prises par les masses de contenus web créés chaque jour par les internautes.

Dans cette recherche se posait alors la question des apports effectifs d'un type de contenu issu du Web social, les traces numériques géolocalisées émises sur la plateforme Twitter, et générées dans le cadre d'une interaction directe entre l'individu producteur de la trace et le territoire aux conditions environnementales perturbées. Autrement dit, la question principale s'oriente autour des connaissances géographiques des interactions entre individus et territoires que pourraient nous fournir les traces numériques géolocalisées émises dans le contexte spécifique de la survenue d'une crise d'origine naturelle. Pendant une première période (2009-2015), force est de constater que le potentiel accordé à ces traces a été annoncé et validé de telle manière que la recherche académique les plaçait comme pilier d'une nouvelle connaissance à construire dans les sciences humaines et sociales (Sui et Goodchild, 2011 ; Elwood *et al*., 2012). Cette assertion, que certains auteurs qualifièrent ensuite de croyance, a finalement été mise en doute par des expériences ou observations concluant à un potentiel initialement surévalué des traces numériques géolocalisées comme les tweets, pour l'étude précise de questions socio-spatiales (Quesnot, 2016).

En ce qui concerne l'évaluation de ce potentiel (réel ou supputé ?) en termes de connaissances lors de la survenue de crises d'origine naturelle, le questionnement général énoncé dans le paragraphe précédent s'est traduit par la définition de trois hypothèses directrices, construites à partir d'une première série de publications focalisées sur les thèmes suivants : les effets de la transformation des données géographiques sous l'impulsion de l'adoption massive et universelle des outils de géolocalisation et du Web social ; les travaux existants fondés sur le recours au matériau tweet géolocalisé dans des thématiques géographiques ; enfin, la question de l'adaptabilité des outils d'analyse traditionnels aux nouvelles formes de données générées sur le Web social, et de la posture de recherche. Pour rappel, les trois hypothèses mentionnées s'articulaient ainsi sur les observations récurrentes relatives au comportement du réseau virtuel d'une part, et d'autre part, sur les propositions quant au cadre à fixer pour l'analyse de ces traces :

- parmi ces observations et constats récurrents, on pouvait mettre en évidence la capacité du réseau à réagir en concomitance avec les phénomènes et événements du monde réel ; le réseau se positionne donc comme un instrument de mesure fiable en termes de suivi des dynamiques du réel ;

- parmi les propositions relatives au cadrage méthodologique de la recherche, on soulignait l'existence d'un retournement épistémologique qui se détachait de la traditionnelle démarche hypothético-déductive au profit de l'association entre logiques inductive et abductive. L'approche déductive préconisée dans les sciences était en effet perçue comme un frein à la découverte de phénomènes imprévus ayant le potentiel d'apporter davantage de connaissances que des comportements virtuels qu'on pourrait intuitivement déduire. En d'autres termes, utiliser la démarche hypothético-déductive avec les traces numériques géolocalisées représenterait la prise de risques suivante : orienter les analyses en fonction d'une théorie préconçue et laisser de côté tous les faits observés qui ne s'y rattachent pas, étant donné la récente disponibilité d'un matériau dont on ne connaît pas l'ensemble des facettes et qui évolue rapidement ;

- enfin, étant donné que le chercheur collecte les traces en dehors de toute connaissance des éléments ou conditions environnementales qui incitent l'activité virtuelle, les émissions de tweets géolocalisés doivent être recontextualisées dans leur environnement (ce qui constitue la condition essentielle afin d'identifier les éventuels facteurs physiques qui déclenchent la motivation de l'individu à tweeter).

En dépit de ces hypothèses directrices (qui seront débattues dans les contributions), il fallait considérer un certain nombre de verrous méthodologiques et thématiques, notamment en ce qui concerne les conditions actuelles d'utilisation des plateformes numériques et leurs répercussions sur les questionnements faisant intervenir des concepts géographiques :

- le choix d'un terrain d'étude était d'ores-et-déjà contraint par sa visibilité numérique : il nous fallait certes sélectionner un territoire à risques, mais qui bénéficiait également d'une visibilité numérique palpable ;

- les méthodologies d'extraction de tweets dont le contenu est lié à un thème particulier ont besoin d'être améliorées car elles ne font pas l'objet, dans la littérature, de recherches approfondies alors que le jeu de tweets trié représente la base des analyses accomplies ;

- on ne sait pas si les résultats ayant été observés sur un territoire précis, dans un contexte particulier, s'observent de manière répétée sur ce même territoire (dans une temporalité autre) ou sur un autre territoire (en d'autres termes, on ne sait pas s'il existe des tendances comportementales virtuelles universelles) ;

- il faut considérer les tweets géolocalisées comme le produit de l'activité d'un échantillon de l'ensemble des utilisateurs de Twitter (et en l'occurrence, il s'agit des utilisateurs qui acceptent l'activation du GPS en émettant des tweets), eux-mêmes formant une population à part entière parmi l'ensemble des individus recensés ou parcourant un territoire. On ne sait donc pas si le discours véhiculé par un tweet sur des conditions environnementales ou sur un objet particulier du territoire est partagé par l'ensemble des

usagers (habitants ou usagers de passage) de ce même territoire. Peut-on alors passer de la pratique individuelle exercée dans le lieu et transmise par le tweet géolocalisé émanant d'un individu seul à la pratique collective cohérente par l'ensemble des tweets émis en un lieu ?

- certains auteurs soulignent (Kitchin, 2013) le fait que les outils et méthodes d'analyse spatiale n'ont finalement peu ou pas évolué : quels outils et quelles données peut-on alors associer pour déployer le potentiel ainsi que les limites des tweets géolocalisés ?

L'orientation générale donnée à cette recherche ne consiste donc pas à proposer des solutions méthodologiques de traitement des tweets géolocalisés dans un contexte de gestion de crise en temps réel mais d'associer outils et méthodes d'analyses statistiques, spatio-temporelles et sémantiques afin de décrypter et de mesurer le potentiel de ces traces en termes d'apports de connaissances géographiques d'un territoire soumis à une crise d'origine naturelle. Par conséquent, notre approche se veut spatiale, cartographique et statistique.

Rappel des lacunes identifiées

L'état de la question a appréhendé un certain nombre de lacunes d'ordre méthodologique et thématique (dont certaines ont d'ores-et-déjà été exposées dans le paragraphe précédent), que nous pouvons résumer par les points suivants :

- *Lacune méthodologique n°1 - méthodes de collecte de jeux de tweets utiles* : d'une manière générale, un grand nombre de publications axées sur la question des risques et catastrophes naturels procèdent très rapidement à l'étape de recherche et d'extraction des tweets dont la sémantique est focalisée sur un thème précis. Ces méthodes sont généralement limitées à la définition, par les chercheurs, d'une poignée de mots-clés ou de hashtags sur lesquels s'appuie une unique requête d'extraction d'un jeu de tweets de crise servant de base aux analyses. Pour autant, les résultats d'analyses effectuées sur un jeu de tweets de crise non exhaustif sont-ils fiables ? L'extraction de tweets de crise étant l'étape primordiale à toute analyse, existe-t-il un risque de biais sur les résultats si le jeu trié est restreint dans sa collecte ? (Steiger *et al*., 2015) soulignaient alors le manque de méthodes explorant les tweets par leur contenu sémantique ou intégrant la dimension spatiale des tweets dans leur extraction.

- *Lacune méthodologique n°2 - associer les traces numériques géolocalisées aux données traditionnelles* : dans la plupart des cas, les méthodologies d'analyse croisant traces numériques géolocalisées et données officielles traditionnelles réduisent l'usage de ces dernières à une simple utilisation de fond ; autrement dit, elles fournissent des éléments de contextualisation de l'activité virtuelle mais ne sont pas nécessairement intégrées à l'analyse spatiale ou aux analyses statistiques (Steiger *et al*., 2015).

- *Lacune méthodologique n°3 - peut-on créer de l'information géographique à l'aide des traces numériques géolocalisées ?* Cette question découle directement de la lacune méthodologique n°2 exposée ci-dessus : il s'agit des possibilités de valorisation des tweets

géolocalisés par GPS, associés aux données traditionnelles, en une information géographique cartographiable destinée à faire émerger des structures spatiales recoupant l'ensemble des variables considérées. En d'autres termes, peut-on, à partir d'une entité spatiale délimitée (comme une maille), dans une résolution spatiale et temporelle donnée, créer la carte thématique classifiant ces entités en fonction d'une série de paramètres combinant données physiques, données sociales et les différents contenus des traces numériques géolocalisées ? Est-il donc envisageable de transformer des traces spatio-temporelles hétérogènes, dont la caractéristique principale est d'associer une dimension quantitative (combien de traces sont inventoriées en un lieu ?) à une dimension qualitative (qu'est-ce que ces traces décrivent ?), en un unique type d'information, restitué sur la carte ?

- *Lacune méthodologique n°4 - comment et par quels outils faire de la recherche géographique fondée sur les tweets géolocalisés ?* Les tweets géolocalisés, et d'une manière générale, l'ensemble des traces numériques géolocalisées, constituent un matériau nouveau dont les propriétés heurtent les cadres et approches professionnels traditionnels vis-à-vis des données : quand on collecte les jeux de données via les portails institutionnels, on a d'ores-et-déjà une idée des contenus que l'on recherche pour résoudre un problème précisément ciblé, et de la production cartographique qui en résultera. Avec les traces numériques géolocalisées, une infinité de domaines peuvent être capturés, offrant ainsi un supposé accès à une nouvelle connaissance des rapports entre populations et territoires. Mais paradoxalement, on éprouve des difficultés à approfondir les analyses existantes et surtout, à valoriser l'ensemble des dimensions des traces numériques géolocalisées. Les questions posées autour de la méthodologie d'exploitation des tweets géolocalisés par GPS s'articulent ainsi autour de deux points :

- *l'approche épistémologique* : face aux propriétés des traces numériques géolocalisées et en considérant leur possible inconstance, certains auteurs (Anderson, 2008 ; Kitchin, 2013 ; Miller et Goodchild, 2015) préconisent logiquement le détachement de la méthode hypothético-déductive (comme on ne sait pas *a priori* ce qu'on peut trouver ou pas, ni ce qu'on peut chercher ou pas) pour l'adoption d'une approche mêlant abduction et induction, c'est-à-dire la formulation d'hypothèses *a posteriori* et à leur test. Cette approche permet-elle de mettre en évidence des observations dont les publications existantes sur les phénomènes naturels (Dashti *et al*., 2014 ; Blanford *et al*., 2014 ; Hertfort *et al*., 2014 ; Alam *et al*., 2018) ne font nulle mention (ou de les nuancer *a minima*) ?
- *la question de l'adaptabilité des outils et méthodes dont on dispose actuellement* : analyser les tweets géolocalisés dans une perspective géographique revient à mobiliser des outils existant depuis quelques décennies (SIG, outils d'analyse spatiale et statistique, clustérisation, *etc.*) auxquels se greffent les outils d'analyse et de représentation lexicale. Le problème reste qu'on ne sait pas si ces outils, développés à l'origine pour traiter des jeux de données structurées et standardisées, offrent des résultats pertinents lorsqu'ils sont appliqués à des traces dont les

propriétés sont à l'opposé des données officielles (en dehors du fait qu'elles disposent d'une composante spatiale et temporelle). Il s'agit alors de discriminer les méthodes (aussi bien spatiales que sémantiques) fournissant des résultats exploitables de celles qui sont trop sélectives ou qui conduisent à des impasses.

- *Lacune thématique n°1 - le tweet de crise géolocalisé est-il un marqueur socio-spatial et temporel pertinent ?* D'une manière générale, les travaux existant dans la thématique des risques naturels et fondés sur les tweets géolocalisés, ont tendance à considérer systématiquement ce matériau comme marqueur des phénomènes et événements consécutifs du réel (Dashti *et al*., 2014 ; Hertfort *et al.*, 2014). Mais selon quels critères appréhender le tweet de crise géolocalisé comme marqueur et quelle est sa pertinence ? En effet, les approches s'intéressant à l'événement virtuel dans l'ensemble de ses dimensions et en rapport aux contextes sociaux et environnementaux sous-jacents restent rares et récentes (Zou *et al*., 2018 ; Samuels *et al.*, 2018). En accord avec l'approche épistémologique proposée, il nous semble essentiel de considérer le comportement de l'activité virtuelle de crise dans sa globalité, c'est-à-dire par la méthode *Qui, Quoi, Où, Quand, Comment, Combien, Pourquoi ?* Une nouvelle fois, comme le matériau est récemment apparu et que le numérique et ses pratiques évoluent en permanence, on ne sait pas si les résultats d'une étude publiée en 2010 sont encore valables en 2020. Au-delà de cette remarque, c'est également la question de la généralisation des résultats d'une étude qui se pose : est-elle seulement envisageable compte-tenu de la variabilité spatiale et sociale de la production de traces numériques géolocalisées d'une part, et d'autre part, si l'on considère la pérennité et l'évolution des usages des plateformes dans le temps ?

- *Lacune thématique n°2 - le tweet de crise géolocalisé a-t-il un potentiel de renseignement de terrain ?* Cette question découle de la lacune exposée ci-dessus : le tweet de crise géolocalisé peut-il, à l'heure actuelle (ou du moins pourrait-il, en énonçant quelques recommandations), avoir le potentiel de constituer une source de renseignements complémentaires de terrain dans un contexte, la crise, où cette information fait justement défaut, aussi bien en termes d'optimisation de la gestion de crise que d'évaluation post-crise ou encore d'étude des rapports entre populations connectées et territoires en crise ? Dans le cas de la gestion de crise en temps réel, le tweet géolocalisé a déjà acquis une certaine notoriété pendant les catastrophes naturelles survenues au début des années 2010 (séisme de Haïti, Fukushima, ouragan Sandy, séisme au Népal). Mais en 2017, après les ouragans Harvey et Irma dans le Golfe du Mexique, le discours académique commençait à nuancer les apports jusqu'alors envisagés comme positifs des tweets de crise géolocalisés, notamment sur fond d'équité socio-spatiale et de visibilité numérique des territoires et des populations affectés. Faut-il ainsi appréhender ce matériau comme une perspective ayant le potentiel de combler un manque, lui bien réel, ou comme un miroir aux alouettes ?

Résumé des contributions

Les contributions de cette recherche se déclinent en fonction de trois points :

- les méthodologies d'extraction de jeux de tweets utiles, destinées à compléter les approches, souvent trop sommaires, rencontrées dans les travaux existants ainsi que la proposition d'un environnement de géovisualisation assurant l'exploration spatiale et sémantique et l'extraction de tweets utiles.
- La mise en évidence des propriétés spatiales, temporelles et sémantiques des événements virtuels consécutifs à des phénomènes naturels d'intensité différente, qui permet de nuancer le potentiel annoncé (cf. chapitre 3) du tweet géolocalisé par GPS dans la problématique des risques et catastrophes naturels.
- la question des méthodologies et des outils d'analyse qui fournissent des résultats exploitables (ou qui conduisent à des impasses) lorsqu'ils sont appliqués à des traces numériques géolocalisées de forme ponctuelle comme les tweets géolocalisés par GPS : comment et avec quels outils analyser ces traces, et pour produire quoi (autant en termes de connaissances que de représentation cartographique) ?

Contribution n°1 - Proposition d'une démarche de recherche et d'extraction d'information de crise et formalisation d'un environnement de géovisualisation assurant la généralisation des étapes proposées

Dans ces travaux, nous avons proposé une démarche de recherche et d'extraction fondée sur trois approches complémentaires dont la recommandation dépend non seulement du type de phénomène considéré mais également de l'échelle du terrain d'étude retenu : l'extraction lexicale par la recherche exclusive de hashtags pour un phénomène extrême de forte intensité à l'échelle globale, l'extraction lexicale toutes catégories de mots-clés confondues pour un phénomène de moindre ampleur mais habituel et pour finir, l'extraction lexicale d'échelle locale, fondée sur une primo-approche spatiale encadrant la sélection des tweets géolocalisés non pas en fonction d'une entité administrative mais en fonction d'éléments environnementaux. Mais surtout, quelle que soit l'approche considérée, la démarche se veut plus exhaustive dans la mesure où elle offre la possibilité d'associer cette fréquente approche supervisée (qui qualifie le recours à une extraction de tweets à partir de mots exclusivement choisis par les chercheurs) à une approche non supervisée, autrement dit la mise en évidence et le recours à une extraction par le vocabulaire contenu dans les tweets de l'événement virtuel.

Les trois approches ont été testées empiriquement sur les phénomènes suivants : la tempête de blizzard Jonas en janvier 2016, à l'échelle des Etats-Unis pour l'approche lexicale par hashtags seuls (49 826 tweets extraits) ; les phénomènes récurrents de pluies/inondations survenus au Texas au printemps 2016 pour l'approche lexicale par toutes les catégories de mots-clés (46 192 tweets extraits) ; la crue de la Seine dans le bassin parisien en juin 2016 (674

tweets extraits) ainsi que les territoires du Texas et de Louisiane frappés par l'ouragan Harvey en août 2017 (47 297 tweets extraits) pour l'approche intégrant la dimension spatiale (en prenant comme références respectives le tracé du fleuve et les cumuls pluviométriques liés à l'ouragan)[1]. Enfin, dans le but d'assurer la reproductibilité de l'ensemble des étapes des approches proposées tout en optimisant leur facilité et leur rapidité d'exécution, nous avons constitué le cahier des charges d'un environnement de géovisualisation destiné à supporter lesdites étapes, et qui a bénéficié d'une première phase de développement intégrée au projet Sakura du LIG[2].

Contribution n°2 - La mise en évidence des multiples facettes et de la complexité du comportement de l'activité virtuelle générée en réponse à un phénomène et à une crise d'origine naturelle

La recherche considère-t-elle à tort ou à raison le tweet de crise géolocalisé par GPS comme un témoin pertinent des perturbations environnementales et de leurs effets sur les territoires et populations ?

- *Une géographie virtuelle sélective de ses territoires*

Sur le terrain d'étude exploré, l'activité virtuelle, qu'on la considère en situation normale ou comme la réponse à une crise d'origine naturelle, se présente avant tout comme l'expression d'une minorité de territoires : statistiquement, cette minorité se manifeste, à l'échelle des comtés, sous la forme d'une poignée d'individus - qui oscille entre 3% en milieu rural et 18% en milieu métropolitain - à l'activité virtuelle hors-normes[3] (cf. figure 5.26) et spatialement, par une cartographie des vides et des pleins, ceci quelles que soient les échelles considérées. Quels sont ces vides et quels sont ces pleins ?

A l'échelle globale du Texas, en situation normale comme en situation perturbée, une dichotomie apparaît rapidement dans l'ancrage territorial de la production de tweets géolocalisés : les hauts-lieux de cette production restent les aires urbaines, et parmi ces aires urbaines, ce sont les aires métropolitaines qui cumulent les plus forts effectifs de tweets géolocalisés. Pour autant, il ne faut pas appréhender le milieu métropolitain comme un territoire dont la production numérique géolocalisée est équilibrée, puisqu'il adopte ce même comportement qu'on observe à l'échelle globale : une minorité de territoires concentrant les

[1] La comparaison des résultats, en termes de nombre de tweets extraits, avec les publications existantes reste cependant difficile : en effet, les travaux consultés et les quelques chiffres cités dans le chapitre 3 du manuscrit, en ce qui concerne les méthodologies d'extraction de tweets de crise, se fondent exclusivement sur le flux *sample* de l'API Streaming de Twitter, autrement dit le flux qui retourne, sans critère de filtrage additionnel des tweets, 1% du trafic mondial. Ce flux se trouve donc automatiquement soumis à un échantillonnage spatial. L'infrastructure du LIG est quant à elle connectée à flux *filter* de cette même API : elle collecte des tweets en fonction de critères sélectifs précis. En fonction de l'application de ce critère d'échantillonnage, le jeu de tweets bruts peut donc avoir un potentiel différent en termes de quantités de tweets de crise contenus.
[2] Lien vers la page du projet : https://sakura-platform.liglab.fr/
[3] Autrement dit, il s'agit des effectifs de tweets de crise géolocalisés inventoriés dans les entités en question.

plus fortes densités de tweets alors que la majorité des territoires restent peu inscrits voire quasiment invisibles. Ainsi, si l'on considère l'ensemble des tweets émis dans l'aire métropolitaine de Houston pour chacun des *census tracts* entre le 27 et le 30 août 2017, 10% de ces entités sont considérées comme des individus hors-normes (dont seulement quatre, parmi lesquels figure le centre des affaires, cumulent 45% des émissions de tweets de crise géolocalisés), et 40% de ces entités sont invisibles de l'événement virtuel (cf. section 6.1.2.1). Malgré ce défaut de représentation des territoires locaux, c'est ce milieu métropolitain qui reste le plus commode à explorer à une échelle spatiale et temporelle fine, à partir des traces numériques géolocalisées par GPS ; pour rappel, dans les milieux situés en dehors des métropoles, et même si les populations étaient considérées comme urbaines, les phénomènes naturels pouvaient se révéler quasiment absents du réseau (ce qui avait été le cas pour les inondations survenues dans le comté de Jefferson en avril 2016 [cf. section 5.2.1.4]), ou trop peu documentés (cas des aires urbaines entourant le *Lake Conroe* au Nord de Houston, pendant le passage de l'ouragan Harvey [cf. figure 5.60] : ici, même si les populations sont considérées comme urbaines, elles sont déjà peu actives sur le réseau virtuel géolocalisé dans des conditions normales).

Comment se structure l'activité virtuelle dans ces milieux métropolitains ? Le maillage de l'activité virtuelle alterne entre des centres aux fortes densités de tweets géolocalisés (que l'on considère l'activité normale ou l'activité de crise), dont le plus volumineux reste le centre des affaires, et des périphéries au maillage plus lâche. Mais lorsqu'on cherche des facteurs explicatifs de ces centres, ils restent difficiles à modéliser : à San Antonio, il apparaissait clairement que l'activité virtuelle se structurait autour d'objets particuliers, alors qu'à Houston, ces pôles s'avéraient plus complexes, associant à la fois des objets ponctuels mais également linéaires (cf. sections 6.1.2.1 et 6.1.2.2).

Par ailleurs, cet effet de sur-représentation d'une minorité d'individus spatiaux se retrouve également dans le comportement des utilisateurs (et quand bien même on sépare les tweets marqueurs d'alerte ou envoyés par les automates), y compris en situation de crise. Mais paradoxalement, la caractérisation de profils de populations créatrices de contenus géolocalisés (ou absentes de ce réseau) se révèle comme une question épineuse, et qui finalement, ne trouve pas de réponse fiable sur le terrain d'étude : en fait, l'examen sémantique des tweets de crise géolocalisés révèle une tonalité qui apparente la majorité des tweets de crise géolocalisés à des contenus de type professionnel (et non générés par des individus lambda dans le cadre de leur activité privée), mais surtout, dans les tests statistiques, l'activité virtuelle géolocalisée apparaît comme une variable non significative lorsqu'elle est mise en relation à des variables socio-démographiques (cf. section 6.1.3).

Pour conclure sur ce premier point, il convient de retenir que si l'on considère le potentiel d'un territoire virtuel à étudier en fonction de la quantité des tweets géolocalisés inventoriés, alors on pratique une géographie sélective de ses terriroires ; en outre, si l'on ne parvient pas à caractériser les populations productrices, alors toute connaissance spatiale est construite en dépit de la représentativité sociale des traces numériques géolocalisées sur le terrain d'étude.

- *Des logiques spatiales virtuelles de crise peu articulées à la spatialisation des phénomènes physiques réels*

Si l'on considère l'Etat du Texas à son échelle globale, les lieux qui s'activent virtuellement pendant la crise correspondent aux lieux d'ores-et-déjà empreints d'une activité virtuelle en temps normal (et à condition que l'on se réfère aux tweets de crise géolocalisés autres que ceux émis par les comptes du *NWS* et des stations météorologiques ou de jaugeage déployées par l'*USGS*). La survenue d'une crise naturelle ne constitue pas un catalyseur qui va modifier les structures fondamentales de l'activité virtuelle géolocalisée par GPS existant en temps normal, c'est-à-dire la tendance à l'hyperconcentration des tweets en milieu urbain et métropolitain. En effet, l'étude de l'événement virtuel enregistré entre le 16 et le 21 avril 2016 témoignait d'une concentration des tweets de crise géolocalisés sur un territoire restreint : 96% des mailles enregistrant des précipitations et agrégées sur dix kilomètres étaient vides de tweets de crise géolocalisés; et parmi les mailles restantes qui contenaient des tweets, 75% d'entre elles en comptaient moins de douze (cf. section 5.2.1.1). De même, pendant l'ouragan Harvey, on retrouvait en moyenne 76% de tweets émis entre le 23 août et le 31 août 2017 dans les aires urbaines et parmi ces tweets urbains, une moyenne de 74% s'avéraient en fait métropolitains (cf. tableau 5.4, et seules Houston et Austin étaient prises en compte dans les métropoles).

Si l'on appréhende alors l'événement virtuel à une échelle locale (comme la métropole et l'aire urbaine) et qu'on le compare à des données spatialisant les phénomènes réels, il apparaît rapidement une déconnexion entre les lieux de crise marqués par les données physiques et les lieux de réactivité virtuelle. En fait, les lieux d'activité virtuelle de crise constituent le miroir de l'activité tweeting normale ; en d'autres termes, les lieux dans lesquels on observe une activité virtuelle en situation normale sont plus enclins à être inscrits virtuellement en situation de crise. Ainsi, à Houston, tout comme à San Antonio, nous avons souligné l'existence d'un comportement identique : l'ensemble des *census tracts* actifs pendant la crise d'avril 2016 l'étaient également en temps normal (cf. tableau 6.3). En outre, si l'on considère les trois aires urbaines situées autour du *Lake Conroe*, on observait une

quantité de tweets de crise géolocalisés quasi négligeable[4], concentrée dans une seule aire urbaine, et ce malgré l'importance des dégâts enregistrés par la *FEMA* ; en situation normale, il s'avèrait que l'activité virtuelle géolocalisée représentait un poids tout aussi négligeable (cf. section 5.2.2.4).

A une échelle fine (les lieux de l'aire urbaine ou de la métropole), on ne peut pas affirmer l'existence systématique d'une logique spatiale de réponse virtuelle à la survenue d'un phénomène local identifié par des marqueurs officiels ou par des données témoins externes au réseau. En analysant la réponse virtuelle à l'émission d'un tweet marqueur d'alerte par le *NWS* ou l'*USGS* en milieu métropolitain, nous n'avons pas constaté l'existence d'une réactivité proche du marqueur en question : à Houston, en avril 2016, les tweets les plus proches de ce type de marqueur se localisaient à quelques centaines de mètres mais n'en constituaient manifestement pas des réponses directes (étant donné le décalage temporel de leur émission, qui se comptait en jour, cf. tableau 6.1). Si l'on se réfère ensuite aux données de la *FEMA* concernant les habitations endommagées par le passage de l'ouragan Harvey, la tendance générale reste identique : en dehors des milieux métropolitains, à Lake Charles (Louisiane), tout comme à Port Arthur-Bridge City (Texas), les territoires concentrant les plus fortes densités d'habitations ayant subi les dégâts les plus lourds restent peu inscrits sur le réseau, voire quasiment absents (cf. section 5.2.2.4).

Par ailleurs, on retrouve, dans les territoires de la métropole, cette tendance à l'hyperconcentration des tweets de crise géolocalisés par une minorité d'individus statistiques : à Houston, entre le 16 et le 21 avril 2016, 5,6% des *census tracts* cumulaient 50% des tweets de crise géolocalisés émis dans l'aire métropolitaine (cf. section 6.1.2.1). Dans le cadre de l'activité consécutive à l'ouragan, l'algorithme DBSCAN fournissait des résultats similaires : quelques îlots restreints de fortes densités de tweets et, à leurs côtés, une activité éparse diffusée dans l'ensemble de la métropole (cf. figure 6.47). En revanche, si l'on agrège cette information éparse pour en mettre en évidence le sens, elle se positionne parfois comme une réactivité directe aux phénomènes et événements locaux du réel (intensité des pluies, niveau des inondations, signalement de dégâts, adoption d'un comportement particulier, *etc*.), le problème restant que ce type d'information locale est très irrégulièrement mis à jour : en conséquence, on ne peut pas assurer le suivi d'un phénomène ou événement local exclusivement à partir des tweets de crise géolocalisés par GPS (cf. section 6.2.2.3).

[4] Et pour rappel, parmi les quatre territoires explorés situés en dehors des aires métropolitaines, les aires urbaines du *Lake Conroe* présentaient le plus faible ratio entre le nombre de tweets de crise géolocalisés et le nombre de bâtiments endommagés inventoriés par la FEMA.

Pour conclure sur ce deuxième point, le tweet de crise géolocalisé par GPS n'apparaît pas comme un témoin spatial systématique de la survenue des phénomènes et événements du réel ; le fait de tweeter pendant une crise d'origine naturelle proviendrait alors davantage d'un réajustement sémantique des lieux de l'activité habituelle sur la crise en cours. C'est pourquoi la localisation et les effectifs de tweets de crise géolocalisés émis en réponse à une crise, sont, à l'échelle de la métropole, de mauvais indicateurs d'intensité des phénomènes : une forte agitation en un lieu précis ne signifie pas un phénomène ou événement intense.

- *Le temps et la sémantique comme marqueurs significatifs avant la localisation et le volume de tweets*

L'existence des dynamiques consécutives à la réactivité aux phénomènes et événements du réel semble davantage perceptible par les pulsations temporelles et par la variabilité du contenu sémantique : à l'échelle d'une journée, qu'il s'agisse du phénomène rare ou des phénomènes récurrents, cette réactivité se manifeste par l'occurrence de thèmes successifs et perceptibles, qu'on soit en milieu métropolitain ou urbain, et que nous avons classés en mesures d'anticipation, alertes, description de phénomènes physiques, mesures de sauvegarde ou de soutien aux sinistrés, ainsi que mesures de résilience (cf. sections 5.2.1.3 et 5.2.2.4). Sur le plan de la réactivité temporelle, on a observé, par les effectifs de tweets émis, l'absence de réactivité systématique face à l'émission de tweets d'alerte mais en revanche, dans le cas de l'ouragan, l'accroissement général des flux de tweets de crise géolocalisés concomitants à la montée des eaux, dans une temporalité fine (si l'on considérait les mêmes tweets agrégés par heure dans une journée entière, les dynamiques d'émission semblaient davantage refléter les temporalités d'émission normales, cf. sections 6.2.1 et 6.2.2.1).

Le premier marqueur significatif qu'on peut avancer correspond à la soudaineté de l'activité virtuelle qui peut indiquer un phénomène ou un événement local subit. Nous avons envisagé le critère de soudaineté d'un événement virtuel selon deux paramètres : d'une part, l'apparition brusque d'un pic d'activité isolé temporellement par des périodes moins actives (cf. figure 6.34) ainsi que la persistance plus ou moins longue de mailles actives lors du passage de l'ouragan (cf. figure 5.52). Dans les deux cas, on distingue des éléments significatifs comme des éléments parasites : l'activation de mailles suite à un événement virtuel local provoqué par le retour des habitants dans leur quartier (et l'émission de tweets décrivant les dégâts identifiés ou l'ampleur des inondations, comme à Kingwood, cf. figure 5.57), ou au signalement direct de phénomènes ou événements détectés à une résolution temporelle fine (intempéries en cours, montée des eaux dans un quartier, routes inondées fermées à la circulation, cf. figures 6.53 à 6.56). Mais on identifie également des comptes automatiques uniques responsables de l'activation d'une maille (cf. tableau 5.7) ou encore des messages ambigus multi-sites émis dans une temporalité précise puis éteints définitivement (cf. figure

6.46). D'autre part, le second paramètre envisagé correspond à la disparition soudaine, temporaire ou définitive, d'un territoire dans la métropole ou l'aire urbaine (soit la dimension du silence virtuel introduite par [Samuels *et al.,* 2018]). Et en effet, on avait déjà constaté que les territoires qui concentrent de fortes densités de propriétés endommagées ne sont pas les plus visibles pendant la période de crise, lien qui a tendance à être confirmé par le test statistique effectué entre le degré de l'activité virtuelle pendant la crise et l'intensité des dégâts recensés par la FEMA (cf. figure 6.66).

Le second marqueur de significativité relevé correspond à la composante sémantique. Dans un premier temps, force est de constater que les nuages constitués d'associations lexicales affichent deux tendances :

- la présence prépondérante de messages impersonnels émanant d'acteurs officiels ou à consonance médiatique ;
- quand l'information tweetée semble plus personnelle, nous sommes confrontés au problème du micro-événement individuel, c'est-à-dire au tweet d'un utilisateur dont le contenu n'est par directement associé à un vécu particulier, le cas typique de ce tweet correspondant au message de prières. Au final, le nuage constitué de telles associations lexicales apparaît comme un écheveau d'informations auxquelles il est difficile d'associer un sens particulier (cf. figure 6.49 et 6.50).

En revanche, le sens de la sémantique se trouve davantage perceptible lorsque, au lieu de nous concentrer sur la représentation des associations lexicales, nous focalisons cette même représentation sur les différentes catégories grammaticales des mots, et en particulier les adjectifs et les verbes (cf. figures 5.59, 5.61 et 6.72). Pendant l'ouragan, si l'on comparait ce vocabulaire en fonction de différents territoires, on observait une variabilité spatiale certaine : des verbes focalisés sur les dégâts à Rockport (*hit*, *destroyed*) alors que les tendances de Port Arthur s'articulaient exclusivement sur l'urgence (*need*, *getting*, *stay*). Enfin, le second comportement sémantique qu'on peut considérer comme témoin correspond à l'hétérogénéité lexicale dans un territoire exploré (cf. figures 6.53 à 6.56) : dans le *CBD* de Houston, on a vu que l'activité virtuelle se montrait hétérogène dans ses thèmes alors que l'événement virtuel de Port Arthur était exclusivement focalisé sur la thématique des inondations et des interventions de secours ; de là, on peut émettre l'hypothèse selon laquelle, lorsque l'événement virtuel ne se réduit pas à une poignée de tweets épars, l'homogénéité sémantique laisse entendre une situation locale plus grave que la mention de thèmes variés en un lieu unique, dans un temps donné.

Pour conclure sur la question de la visibilité temporelle et de la sémantique, nous nous référons à l'exemple d'un territoire en marge de la métropole de Houston, qui avait alors attiré notre attention (et dont l'exploration résume l'intérêt de l'étude des composantes temporelle et sémantique avant le spatial et le quantitatif) : le quartier de l'extrême sud-ouest de l'aire métropolitaine, Sugarland-Rosenberg, qui alternait, pendant l'ouragan, entre des périodes d'activité et des périodes d'invisibilité virtuelle. Pour rappel, ce quartier est structuré comme suit : au nord-ouest, des lieux aux populations moins favorisées et classées plus vulnérables ; au sud-est et à l'est, des lieux aux populations plus aisées et classées moins vulnérables. L'indice statistique alors créé pour atténuer l'effet de l'activité virtuelle normale mettait en évidence une structure miroir de cette logique socio-spatiale, en situation de crise : les lieux aux populations vulnérables témoignaient d'une activité inférieure à la normale alors que la plupart des lieux aux populations moins vulnérables s'activaient en temps de crise.

L'analyse lexicale par catégories grammaticales révélait ainsi une spatialisation des verbes et adjectifs, associée à ces paramètres descriptifs des lieux et de leurs habitants (cf. figure 6.72) : les mots comme *affected*, *help*, *need*, *flooding* se trouvaient dans ces territoires vulnérables du nord-ouest alors que le vocabulaire du sud-est s'affichait beaucoup moins dans l'urgence (*fortunate*, *fine*, *safe*, *stay strong*, *praise*). Au final, la composante temporelle, à une résolution fine, a permis d'identifier un lieu d'intérêt par l'alternance des périodes d'activité et de silence ; dans un second temps, la spatialisation du vocabulaire par catégories grammaticales significatives indique l'existence d'une logique spatiale qui concorde avec des données externes au virtuel, en l'occurrence l'indice de vulnérabilité sociale.

- *Le tweet et les données officielles externes : une complémentarité restreinte*

Dans l'introduction du manuscrit, nous avions émis l'hypothèse selon laquelle le tweet de crise géolocalisé ne pouvait s'autosuffire dans la mesure où la trace numérique géolocalisée était fournie au chercheur en étant détachée de tout contexte social et environnemental de production ; nous avons alors positionné le recours aux jeux de données externes et officiels comme solution pour pallier ce défaut d'a-contextualisation.

Contre toute attente, et à défaut de révéler le potentiel des tweets de crise géolocalisés, le recoupement entre les diverses sources de données a principalement mis en évidence les problèmes supplémentaires qu'implique l'adoption d'une approche géographique dans l'étude des tweets de crise géolocalisés. En fait, on ne met pas systématiquement en évidence d'articulation intuitive et répétitive entre les modalités de l'activité virtuelle et les structures spatiales constatées en ayant recours aux données externes officielles :

- même si, depuis les cartes, se dégageait l'impression générale d'une moindre représentation des territoires concentrant des minorités ethniques, ou des populations précaires moins diplômées, nous n'avons pas pu mettre en évidence (par les outils statistiques multivariés), ni pour Houston, ni pour San Antonio, l'existence de corrélations positives ou négatives entre de telles variables socio-économiques et l'activité tweeting géolocalisée (qui apparaît comme une variable statistique non significative, cf. section 6.1.3). Tout au plus, nous avons pu montrer l'existence d'une relation entre le profil des populations (urbain/rural) et la contribution à la création de tweets de crise géolocalisés (cf. section 5.2.1.2). En conséquence, sur le terrain d'étude, on ne sait pas quels individus sont à l'origine des traces qu'on analyse.

- si l'on compare tweets de crise géolocalisés et données environnementales, il s'avère que les foyers de l'activité virtuelle de crise ne constituent pas le miroir de l'intensité des phénomènes du réel (ce n'est pas dans les territoires qui subissent les plus forts cumuls pluviométriques qu'on trouve une activité significative, cf. sections 5.2.1.3 et 6.1.1), ni des événements qui leur sont consécutifs (les territoires qui concentrent de fortes densités de dommages inventoriés ne sont pas régulièrement visibles pendant la période de crise ; ils peuvent même être quasiment absents du réseau géolocalisé, cf. figures 6.63 et 6.65).

En fait, il semble que, dans un premier temps, l'événement virtuel peut se suffire à lui-même afin d'identifier des lieux d'intérêt : par la création d'indices, on peut mettre en exergue les mailles à l'activité virtuelle de crise inexistante ou élévée (autrement dit il s'agit des mailles dans lesquelles on constate la présence exclusive de tweets de crise, ou dont cette même activité virtuelle est supérieure ou égale à l'activité géolocalisée normale). Il s'avère que ces mailles d'intérêt correspondent en fait à ces territoires de marges virtuelles (cf. figure 6.64), qu'une simple représentation quantitative ou statistique (discriminer les mailles selon qu'elles contiennent un nombre inférieur ou supérieur à une valeur de référence, comme la médiane) des tweets ne permet pas de mettre en évidence (cf. figure 6.57). C'était le cas du quartier de Sugarland-Rosenberg de Houston, dans lequel on observait des discontinuités dans le degré de l'activité virtuelle de crise : l'appel aux données externes représentant l'indice de vulnérabilité sociale permettait ensuite de constater l'existence de lieux aux populations plus vulnérables ; comme indiqué dans le point précédent, la variabilité spatiale de la sémantique était alors significative de la variabilité du profil socio-spatial et virtuel des territoires et de leurs populations. Il ne s'agit en revanche que d'un unique test et il conviendrait de le répéter sur d'autres marges de la métropole pour constater l'éventuelle répétitivité de ce comportement.

Pour conclure sur le point du recours aux données externes afin de recontextualiser l'activité virtuelle de crise, et en considérant les décalages observés entre propriétés de l'événement virtuel et structures perçues dans les données externes, il conviendrait de séparer les traces des données : dans un premier temps, l'événement virtuel est étudié comme un objet à part entière (dans la mesure où les facteurs environnementaux ou sociaux ne sont pas systématiquement explicatifs des modalités de l'événement virtuel) puis la connaissance spatialisée produite à partir de cet objet est recoupée aux éléments environnementaux. L'une des questions qu'on peut alors soumettre est la suivante : peut-on modéliser, et donc prédire, à une résolution spatiale et temporelle fine, un comportement virtuel particulier (la présence d'un certain vocabulaire, les phases de silence ou d'activité d'un lieu, le passage d'un lieu dans différentes phases) en fonction des éléments environnementaux ?

Enfin, pour répondre à la question soulevée au début de la présentation de cette contribution thématique (*La recherche considère-t-elle à tort ou à raison le tweet de crise géolocalisé par GPS comme un témoin pertinent des perturbations environnementales et de leurs effets sur les territoires et populations ?*), un événement virtuel constitué de tweets de crise géolocalisés par GPS contient des témoins incontestables de phénomènes et événements en cours, émis par des individus capteurs subissant la crise, mais ceux-ci sont irréguliers (spatialement et temporellement) et noyés dans les tweets à consonance officielle et les tweets que nous avons qualifiés de micro-événements individuels. Ces témoins, même s'ils sont épars et ne constituent que les bribes d'un récit, peuvent être valorisés par la carte dans la mesure où un tweet émis par un individu capteur réagit à l'environnement immédiat en fournissant des indices qualitatifs qui permettent de hiérarchiser l'information (à l'image de ces tweets que nous avons mentionnés et qui indiquaient qu'un utilisateur avait de l'eau jusqu'aux genoux ou qu'un autre considérait le niveau de l'inondation comme "normal" au moment où il tweetait). En outre, si l'on considère l'utilisation des tweets de crise géolocalisés par GPS dans une situation de gestion de crise, nous avons vu que la présence d'une rupture spatiale du degré d'activité virtuelle de crise pouvait se montrer significative sur les territoires de marges virtuelles.

En fait, nous pensons désormais, à l'issue de ce travail, qu'il est nécessaire de nous détacher de l'approche qui a systématiquement été adoptée pendant les dix années écoulées vis-à-vis des tweets de crise géolocalisés : sélection des tweets, approche de traitement fondée sur les effectifs de tweets, ou encore recours systématique aux données externes pendant les étapes de l'analyse des tweets[5].

[5] Les réflexions à ce propos sont présentées dans la dernière partie de la conclusion "Orientation des perspectives de la recherche".

Contribution n°3 - Proposition de démarches méthodologiques à cibler dans l'analyse des tweets géolocalisés

Comme souligné dans l'introduction du manuscrit, notre objectif ne consistait pas à proposer une solution clé en main d'analyse des tweets de crise géolocalisés, c'est-à-dire un outillage logiciel conçu sous la forme d'un environnement de géovisualisation. Il s'agissait de distinguer, au préalable, les méthodes qui fourniraient des résultats exploitables de celles qui conduiraient éventuellement à des impasses. Nous proposons maintenant un bilan de ces outils et démarches méthodologiques de conduite d'expériences :

- *Selon une approche spatiale*

Au regard du maillage spatial irrégulier formé par un semis de tweets géolocalisés, il faut considérer les faits suivants dans le choix des outils d'agrégation et de représentation cartographique. En premier lieu, l'algorithme DBSCAN, appliqué à une échelle spatiale fine, se montre beaucoup trop sélectif : pour rappel, lors des tests effectués sur les tweets de crise émis en réponse à l'ouragan, environ 50% des points étaient considérés comme du bruit résiduel (et donc en théorie écartés des analyses ultérieures à la partition spatiale, cf. figure 6.47). Dans un second temps, nous avons vu que les effectifs de tweets géolocalisés inclus dans une entité spatiale étaient un élément trompeur en ce qui concerne le rapport au réel. En conséquence, pour la question de la représentation et de l'étude des dynamiques virtuelles, nous préconisons la définition de mailles d'agrégation spatio-temporelle des tweets, restituées selon deux cartographies envisageables : l'activité (la maille contient au moins un tweet dans le temps donné) ou le silence (la maille ne contient aucun tweet) d'une part, et le recours aux paramètres statistiques d'autre part. Les mailles peuvent alors être représentées soit par un indice mesurant le degré d'une activité virtuelle de crise en fonction d'une valeur de référence, soit par des paramètres statistiques univariés[6] (par exemple, les mailles dont les effectifs sont supérieurs/inférieurs aux différents quartiles).

- *Selon une approche sémantique*

Comment explorer et représenter la sémantique des ces mailles afin de les comparer ? Nous avons testé la LDA afin de dégager l'éventuelle existence de *topics* précis sur un territoire en crise ; les résultats ne se sont pas montrés pertinents : le contenu lexical des tweets de crise se montre trop hétérogène (rien ne garantit systématiquement que deux tweets émis dans un environnement proche se ressemblent sémantiquement). Par ailleurs, si cette méthode est appliquée dans une maille ne contenant qu'une poignée de tweets de crise, le matériau fourni à l'algorithme sera quantitativement insuffisant pour donner des résultats exploitables.

[6] Mais en excluant la moyenne qui se trouve beaucoup trop influencée par les valeurs extrêmes fortes.

Dans un deuxième temps, comme indiqué précédemment dans la contribution n°2, le nuage de mots contruit à partir d'associations lexicales se montre fréquemment peu significatif ; en revanche, les nuages fondés sur les catégories grammaticales que nous avons mises en évidence (verbes et adjectifs) semblent apporter des renseignements spatiaux locaux et comparables (cf. figures 5.59, 5.61 et 6.72). De tels nuages catégorisés construits à partir des mailles peuvent alors être complétés par des matrices permettant de mesurer la similarité lexicale entre différentes mailles localisées sur la carte.

En ce qui concerne la représentation de la sémantique sur la carte, nous avons testé deux approches complémentaires, à partir d'une même base, c'est-à-dire la classification des tweets par thèmes :

- nous avons défini des icônes pour représenter chaque thème défini et utilisé le principe de la lecture du nuage de mots : dans chaque agrégat de tweets de crise géolocalisés est représenté l'ensemble des thèmes identifiés dans les tweets ; plus le thème est marqué, plus le symbole est grand (NB : si un tweet contient plusieurs thèmes, alors l'ensemble des icônes correspondantes sont affichées sur la carte). Cette représentation offre une lecture rapide de l'agrégat mais masque l'éventuelle diversité lexicale (cf. figures 5.62 et 5.63).
- Nous avons également conservé le nuage d'associations lexicales, mais colorées en fonction des thèmes identifiés ; cette représentation offre l'avantage de voir rapidement s'il existe une thématique dominante sur le territoire global ou dans un lieu précis, ainsi que la possibilité d'identifier la survenue soudaine d'un thème particulier (cf. figure 6.53) et en quels lieux (ou au contraire, elle peut témoigner de l'existence d'un brouhaha virtuel général : divers thèmes qui apparaissent en tous lieux et dans une même temporalité [cf. figure 6.55]). En revanche, en cas d'agitation soudaine, une telle représentation aboutit rapidement à une surcharge visuelle de la carte (cf. figure 6.62).

- *Et selon quelle approche épistémologique ?*

L'idée générale de la démarche épistémologique de cette recherche était la suivante (cf. section 4.1.3.1 et figure 4.1) : (1) on propose une première exploration des traces ; (2) on observe les structures qui en ressortent ; (3) on essaye de distinguer les facteurs qui peuvent être à l'origine des constats établis ; (4) on teste la validité de ces facteurs ; (5) si le test est concluant, la théorie peut être formulée puis on teste sa validité dans une nouvelle expérience ; dans le cas contraire, on explore d'autres possibilités. Cette démarche a été mise en œuvre, telle que décrite ci-avant, de la manière suivante (deux exemples sont donnés) :

- dans les sections 6.1.1 et 6.1.2 : en cartographiant les émissions de tweets de crise géolocalisés par rapport aux témoins de précipitations intenses ou aux données externes pluviométriques (1), on ne trouve pas de logique spatiale et quantitative virtuelle qui serait la réponse directe aux perturbations environnementales (2). On compare alors l'activité virtuelle de crise à l'activité virtuelle normale (3), qui semble davantage se référer à une logique de centre-périphéries. On effectue le premier test (pour rappel, il s'agissait alors de San Antonio)

qui se montre concluant (4). On répète le test dans un autre lieu, à Houston qui donne des résultats plus mitigés ; en fait, il s'avère même qu'en situation de crise, on observe une indépendance spatiale entre les entités considérées (5). En conséquence, la théorie formulée dans une ville A n'est pas généralisable à la ville B ; de même que la logique spatiale observée dans un contexte donné n'est pas transposable à un autre contexte.

- Dans les sections 6.3.1 et 6.3.2 : d'une première série d'expériences (menées dans le chapitre 5), on a conclu que les dimensions spatiales et quantitatives seules d'un événement virtuel n'étaient pas significatives par rapport aux phénomènes et événements du réel. Dans une nouvelle représentation cartographique, on crée des indices pour prendre en compte le biais de l'activité habituelle des entités spatiales (1). La carte résultante met en évidence les marges virtuelles dans lesquelles on avait constaté, dans une série de cartes précédentes (en l'occurrence, l'exploration sémantique dans un pas de quinze minutes [cf. figure 6.53 à 6.56 et 6.60 à 6.62]), des éléments particuliers : des périodes alternées de silence et d'activité, l'existence d'un vocabulaire spécifique, *etc.* (2). On compare les structures identifiées avec une source de données externes officielles, l'indice de vulnérabilité sociale : on constate une activité inférieure à la normale dans des entités plus vulnérables (3). On teste alors la possibilité de dégager un sens sémantique dans les entités identifiées en appliquant la méthode de représentation lexicale qui fournissait des résultats discriminants (les catégories grammaticales par nuages de mots) (4). Dans le quartier de Sugarland-Rosenberg, la sémantique prend un sens spatial particulier, semblant ainsi indiquer que les marges métropolitaines vulnérables sont quantitativement moins actives mais contiennent un vocabulaire spécifique d'une situation de crise alors que les lieux voisins, moins vulnérables, témoignent d'un contenu plus diversifié et davantage apparenté au micro-événement individuel (5), (cf. figure 6.72).

L'ensemble des remarques présentées ci-dessus sont destinées à fixer un cadre méthodologique et technique pour l'éventuel développement d'un environnement de géovisualisation. En effet, l'inconvénient majeur de l'approche épistémologique testée réside dans le fait qu'elle nécessite qu'on s'attache aux détails identifiés et qu'on effectue de multiples tests afin de savoir si un comportement virtuel identifié dans un lieu précis peut être observé dans un autre lieu, ou si chaque lieu adopte son propre comportement (auquel cas, aucune théorie ne serait alors généralisable). Mise en œuvre par la combinaison des outils bases de données/SIG/outil de traitement statistique et sémantique de l'information (R), et donc non automatisée, elle reste longue et fastidieuse à appliquer ; d'autant plus que certains constats indiqués dans la contribution n°2 (et notamment sur les comportements sémantiques et temporels) peuvent être considérés comme des hypothèses posées en réponse à l'ensemble des analyses effectuées dans cette recherche, qui mériteraient d'être testées en ayant recours à des traces plus récentes, dans d'autres lieux : par exemple, pourrait-on observer des résultats identiques (et donc supposer l'existence des hypothèses formulées) avec les ouragans des saisons récentes, comme Michael (octobre 2018) et Dorian (août-septembre 2019) ?

Pour conclure la section des contributions de ce travail, nous pourrons les résumer selon deux points :

- une méthodologie reproductible de recherche et d'extraction de tweets utiles qu'on pourra étendre à d'autres contextes événementiels, et notamment l'approche spatiale (en identifiant des objets ou lieux du territoire et en observant comment les structures spatiales et le discours évoluent autour de ces objets en fonction du temps ;
- une méthodologie exploratoire reproductible dont les apports et les pistes valorisables ont été discutés ci-avant ; étant donné que certaines conclusions de la recherche demeurent à l'état d'hypothèses et en adéquation avec la méthodologie proposée, il conviendrait d'effectuer de nouveaux tests à partir d'autres phénomènes afin de savoir si l'on peut identifier les comportements virtuels mis en évidence sur notre terrain d'étude à d'autres lieux (***autrement dit, est-il possible de généraliser, en loi, un comportement virtuel formalisé par une hypothèse, ou le comportement observé est-il spécifique d'un lieu ?***).

Orientation des perspectives de la recherche

Au regard des contributions détaillées dans le paragraphe précédent, nous soumettons, pour terminer, les pistes qu'il conviendrait selon nous d'appréhender en vue de la définition d'une géographie fondée sur les traces numériques géolocalisées (considérées dans leur ensemble). Nous envisageons ces pistes en fonction de quatre points :

- *Approche méthodologique spécifique aux tweets géolocalisés par GPS*

Quels tweets géolocalisés utiliser dans les analyses ? Si l'on considère les tweets de crise géolocalisés par GPS dans un premier temps, on s'est rapidement aperçu des tendances suivantes :

- de nombreuses émissions concentrées par des acteurs officiels et leurs conséquences :
- une poignée de comptes hyperactifs alors que les individus qui tweetent à titre privé n'exercent qu'une activité sporadique ;
- une sur-représentation de l'information à consonance officielle ;
- une déconnexion entre lieux violemment affectés et foyers d'activité virtuelle de crise.

En conséquence, l'information focalisée sur les activités ou le ressenti de l'utilisateur dans son environnement perturbé, et dont le contenu sémantique pourrait être valorisé dans la cartographie, reste très marginale en termes de poids.

D'après les recherches bibliographiques (Dashti *et al*., 2014 ; Hertfort *et al*., 2014 ; Saravanou *et al*., 2015 ; Steiger *et al*., 2015), nous sommes partis du postulat selon lequel l'analyse d'un phénomène naturel à partir de tweets géolocalisés ne peut pas s'effectuer sans

cette primo-étape de filtrage lexical aboutissant à la création d'un corpus de tweets de crise. Mais ce système introduit sans doute un biais quand bien même l'on adopte une approche non-supervisée impliquant une fouille lexicale des tweets. En effet, nous avions vu que la fouille de texte mettait en exergue les tendances majoritaires contenues dans les tweets, celles-ci correspondant en fait à des tweets à consonance officielle. Le tweet de l'individu peut, quant à lui, être mis en évidence par la recherche des collocations éparses ; pour autant, en raison de la grande hétérogénéité de ces tweets émanant des individus, et si l'on considère les quelques milliers voire dizaines de milliers de tweets à fouiller, on risque de pas souligner certains types de contenus individuels.

En revanche, on avait vu que des cellules d'activité virtuelle s'activaient ponctuellement pendant la période de crise (on pouvait d'ailleurs les détecter précisément en repérant un degré d'activité virtuelle inhabituelle ou inexistante par un indice statistique) ; d'autre part, en temps de crise et quel que soit le milieu considéré, on observait un cadrage du contenu sémantique de l'événement virtuel en fonction du réel. En conséquence, la nouvelle priorité à donner aux étapes de primo-analyse est, selon nous, focalisée sur deux axes :

- un fractionnement préalable du territoire en cellules afin d'observer l'évolution de l'activité virtuelle dans un contexte perturbé : les mailles sont représentées en fonction de leur activité ou de leur silence selon un pas de temps défini. L'exploration sémantique des mailles peut être envisagée à partir de l'analyse grammaticale des tweets, en portant attention aux éventuels changements de tonalité. Ici, ce n'est donc pas le nombre de tweets qui est indicateur de perturbations localisées mais la dimension qualitative des tweets. Il s'agit alors d'explorer les modalités selon lesquelles l'activité virtuelle se recentre (ou pas) spatialement et lexicalement, quand survient une perturbation locale.
- quand nous avons pratiqué le test d'extraction spatiale sur la crue de la Seine, nous avons identifié un certain nombre d'objets ou de lieux polarisant l'activité virtuelle de crise. De la même manière, nous avions identifié de tels objets à San Antonio. L'approche spatiale pourrait ainsi être recentrée sur l'identification préalable de ces pôles d'activité et l'observation, sans filtrage sémantique, de l'évolution spatiale de l'activité virtuelle globale autour de ces pôles, en période de crise.

Au final, la primo-étape de nettoyage qu'il conviendrait d'appliquer consisterait à supprimer les tweets émis par ces utilisateurs au profil hors-normes, en s'assurant qu'il s'agit de comptes institutionnels ou médiatiques (et ce, afin de nous centrer exclusivement sur les tweets des individus).

Quelles perspectives cartographiques pour la valorisation des tweets géolocalisés ? Dans la première partie du manuscrit, nous avions signalé les questions méthodologiques suivantes par rapport à la place des tweets géolocalisés par GPS dans les données et l'information géographique : comment considérer cette trace numérique dans l'information géographique et est-il possible de la transformer, en l'associant aux données externes officielles, en

information géographique cartographiable ? Si oui, à partir de quelles composantes des traces ?

La réponse à cette question ne s'impose pas de manière évidente ; pour l'heure, on ne propose de valoriser le tweet qu'en donnée transformée et en ne représentant que ses composantes propres (par exemple : un agrégat spatio-temporel localisé ou une maille à l'activité anormale dont on représente la sémantique sur la carte, soit sous forme de nuages de mots, soit sous forme de symboles). Les données externes restent des éléments de recontextualisation des émissions qui ne sont pas intégrés dans la représentation spatiale des tweets.

En fait, le cœur du problème de la transformation des tweets en information géographique précisément délimitée intégrant éventuellement les données externes de contexte est le suivant : comment prétendre délimiter des territoires supposés avoir un comportement spatialement homogène quand on considère les faits suivants mis en évidence dans la recherche ?

- La structure de l'activité virtuelle n'est pas un échantillon spatial : elle est irrégulière en temps normal comme en temps de crise. En conséquence, certains lieux sont sur-représentés alors que d'autres sont invisibles ;
- on ne parvient pas à modéliser des facteurs explicatifs des lieux de l'activité virtuelle qui s'avèrent généralisables d'un territoire à l'autre, ou d'un contexte à l'autre ;
- on peut présupposer l'existence de biais socio-spatiaux qui se répercutent sur la volonté et les possibilités techniques et financières d'un individu à contribuer régulièrement à la production de tweets géolocalisés mais on ne parvient pas précisément à déterminer les profils des populations qui tweetent avec le GPS ou qui sont exclues (ou s'excluent) de cette activité virtuelle.
- enfin, en raison de la présence de ces micro-événements individuels, la première loi de Tobler n'est pas systématiquement vérifiée (deux tweets voisins émis dans une temporalité proche ne portent pas nécessairement un discours à tonalité identique).

Quand, dans le chapitre 1, on présentait la carte de vigilance météorologique des départements français métropolitains, il ne figurait aucune entité dans laquelle on n'avait pas d'information qualitative (en revanche, celle-ci reste générale et n'indique pas les éventuels phénomènes locaux violents). Doit-on adopter la même démarche avec les tweets de crise géolocalisés ? Autrement dit, convient-il d'agréger ces tweets dans des entités polygonales de telle manière qu'aucun lieu d'une métropole n'apparaisse vide de tweets ? et est-il approprié de considérer que le témoin unique de la maille, ou le discours majoritaire identifié si la maille contient plusieurs tweets, est représentatif de l'ensemble des situations locales envisageables dans une maille ?

A ce stade, nous envisageons plutôt de conserver un maillage régulier (en identifiant les mailles invisibles) et d'appréhender trois pistes de représentation et de mesure de l'activité

dans la maille par rapport à son contexte socio-environnemental. En premier lieu, dans les mailles enregistrant une activité virtuelle, l'objectif consisterait à associer un score local à la maille en fonction de l'intensité ressentie (ou de la phase dans laquelle se situe la maille pendant la crise). L'attribution de ce score serait fondé sur la présence de catégories grammaticales précises et de mots de vocabulaire marqueurs d'intensité ou du phasage de l'événement virtuel, que nous avons identifiés ici (*pray*, *staystrong* [pour les micro-événements individuels], *need*, *help*, *rescue*, *boats*, *hard*, *hit*, *destoyed* [dans les territoires où la situation semblait plus critique]). L'objectif consisterait ainsi à construire non plus des glossaires d'extraction de tweets de crise, mais un glossaire de mots marqueurs auxquels on attribue un score positif ou négatif en fonction de leur contenu, et ce afin de hiérarchiser les mailles en fonction du score obtenu par la sémantique des tweets. En revanche, le bon déroulement de cette étape nécessite une étude préalable spatiale et sémantique complète et approfondie de l'ensemble du terrain d'étude, les cartographies restituées dans le chapitre 6 indiquant que le lexique employé varie d'un lieu à l'autre. En outre, pour attribuer un score à une unité en gommant l'effet du maillage irrégulier, on peut envisager deux possibilités :

- prendre en compte une seule fois la présence d'un mot qui se répéterait ;
- créer la typologie des profils d'activité des mailles en fonction des lieux et de leur activité, qu'on analysera séparément.

Dans un deuxième temps, il faudrait également évaluer le poids que représente l'information associée au contexte perturbé par rapport à l'ensemble des autres tweets qui peuvent graviter autour de cette information dans chaque maille. L'objectif est articulé autour de la caractérisation du signal : on peut représenter, sur la carte, la proportion de tweets incluant des mots marqueurs dans la maille et essayer de détecter, à un moment précis, une période de rupture, c'est-à-dire un recadrage exclusif de l'activité virtuelle sur l'environnement perturbé ou encore la survenue d'un silence. Enfin, après observation du comportement virtuel spatio-temporel et sémantique, il reste à intégrer les données externes, afin d'explorer les liens éventuels entre la réactivité virtuelle et le contexte réel, et de savoir s'il est possible de créer la typologie généralisée des lieux de réactivité en fonction du contexte. Si l'on reprend les deux exemples évoqués ci-avant, on peut envisager :

- en considérant la possibilité de disposer de données pluviométriques ou hydrométriques (si la maille inclut une station météorologique ou de jaugeage de l'*USGS*), on peut créer un indice intégrant la proportion de tweets aux mots marqueurs d'une part, et la valeur des cumuls pluviométriques ou de la hauteur du cours d'eau d'autre part ;
- si l'on a hiérarchisé, en période de crise, les mailles en fonction d'un score calculé à partir des mots marqueurs, on peut les associer à l'indice de vulnérabilité sociale.

- *Intégration des autres formes des traces numériques et évolution des outils*

La recherche en géographie s'est sans doute trop rapidement focalisée sur le tweet géolocalisé par GPS (et par conséquent exclusivement émis par smartphone) pour son exploitabilité cartographique et spatiale directe ; mais le recours exclusif à un type de trace numérique pour ses propriétés avantageuses n'est-il pas, au fond, restrictif ? Pour rappel, en dehors des acteurs officiels, on ne connaît pas très précisément le profil ou les motivations des utilisateurs qui se géolocalisent ; de même, on ne sait pas si les résultats mis en évidence dans un cas d'étude peuvent se généraliser à d'autres territoires. Ensuite, les logiques de réactivité spatio-temporelle en temps réel ne sont pas évidentes à mettre en exergue car elles se trouvent bruitées par les tweets à consonance médiatique ou des comptes institutionnels qui constituent la majorité des tweets géolocalisés ; en conséquence, on ne sait pas si les tweets émis par les individus au cœur de la crise sont bien représentés dans les analyses et ainsi dans les résultats.

Aux côtés de ces constats, il faut encore considérer la dépendance de la recherche académique aux politiques des plateformes du Web et aux pratiques de leurs utilisateurs : le tweet géolocalisé par GPS est d'ailleurs peut-être un produit en train de disparaître. En effet, depuis juin 2019, Twitter a supprimé l'option d'ajout direct d'une localisation aux tweets émis (la géolocalisation d'un tweet reste toutefois possible mais il faut alors avoir recours à d'autres réseaux sociaux), en invoquant une popularité de la fonctionnalité bien inférieure à celle qu'avait escomptée l'entreprise[7].

Mais aux côtés de ces tweets géolocalisés émis par smartphone, une pléthore de contenus ont été délaissés par la recherche en géographie : il s'agit de ces tweets dont l'utilisateur inclut une mention sémantique de localisation (qu'on peut également retrouver dans les messages de Facebook ou encore dans les *tags* des photographies de Flickr ou d'Instagram). Ainsi, l'enjeu primordial qui encadre l'exploitation des traces numériques reste, à notre sens, le suivant : lorsque la géolocalisation directe est impossible, faut-il orienter la recherche vers une géographie où la mention du lieu et de ce que l'utilisateur pense, fait ou voit du lieu est plus importante que le marqueur de sa présence réelle dans le lieu dont il est question à travers la trace ? Doit-on donc, dans un premier temps, déconnecter l'analyse du territoire afin de souligner ce qui se dit ou se fait dans un lieu par les traces sémantiques laissées à son propos, pour, dans un second temps, représenter sur la carte (par des symboles ou des mots) ce qu'on a extrait du lieu dont il est question ?

En dernier lieu, même si nous avons commencé à évoquer ce fait dans le dernier point de la contribution n°3 (cf. *Et selon quelle approche épistémologique ?*), nous rappelons que l'approche traditionnelle de gestion et d'analyse des données par le triptyque gestionnaire de base de données/SIG/logiciels de traitements (autres que cartographiques et spatiaux) est

[7] Source : https://siecledigital.fr/2019/06/19/twitter-ne-permet-plus-de-partager-la-geolocalisation-exacte/ (Consulté pour la dernière fois le 25/11/2019)

désormais trop fastidieuse à mettre en œuvre face à la quantité et à la diversité des situations à explorer dans les traces numériques géolocalisées. C'est pourquoi nous avons proposé, conformément à l'un des objectifs que nous avions annoncés, de distinguer les outils spatiaux et lexicaux à intégrer dans une interface de géovisualisation, qui fournissent des résultats exploitables au regard des constats que nous avons établis après chaque analyse test. Comme en témoignait la version la plus récente de l'interface de *SensePlace3* (cf. [Pezanowski *et al*., 2017] dans le chapitre 3 du manuscrit), il n'y a pas nécessairement besoin d'une grande diversité ou complexité d'outils pour mettre en évidence des structures ; ainsi, dans la présentation des contributions, nous avons commencé à introduire les outils et approches, qui selon nous, permettent la mise en évidence de la variabilité spatio-temporelle et sémantique des événements virtuels. En conséquence, s'il fallait désormais spécifier le cahier des charges préparatoire au développement d'un environnement exploratoire automatisé et performant, celui-ci devrait intégrer *a minima* :

- au niveau spatial : la possibilité de créer des maillages spatiaux de différentes résolutions, d'agréger les tweets selon ces mailles, et de les visualiser en fonction de paramètres statistiques ou d'indices ;
- au niveau sémantique : la possibilité d'afficher des nuages de mots en fonction de catégories grammaticales précises, de représenter des nuages d'associations lexicales dont la couleur varie en fonction du thème ou encore de représenter le contenu des tweets par le recours aux symboles iconographiques, mais encore de générer des matrices de similarité lexicale en fonction de mailles sélectionnées.
- au niveau temporel : intégrer le graphique temporel des émissions avec la possibilité de modifier les dates et heures de sélection ainsi que la résolution temporelle d'affichage des tweets géolocalisés.

Enfin, si l'on considère également les tweets contenant une spatialisation sémantique, alors on sort du cadre régulier des mailles pour intégrer une dimension spatiale qui peut être multi-scalaire et multiforme (un territoire global, un lieu ou encore un objet du territoire) ; il est alors nécessaire d'intégrer une fonctionnalité de *geoparsing*[8] (soit un outil qui convertit une information spatiale sémantique en information spatiale géolocalisée), puis de géolocaliser le lieu dont il est question et de déterminer la forme (polygonale, linéaire ou ponctuelle) associée au discours.

- *Une géographie du numérique pour qui et pour quels territoires ?*

Au regard des résultats présentés dans cette recherche, les propos tweetés du géomaticien Anthony Robinson, cités dans l'introduction (*"The tragedy in Texas will magnify the fact that the most vulnerable ppl*[9] *are not going to tweet about it. Response can't be driven by tweets"*) s'avèrent bien fondés. En effet, d'après nos résultats, même si nous ne savons pas

[8] Source : https://en.wikipedia.org/wiki/Toponym_resolution
[9] Abréviation du mot *people*

précisément quelles populations sont (ou ne sont pas) représentées par l'activité virtuelle géolocalisée, nous avons néanmoins mis en évidence la persistance indiscutable de la fracture socio-spatiale du numérique, à travers deux comportements :

- des personnes ayant besoin d'aide pour être évacuées mais signalées par des utilisateurs tiers (Port Arthur pendant le passage d'Harvey) ;
- à toutes les échelles spatiales considérées et quel que soit le type de phénomène, des territoires frappés par les phénomènes naturels mais invisibles de l'événement virtuel.

Construire une connaissance des territoires par le numérique représente ainsi un risque de marginalisation des territoires, objets locaux ou populations non représentées dans la production de contenus. Et en effet, Web et réseaux sociaux apparaissent désormais comme les agents et garants de ce qu'on pourrait nommer la *visibilité numérique* sur laquelle s'appuie une pré-expérience du territoire réel (Valentin *et al*., 2011). A l'échelle d'un individu, cela signifie que l'appropriation d'un territoire méconnu et des objets qui le composent passe par une primo-connaissance virtuelle du terrain, acquise par l'intermédiaire des outils cartographiques du Web et des applications nomades ; en conséquence, l'un des effets pernicieux du numérique consiste en ce que tout objet ou lieu mal référencé dans l'espace virtuel prend le risque d'être invisible dans le réel pour une partie des individus. Or il s'avère que les chercheurs, lorsqu'ils définissent un terrain d'étude local, adoptent cette même pratique de sélection d'une cible en fonction de son inscription dans le réseau virtuel géolocalisé (en témoignent, dans l'état de l'art, la présence exclusive des métropoles dès lors qu'une étude ou un projet est ciblé sur une échelle géographique locale : Seattle, Chicago, Columbus, Bangkok, Jakarta, Londres, Manaus, Madrid, Paris, *etc*.). En conséquence, il est désormais nécessaire de soulever cette question : peut-on prétendre à une prise de décisions raisonnées sur la base d'une géographie du numérique construite à partir des pratiques actuelles qui s'avèrent sélectives de territoires et donc discriminantes ?

Dans le cas précis des dynamiques et réponses aux risques naturels, nous avons vu qu'un unique tweet pouvait parfois apporter plus de sens, à propos d'une situation locale, qu'un groupe de tweets apparentés à des micro-événements individuels. C'est pourquoi il nous paraît alors plus rationnel d'envisager une participation effective des individus capteurs à la production de traces utiles à la connaissance et au suivi des objets du territoire selon l'approche des sciences participatives, c'est-à-dire la participation active d'un public à un processus scientifique, dans la collecte des données et dans leur analyse (Paul *et al*., 2018). En d'autre termes, l'adoption de cette approche impliquerait qu'on se détache de la logique de visibilité numérique qui nous fait aujourd'hui cibler l'objet ou le lieu de recherche en fonction de son adaptation au numérique, au profit d'une logique de formation et de mobilisation d'individus comme base de connaissances d'un territoire local et de ses objets, dont la définition n'est pas contrainte pas cette présence virtuelle massive.

Bibliographie

Bibliographie

Aguiléra A., Belton-Chevallier L. (2017) "Mobilités et (R)évolutions numériques". *Netcom*, Volume 31, n°3-4, pp.275-280.

Alam F., Ofli F., Imran M., Aupetit M. (2018). "A Twitter tale of three Hurricanes: Harvey, Irma and Maria". *Proceedings of the 15th ISCRAM Conference, Rochester, USA,* mai 2018.

Allen D. McAleer M. (2018)."Fake news and indifference to scientific fact: President Trump's confused tweets on global warming, climate change and weather". *Scientometrics*, Vol. 117, n°1, pp, 625-629.

Amer-Yahia S. (2019). "De l'humain dans les marchés virtuels", *Colloque Les sciences de l'information en interaction avec l'humain, Laboratoire d'Informatique de Grenoble, février 2019*, présentation accessible en ligne sur : https://insu.cnrs.fr/sites/default/files/ressource-file/humain_numerique_interaction_sihem_amer_yahia_compresse.pdf

Anderson C. (2008). "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, publié en ligne sur : http://www.wired.com/2008/06/pb-theory/

Andrienko G., Andrienko N., Bosch H., Ertl T., Fuchs G., Jankowski P., Thom D. (2013). "Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics", *Computing in Science and Engineering*, Volume 15, n°3, pp. 72-82.

Arnaud A. (2009). "Valorisation de l'information dédiée aux événements de territoires à risque. Une application cartographique et géovisualisation de la couronne grenobloise". *Thèse de doctorat en géographie, Université Joseph Fourier Grenoble I*, 533 pages.

Audard F., Carpentier S., Oliveau S. (2014). "Les Big Data sont-elles l'avenir de la géographie [théorique et quantitative] ?", *20ème Biennale de géographie d'Avignon, juin 2014, Avignon, France*, pp. 1-4.

Bakis H. (2012). "Le numérique territorial en ses lieux". *Netcom*, Volume 26, n°3-4, pp.149-168

Bakis H., Schon A. (2012). "Ville de la connaissance et terreau numérique : le cas de Montpellier, France". *Netcom,* Volume 26, n°3-4, pp.275-306

Baud P., Bourgeat S., Bras C. (2009). *Dictionnaire de géographie*. Hatier, 4ème édition, Paris, France, 608 pages.

Béguin M., Pumain D. (2017). *La représentation des données géographiques*. Armand Colin, 4ème édition, Malakoff, France, 263 pages.

Blanford J., Bernhardt J., Savelyev A., Wong-Parodi G., Carleton A., Titley D., MacEachren A. (2014). "Tweeting and tornadoes". *Proceedings of the 11th International ISCRAM Conference, University Park, USA,* mai 2014.

Borromeo R. M., Alsayasneh A., Amer-Yahia S., Leroy V. (2017). "Crowdsourcing strategies for text creation tasks", *Proceedings of the 20th International Conference on Extending Database Technology, Venice, Italy*, mars 2017.

Bossu R., Roussel F., Landès M., Steed R., Dupont A., Roch J., Fallou L., Fuenzalida A., Matrullo E., Petersen L. (2018). "Lastquake, un système d'information multicanal pour la réduction du risque sismique global". *Congrès Lambda Mu 21 Maîtrise des risques et transformation numérique : opportunités et menaces, octobre 2018, Reims*, France.

Brovelli M., Zamboni G., Muñoz C., Bonetti A. (2014). "Exploring Twitter georeferenced data related to flood events: an initial approach", *Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, Spain*, juin 2014.

Capdevila J., Cerquides J., Nin J., Torres J. (2017). "Tweet-SCAN: An event discovery technique for geo-located tweets". *Pattern Recognition Letters*, Vol. 93, pp. 58-68.

Capineri C. (2016). "The Nature of Volunteered Geographic Information", in Capineri C., Haklay M., Huang H., Antoniou V., Kettunen J., Ostermann F. and Purves R. (eds.) *European Handbook of Crowdsourced Geographic Information,* Ubiquity Press, London, UK, pp. 15-33.

Cartron F., Fouché A., Jacquin O., Rambaud D., Vullien M. (2018). *Rapport d'information fait au nom de la délégation sénatoriale à la prospective sur les nouvelles mobilités*. Enregistré à la Présidence du Sénat le 8 novembre 2018, 155 pages.

Cavalière C., Davoine P.-A., Lutoff C., Ruin I. (2016). "Analyser des tweets géolocalisés pour explorer les réponses sociales face aux phénomènes météorologiques extrêmes : Réflexions épistémologiques et verrous méthodologiques". *SAGEO'2016, décembre 2016, Nice, France*.

Cebeillac A., Daudé E., Huraux T., (2017). "Where ? When ? And how often ? What can we learn about daily urban mobilities from Twitter data and Google POIs in Bangkok (Thailand) and which perspectives for dengue studies ?", *Netcom*, Volume 31, n°3-4, pp. 283-308.

Cebeillac A., Rault Y.-M. (2016). "Contribution of geotagged Twitter data in the study of a social group's activity space: The case of the upper middle class in Delhi, India", *Netcom*, Volume 30, n°3-4, pp. 231-248."

Chatfield A., Brajawidagda U. (2012). "Twitter tsunami early warning network: A social network analysis if Twitter informations flows", *Proceedings of the 23rd Australian Conference on Information Systems, Geelong, Australia*, décembre 2012.

Chatfield A., Scholl H., Brajawidagda U. (2014). "#Sandy tweets: Citizens' co-production of time-critical information during an unfolding catastrophe", *Proceedings of the 47th Hawaii International Conference on System Science,* janvier 2014.

Chen X., Yang X. (2014). "Does food environment influence food choices? A geographical analysis through tweets", *Applied Geography*, Vol. 5, n°1, pp. 82-89.

Colley A., Thebault-Spieker J., Lin A., Degraen D., Fichman B., Häkkilä J., Kuehl K., Nisi V., Nunes N.-J., Wenig N., Wenig D., Hecht B., Schöning J.(2017). "The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement". *Proceedings of the International Conference on Human Factors in Computing Systems, Denver, USA*, mai 2017.

Croitoru A., Crooks A., Radzikowski J., Stefanidis A. (2017). "Geovisualization of Social Media", in Richardson D., Castree N., Goodchild M., Kobayashi A., Weidong L., Marston R. (eds), *The International Encyclopedia of Geography*, John Wiley and Sons Ltd.

Dashti S., Palen L., Heris M., Anderson K., Anderson S., Anderson T. (2014). "Supporting Disaster Reconnaissance with Social Media Data: A Design-Oriented Case Study of the 2013 Colorado Floods", *Proceedings of the 11th International ISCRAM Conference, University Park, USA,* mai 2014.

Davoine P.-A. (2014). "Contributions géomatiques pour la gestion des risques naturels : modélisation, géovisualisation, acquisition". *Habilitation à diriger des recherches, Université de Grenoble*, 201 pages.

De Blomac F. (2012). "Ushahidi en Haïti : encore des leçons à tirer", *Humanitaire*, publié en ligne sur : https://journals.openedition.org/humanitaire/1306

De Chiara D., Del Fatto V., Sebillo M. (2012). "Visualizing Geographical InformationThrough Tag Clouds" in De Marco M., Te'eni D., Albano V., Za S. (eds), *Information Systems: Crossroads for Organization, Management, Accounting and Engineering*, Springer Verlag, Berlin, Heidelberg.

De Longueville B., Smith R., Luraschi G. (2009). "OMG, from here, I can see the flames!: a case study of mining location based social networks to acquire spatio-temporal data on forest fires". *Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, USA*, novembre 2009.

Douvinet J., Kouadio J., Saint Martin C., Martin G., Gisclard B. (2017). "Une place pour les smartphones et les réseaux sociaux numériques (RSN) dans les dispositifs institutionnels de l'alerte aux inondations en France ?". *Cybergeo: European Journal of Geography*, mis en ligne le 5 janvier 2017, URL : http://cybergeo.revues.org/27875

Dykes J., MacEachren A., Kraak M.-J. (2005). *Exploring Geovisualization*. Elsevier Ltd.

Elwood S., Goodchild M., Sui D. (2012). "Researching volunteered geographic information : spatial data, geographic research and new social practice". *Annals of the Association of American Geographers,* Vol. 102, n°3, pp. 571-590.

Emrich C., Cutter S. (2011). "Social Vulnerability to Climate-Sensitive Hazards in the Southern Unites States". *Weather, Climate and Society*, Vol. 3, n°3, pp.193-208.

Eskenazi M., Pierre M., Boutueil V., Escoffier C. (2017). "Le numérique, indispensable émancipateur de la mobilité électrique ?". *Netcom*, Vol. 31, n°3-4, pp.403-430.

Gal S., Lazier I. (2017). *Les Alpes de Jean de Beins*, Musée de l'Ancien Evêché, imprimerie Les Deux-Ponts, France.

Galinon-Mélénec B., Zlitni S. (2013). "L'Homme-trace, producteur de traces numériques". *Traces numériques, de la production à l'interprétation*, CNRS éditions, pp. 7-19.

Ghosh D., Guha R. (2013). "What are we tweeting about obesity? Mapping tweets with topic modeling and geographic information system", *Cartography and Geographic Information Science*, Vol. 40, n°2, pp. 90-102.

Godfrey D., Johns C., Meyer C., Race S., Sadek C. (2014). "A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets", publié en ligne sur : https://arxiv.org/pdf/1408.5427.pdf

Gomide J., Veloso A., Meira W., Almeida V., Benevenuto F., Ferraz F., Teixeira M. (2011). "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter", *Proceedings of WebSci'11, Koblenz, Germany,* juin 2011.

Goodchild M. (2007). "Citizens as Sensors: The World of Volunteered Geography". *GeoJournal*, Vol. 69, n°4, pp.211-221.

Goodchild M. (2009). "NeoGeography and the nature of geographic expertise", *Journal of Location Based Services*, Vol. 3, n°2, pp.82-96.

Goodchild M., Glennon J. (2010). "Crowdsourcing geographic information for disaster response : a research frontier". *International Journal of Digital Earth,* Vol. 3, n°3, pp.231-241.

Goodchild, M. (2013). "The quality of big (geo)data", *Dialogues in Human Geography*, Vol. 3, pp. 280-284.

Gore R., Diallo S., Padilla J. (2015) "You are what you tweet: connecting the geographic variation in America's obesity rate to Twitter content", *PLoS ONE*, Vol. 10, n°9, e0133505.

Graham M., Hale S., Gaffney D. (2013). "Where in the world are you? Geolocation and language identification in Twitter", *The Professional Geographer*, Vol. 66, n°14, pp. 568-578.

Grataloup C. (2011). "Représenter le monde". *La documentation photographique*, n°8084, 63 pages.

Guermond Y. (2004). "Informatique et géographie", in Bailly A. (dir.), *Les concepts de la géographie humaine*. Masson, 4ème édition, Paris, France, 247 pages.

Guillemette F. (2006). "L'approche de la Grounded Theory ; pour innover ?". *Recherches qualitatives*, Vol. 26, n°1, pp. 32-50.

Haklay M. (2010). "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets", *Environment and Planning B: Planning and Design,* Vol., 37, n°4, pp. 682–703.

Han J., Kamber M., Pei J. (2012). *Data Mining Concepts and Techniques*, 3rd edition, Morgan Kaufmann Publishers, Elsevier, Waltham, Massachusetts, USA.

Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., Ratti C. (2013). "Geo-located Twitter as the proxy for global mobility patterns", *Cartography and Geographic Information Science,* Vol. 41, n°3, pp. 260-271.

Hecht B., Stephens M. (2014). "A Tale of Cities: Urban Biases in Volunteered Geographic Information", *Proceedings of the 8th International Conference on Weblogs and Social Media, Ann Arbor, Michigan, USA*, juin 2014.

Hecker M. (2014). "Le tsunami numérique : gérer les catastrophes naturelles à l'heure des réseaux sociaux". *Etudes*, Volume 7, pp. 9-18.

Herfort B., Porto de Albuquerque J., Schelhorn S.-J., Zipf A. (2014). "Exploring the geographical relations between social media and flood phenomena to improve situational awareness", in Huerta J., Schade S., Granell C. (eds.), *Connecting a digital Europe through Location and Place*, Springer International Publishing Switzerland, pp. 55-71.

Hoffmann M. (1999). "Problems with Peirce's concept of abduction", *Foundations of Science*, Vol. 4, n° 3, pp. 271-305.

Hofmann C., Blais H., Haguet L., Laboulais I., Palsky G., Pansini V. (2012). *Artistes de la carte - De la Renaissance au XXIème siècle*. Autrement, Paris, France, 223 pages.

Huang X., Wang C., Li Z., Huan N. (2019). "A visual–textual fused approach to automated tagging of flood-related tweets during a flood event", *International Journal of Digital Earth*, Vol. 12, n°11, pp. 1248-1264.

Java A., Song X., Finin T., Tseng B. (2007). "Why we Twitter: Understanding Microblogging Usage and Communities", *9th International Workshop on Knowledge Discovery on the Web 2007, and 1st International Workshop on Social Network Analysis 2007, San Jose, California,* août 2007.

Joachims T. (1998). "Making Large-Scale SVM Learning Practical", in Schölkopf B., Burges C., Smola A. (eds.) *Advances in Kernel Methods*, MIT Press, Cambridge, USA.

Joliveau T. (2011). "Le géoweb, un nouveau défi pour les bases de données géographiques". *L'Espace Géographique*, Vol. 40, n°2, pp.151-163.

Joliveau T. (2012). "La géographie et la géomatique au crible de la néogéographie", *Tracés. Revue de Sciences humaines*, mis en ligne le 30 novembre 2012, URL : http://journals.openedition.org/traces/4847

Joliveau T., Noucher M., Roche S. (2013). "La cartographie 2.0, vers une approche critique d'un nouveau régime cartographique". *L'Information géographique*, Vol. 77, n°4, pp.29-46.

Jung J.K. (2014). "Code clouds: Qualitative geovisualization of geotweets". *The Canadian Geographer*, Vol. 59, n°1, pp. 52-68.

Kaplan A., Haenlein M. (2010). "Users of the world, unite! The challenges and opportunities of Social Media". *Business Horizons*, Volume 53, n°1, pp.59-68.

Keim D., Panse C., Sips M. (2005). "Information visualization: scope, techniques and opportunities for geovisualization". in Dykes J., MacEachren A., Kraak M.-J. (eds), *Exploring Geovisulization*, Elsevier.

Kell D., Oliver S. (2003). "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era". *Bioessays,* Vol. 26, n°1, pp. 99-105.

Kitchin R. (2013). "Big data and human geography. Opportunities, challenges and risks". *Dialogues in Human Geography*, Vol. 3, n°3, pp.262-267.

Klein J.-L., Huang P.(2012). "L'espace numérique, un enjeu pour les collectivités locales". *Netcom*, Vol. 26, n°3-4, pp,329-342.

Kounadi O., Lampoltshammer T., Groff E., Sitko I., Leitner M. (2015). "Exploring Twitter to analyse the public's reaction patterns to recently reported homicides in London", *PLos ONE*, Vol. 10, n°3, e01211848.

Lambert N., Zanin C. (2016). *Manuel de cartographie*. Armand Colin, Malakoff, France, 221 pages.

Lebreton C. (2013). "Les territoires numériques de la France de demain". *Rapport à la ministre de l'égalité des Territoires et du Logement, Cécile DUFLOT*, 211 pages.

Lee Hughes A., Palen L. (2009). "Twitter adoption and use in mass convergence and emergency events", *Proceedings of the 6th ISCRAM Conference, Gothenburg, Sweden*, mai 2009.

Lesteven G., Godillon S. (2017). "Les plateformes numériques révolutionnent-elles la mobilité urbaine ?". *Netcom*, Vol. 31, n°3-4, pp.375-402.

Li L., Goodchild M., Xu B. (2013). "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr", *Cartography and Geographic Information Science*, Vol. 40, pp. 61-77.

Lin Y.-R., Margolin D., Keegan B., Baronchelli A., Lazer D. (2013). "#Bigbirds never die : understanding social dynamics of emergent hashtags". *Proceedings of the 7th International AAAI Conference on Weblogsand Social Media, Cambridge,* juillet 2013.

Lucchini F., Elissalde B., Grassot L., Baudry J. (2016). "Paris tweets, données numériques géolocalisées et événements urbains". *Netcom*, Vol. 30, n°3-4, pp.207-230.

Luo W., McEachren A. (2013). "Geo-social visual analytics". *Journal of Spatial Information Science*, Vol. 8, pp. 27-66.

MacEachren A., Gahegan M., Pike W., Brewer I., Cai G., Lengerich E., Hardistry F. (2004). "Geovisualization for knowledge construction and decision support". *IEEE Computer Graphics and Applications,* Vol. 24, n°1, pp. 13-17.

MacEachren A., Kraak M.-J. (2001). "Research Challenges in Geovisualization". *Cartography and Geographic Information Science,* Vol. 28, n°1.

McDougall K. (2012). "An assessment of the contribution of volunteered geographic information during recent natural disasters. Spatially enabling government, industry and citizens: research and development perspectives", in RAJABIFARD A., COLEMAN D. (eds), *Spatially Enabling Government, Industry and Citizens: Research and Development Perspectives*, GSDI Association Press, Needham, USA, 2012.

Meier P. (2015). *Digital Humanitarians: how Big Data is changing the face of humanitarian response*. CRC Press, Inc. Boca Raton, USA, 259 pages.

Mericskay B., Noucher M., Roche S. (2018), "Usages des traces numériques en géographie : Potentiels heuristiques et enjeux de recherche", *L'information géographique*, Vol.82, pp. 39-61.

Mericskay, B. (2016). "La cartographie à l'heure du géoweb : retour sur les nouveaux modes de représentation spatiale des données numériques", *Cartes et géomatique*, n°229-230, pp. 37-50.

Miller H. (2007). "Place-based versus People-based Geographic Information Science". *Geography Compass*, Vol. 1, n°3, pp. 503-535.

Miller H., Goodchild M. (2015). "Data-driven Geography". *GeoJournal*, Vol. 80, n°4, pp.449-461.

Mislove A., Lehmann S., Ahn Y.-Y., Onnela J.-P., Rosenquist J. (2011). "Understanding the demographics of Twitter users", *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain*, juillet 2011.

Monmonnier M. (1991). *How to lie with map*. The University of Chicago Press, Chicago, USA, 176 pages.

Morin-Desailly C. (2018). *Rapport d'informations fait au nom de la commission de la culture, de l'éducation et de la communication sur la formation à l'heure du numérique*. Enregistré à la Présidence du Sénat le 27 juin 2018, 170 pages.

Naz S., Sharan A., Malik N. (2018). "Sentiment classification on Twitter data using Support Vector Machine". *Proceedings of the 2018 IEEE International Conference on Web Intelligence (WI), Santiago, Chile,* décembre 2018.

Ngo P., Wijesekera D. (2012). "Emergency Messages in the Commercial Mobile Alert System", *Proceedings of the 6thInternational Conference on Critical Infrastructure Protection (ICCIP), Washington DC, United States,* mars 2012.

Nguyen D.-Q., Schumann H. (2010). "Taggram: Exploring Geo-data on Maps through a Tag Cloud-Based Visualization", *Proceedings of the IEEE 14th International Conference Information Visualisation, London UK*, juillet 2010.

Nguyen L., Yang Z., Li J., Pan Z., Cao G., Jin F. (2019). "Forecasting people's needs in hurricane events from social network", *IEEE Transactions on Big Data*, publié en ligne sur : https://arxiv.org/abs/1811.04577

Noucher M. (2017). *Les Petites Cartes du web. Analyse critique des nouvelles fabriques cartographiques.* Editions Rue d'Ulm, Paris, France, 70 pages.

Nunez Moscoso J. (2013). "Et si l'on osait une épistémologie de la découverte ? La démarche abductive au service de l'analyse du travail enseignant", *Penser l'éducation*, Laboratoire CIVIIC, pp. 57-80.

Paul J., Buytaert W., Allen S., Ballesteros-Canovas J., Bhusal J., Cieslik K., Clark J., Dugar S., Hannah D., Stoffel M., Dewulf A., Dhital M., Liu W., Lal Nayaval J., Neupane B., Schiller A., Smith P., Supper R. (2018). "Citizen science for hydrological risk reduction and resilience building", *WIREs Water*, Vol. 5, e1262.

Pélissier D. (2015). "De quoi les traces numériques sont-elles le nom ?" Publié en ligne le 27/08/2015 sur : https://presnumorg.hypotheses.org/94

Pezanowski S., MacEachren A., Savelyev A., Robinson A. (2017). "SensePlace3: a geovisual framework to analyze place–time–attribute information in social media", *Cartography and Geographic Information Science,* Vol. 45, n°5, pp. 420-437.

Pirolli F., Crétin-Pirolli R. (2011). "Web social et multimédia : propriétés d'une relation symbiotique". *Les enjeux de l'information et de la communication*. Vol. 2, n°12, pp. 73-82.

Plumejeaud C. (2013)."Modèles et méthodes pour l'information spatio-temporelle évolutive". *Cartes & géomatique*, Comité français de cartographie, pp. 33-38.

Pornon H. (2015). SIG La dimension géographique du système d'information. Dunod, 2ème édition, Paris, France, 303 pages.

Quesnot T. (2016). "L'involution géographique : des données géosociales aux algorithmes". *Netcom*, Vol. 30, n°3-4, pp.281-304.

Rani S., Singh J. (2017). "Sentiment analysis of tweets using support vector machine", *International Journal of Computer Science and Mobile Applications*, Vol. 5, n°10, pp. 83-91.

Reix R., Fallery B., Kalika M., Rowe F. (2011). *Systèmes d'information et management des organisations.* 6ème édition, Vuibert, Paris, France, 480 pages.

Ripberger J., Jenkins-Smith H., Silva C., Carlson D., Henderson M. (2014). "Social Media and Severe Weather: Do Tweets Provide a Valid Indicator of PublicAttention to Severe Weather Risk Communication?", *Weather, Climate and Society*, Vol. 6. pp. 520-530.

Roberts H. V. (2017). "Using Twitter data in urban green space research: a case study and critical evaluation", *Applied Geography*, Vol. 81, pp. 13-20.

Roberts J.C. (2005). "Exploratoty Visualization with multiple linked views". in Dykes J., MacEachren A., Kraak M.-J. (eds.) *Exploring Geovisualization*, Pergamon, 10 février 2005.

Rodriguez Dominguez D., Diaz Redondo R., Fernandez Vilas A., Ben Khalifa M. (2017). "Sensing the city with Instagram: Clustering geolocated data for outlier detection". *Expert Systems with Applications*, Vol. 78, pp.319-333.

Roy Chowdhury S., Amer-Yahia S., Castillo C. (2013). "Tweet4act : using incident specific profiles for classifying crises-related messages". *Proceedings of the 10th International Conference on Information Systems for Crises response and Management, Baden-Baden, Germany,* mai 2013.

Saint-Marc C. (2017). "Formalisation et géovisualisation d'événements historiques issus de risques naturels pour la compréhension des dynamiques spatiales : Application aux inondations ayant touché le système ferroviaire français ". *Thèse de doctorat, Université Grenoble Alpes,* 385 pages.

Sakaki T., Okazaki M., Matsuo Y. (2010). "Earthquake shakes Twitter users: Real-time event detection by social sensors", *Proceedings of the World Wide Web 2010 Conference, Raleigh, North Carolina, USA,* avril 2010.

Sakkari M., Algarni A., Zaied M. (2019). "Urban Crowd Detection Using SOM, DBSCAN and LBSN Data Entropy: A Twitter Experiment in New York and Madrid". *Electronics*, Vol. 8, n°6.

Salas-Olmedo M., Moya-Gómez B., García-Palomares J., Gutiérrez J. (2017). "Tourists' digital footprint in cities: Comparing Big Data sources", *Tourism Management*, Vol. 6, n°6, pp. 13-25.

Samuels R., Taylor J., Mohammadi N. (2018). "The Sound of Silence: Exploring How Decreases in Tweets Contribute to Local Crisis Identification". *Proceedings of the 15th ISCRAM Conference, Rochester, USA*, mai 2018.

Saravanou A., Valkanas G., Gunopulos D., Andrienko G. (2015). "Twitter floods when it rains: a case study of the UK floods in early 2014". *Proceedings of the 24th International Conference on World Wide Web, Florence, Italie,* mai 2015.

Schade S., Diaz L., Ostermann F., Spinsanti L., Luraschi G., Cox S., Nunez M., De Longueville B. (2013). "Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information". *Applied Geomatics*, Vol. 5, n°1, pp. 3-18.

Severo M., Romele A. (2015). "Soft Data" in Severo M., Romele A. (dir.), *Traces numériques et territoires*. Presses des Mines, Paris, France, 270 pages.

Sloan L., Morgan J. (2015). "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter", *PLos ONE*, Vol. 10, n°11, e0142209.

Soliman A., Yin J., Soltani K., Padmanabhan A., Wang S. (2015). "Where Chicagoans tweet the most: Semantic analysis of preferential return locations of Twitter users", *Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics UrbanGIS'15, Bellevue, USA,* novembre 2015.

Steiger E., De Albuquerque J. P., Zipf A. (2015). "An Avanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data", *Transactions in GIS*, Vol. 19, n°6, pp. 809-834.

Sui D., Goodchild M. (2012). "The convergence of GIS and social media: challenges for GIScience", *International Journal of Geographical Information Science*, Vol. 25, n°11, pp. 1737-1748.

Thom D., Bosch H., Koch S., Wörner M., Ertl T. (2012). "Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages", *Proceedings of the IEEE Pacific Visualization Symposium, Songdo, South Korea,* février-mars 2012.

Valentin J., Georges F., Boumenir Y., Dresp-Langley B. (2011). "Espaces virtuels et pré-expérience de l'espace géographique". *Netcom*, Vol. 25, n°1-2, pp.9-32.

Walters B. (2012). "An event-based methodology for climate change and human-environment research". *Geografisk Tidsskrift-Danish Journal of Geography*, Vol. 112, n° 2, p. 135-143.

Widener M., Wenwen L. (2014). "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US", *Applied Geography*, Vol. 5, n°4, pp. 189-197.

Wilson M., Corey K. (2012). "The role of ICT in Arab spring movements". *Netcom*, Vol. 26, n°3-4, pp.343-356.

Yi J. S., Kang Y., Stasko J.(2007). "Towards a deeper understanding of the role of interaction in information visualisation", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, n°6, pp. 1224-1231.

Zou L., Lam N., Shams S., Cai H., Meyer M., Yang S. (2018). "Social and geographical disparities in Twitter use during Hurricane Harvey", *International Journal of Digital Earth*, Vol. 12, n°11, pp. 1300-1318.

Annexes

Table des annexes

Annexe 1 : Outils d'analyse et étapes de préparation des tweets dans la base de données

Le tableau ci-dessous liste l'ensemble des fonctions, extensions et *packages* employés afin de préparer et de traiter les tweets géolocalisés et les jeux de données complémentaires :

Logiciel/ Interface	Fonction/Extension /Package	Utilisations
PostgreSQL + Postgis	Gestion des champs temporels Requêtes spatiales *DBSCAN*	Préparation des tables de tweets pour l'analyse Analyse spatiale
QGIS	*Time Manager*	Animation de données temporelles
R	*RPostgreSQL*	Connexion de la base de données à R
	postGIStools	Exécuter des requêtes spatiales depuis R
	sqldf	Exécuter des requêtes SQL sur un objet *data.frame* (tableau) de R
	tm	Fouille de texte : assure le nettoyage et l'extraction d'un corpus textuel
	Rweka	Associé à *tm*, permet de créer des tableaux d'associations lexicales en *n-grammes*
	RColorBrewer	Création des échelles de couleurs
	wordcloud	Construction des nuages de mots
	topicmodels	Exécuter une LDA
	spacyr	Analyse des catégories grammaticales des mots
	factomineR	Lancer des analyses multivariées
	factoextra	Visualiser les résultats d'une analyse multivariée
	spdep	Test d'autocorrélation spatiale
	leaflet	Cartographie interactive

L'utilisation de l'extension *TimeManager* d'animation des données sous *QGIS* requiert une étape de préparation : il est nécessaire de tronquer l'heure d'émission de chaque tweet selon la résolution temporelle appropriée. La requête suivante est appliquée afin de tronquer le champ *local_timestamp*[1] de la table *tweets_crise* (tweets filtrés et donc rattachés au sujet) à l'échelle de l'heure, dans un nouveau champ intitulé *local_time_hour* : *UPDATE tweets_crise SET local_time_hour = DATE_TRUNC('hour', local_timestamp)* ; on peut également exécuter des requêtes pour agréger les tweets à une résolution plus fine : *UPDATE tweets_crise SET local_time_30min = date_trunc('hour', local_timestamp) + trunc(extract(minutes from local_timestamp)/30) * '30 minutes'::interval* ; cette nouvelle requête va tronquer le champ *local_timestamp* à l'échelle de la demi-heure inférieure (si un tweet est émis à 8h09, alors sa nouvelle heure tronquée sera 8h00 ; en revanche, un tweet émis à 8h31 sera noté à 8h30) et les résultats seront consignés dans le champ nouvellement créé, *local_time_30min.*

[1] Pour rappel, le champ *local_timestamp* correspond à l'horodatage de la création du tweet, au fuseau horaire local.

En ce qui concerne les outils spatiaux spéficiques à PostGIS, nous avons recours aux fonctions assurant les requêtes et jointures spatiales, ou encore la création de zones tampon (*ST_Intersects*, *ST_Buffer*, *JOIN*/*ON*/*GROUP BY*, *etc*.) ; celles-ci offrent l'avantage de temps d'exécution moins longs que sur le logiciel SIG *QGIS* lorsque les tables de données à traiter sont volumineuses (ce qui est le cas d'un jeu de tweets bruts, contenant plusieurs millions d'entités). A partir de sa version 2.3, PostGIS dispose également de la fonction *ST_ClusterDBSCAN(geometry, eps := n, minpoints := n)* qui permet de lancer directement l'algorithme sur une table de données ponctuelles : l'utilisateur doit fournir trois paramètres à la fonction : le champ contenant les données de géolocalisation des objets (*geometry*), la distance seuil en mètres (*eps*) et le nombre minimal de points attendus pour former un cluster (*minpoints*).

Annexe 2 : Apparence et fonctionnalités de l'interface de recherche et d'extraction des tweets géolocalisés

L'interface actuelle est connectée à la base de données *twitterdb* et est structurée en tout premier lieu en quatre parties (figure 1) : le fond correspond à la carte des tweets que l'utilisateur peut afficher sous forme de clusters ou de carte de chaleur (*heatmap*), depuis le volet de paramétrage situé à droite de l'écran. Depuis ce même volet, l'utilisateur peut sélectionner un fond cartographique précis : fond topographique (*Simple*), fond *OpenStreetMap* (*Plan*), imagerie satellite (*Satellite*), ou encore modèle numérique de terrain (*Rivers*). La barre en haut à gauche permet de rechercher un lieu précis sur lequel focaliser la carte (ici, la métropole parisienne). Toujours dans ce volet de droite, l'utilisateur peut créer ses propres couches de données (rubrique *User Layer*).

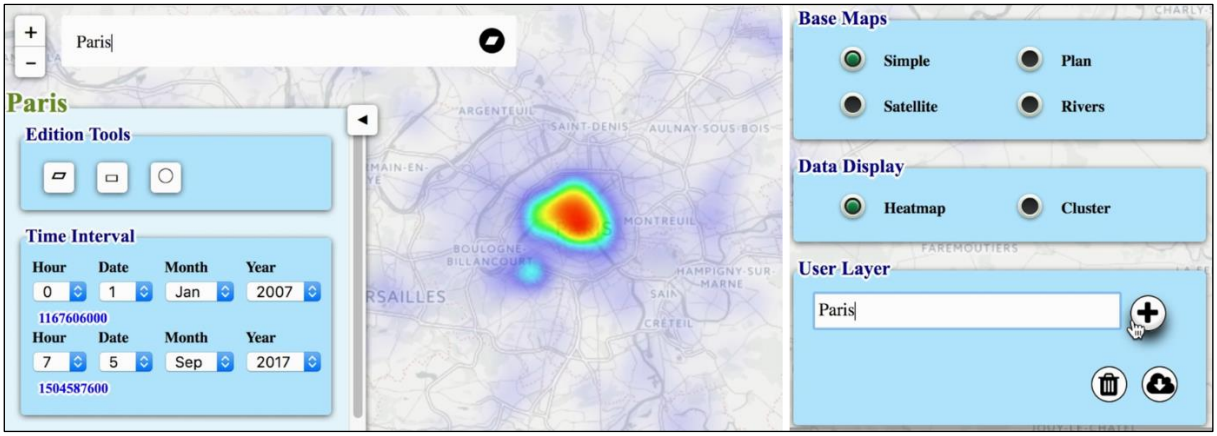


Figure 1 : Apparence de base de l'interface d'extraction des tweets géolocalisés – fond cartographique et volet de paramétrage

L'interface englobe des outils d'édition de régions d'intérêt, dans le volet de filtrage et d'édition de gauche (ici nommé *Paris*, il reprend le nom de la couche que l'utilisateur a saisi dans la rubrique *User Layer* du volet de droite). Ce nouveau volet de filtrage et d'édition contient des outils destinés à dessiner des régions d'intérêt (*Edition Tools*) et à appliquer un filtrage temporel aux tweets affichés sur la carte (*Time Interval*). L'utilisateur peut tracer autant de régions d'intérêt souhaitées afin de sélectionner les tweets contenus à l'intérieur : la gestion des ROI passe de nouveau par le volet de droite et la rubrique *User Layer* et s'apparente au système de calques des DAO (figure 2). Un item apparaît pour chaque région tracée (ici, comme l'utilisateur a créé quatre régions, on trouve quatre items dans le volet). De là, l'utilisateur peut accomplir une série d'actions : masquer la région, la supprimer ou encore en modifier les couleurs d'affichage. Depuis ce même volet, l'utilisateur a également la possibilité d'exporter les régions et les tweets dans le format CSV.

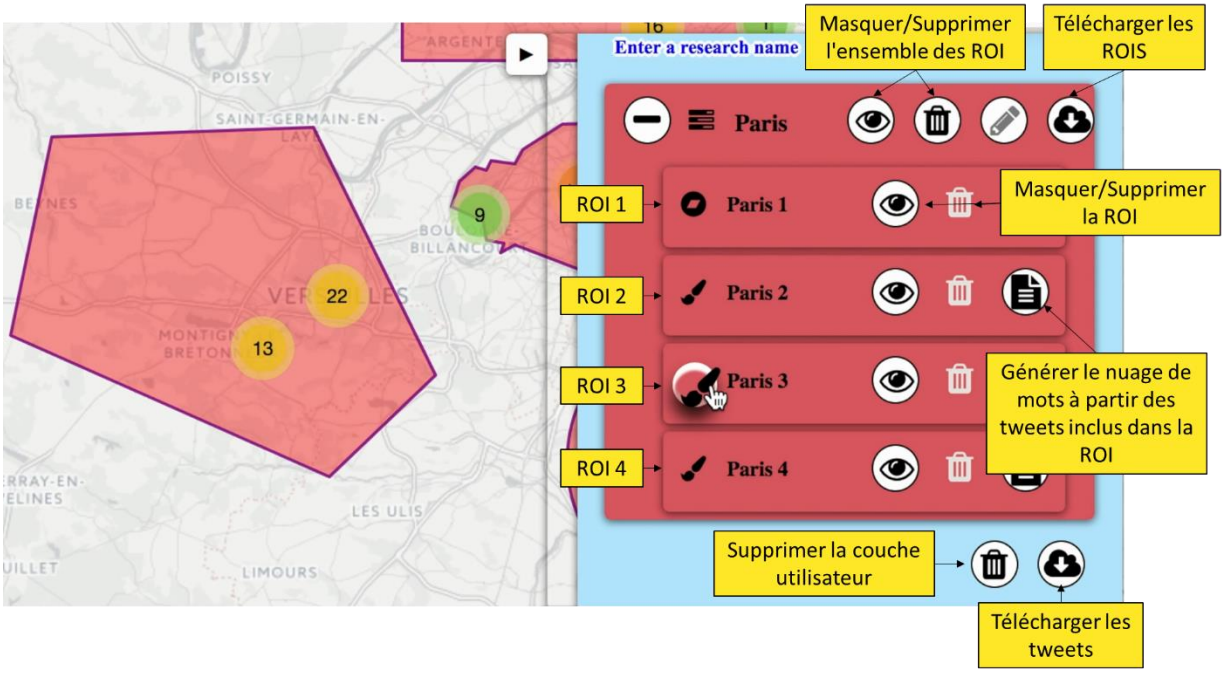
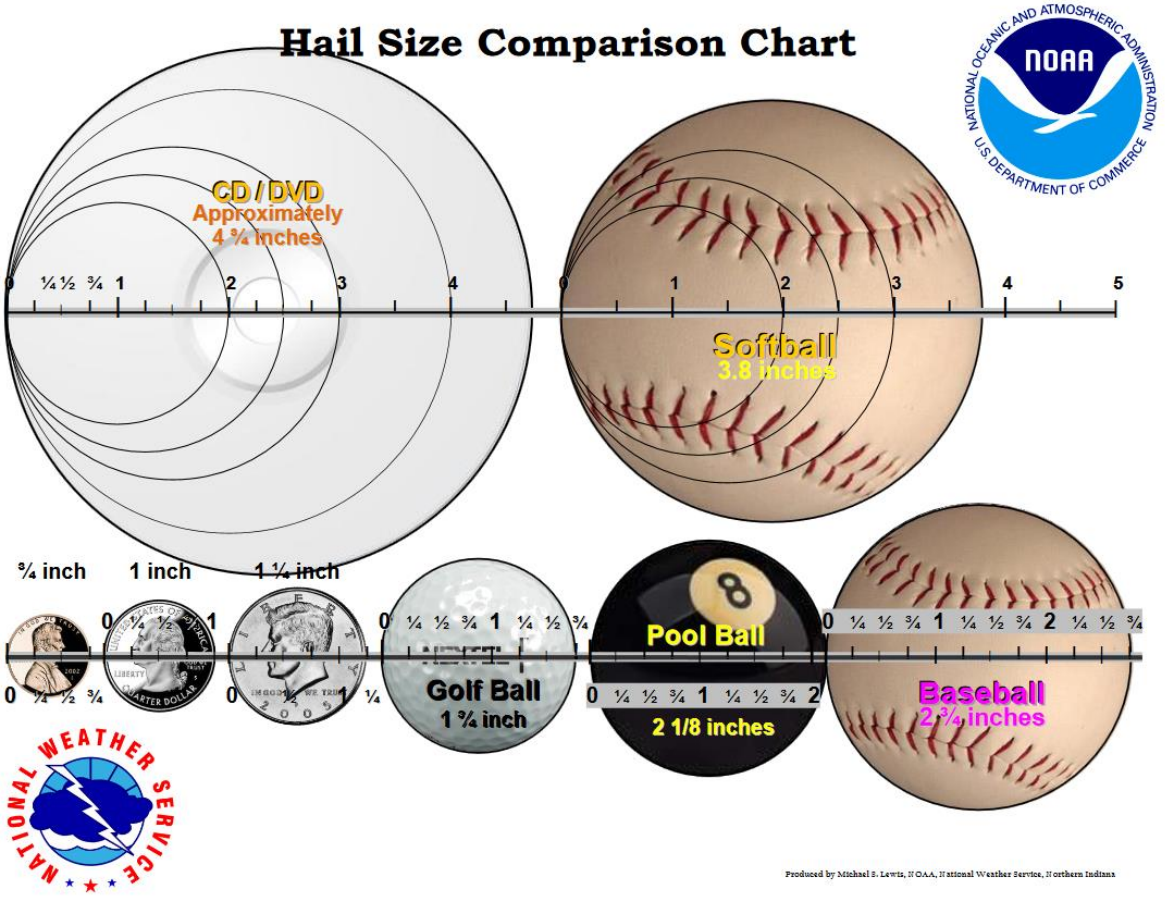


Figure 2 : Volet de gestion des régions d'intérêt créées par l'utilisateur

Comme indiqué sur la figure 2, l'utilisateur a également la possibilité, depuis ce volet, d'obtenir l'aperçu du contenu lexical des tweets, par la génération d'un nuage de mots construit à partir de la sémantique des tweets inclus dans la région considérée (figure 3).

Figure 3 : Affichage d'un nuage de mots représentant les tendances lexicales dans une région d'intérêt

Annexe 3 : Echelle d'estimation de la taille des grêlons - *National Weather Service*



Hail Size Description Chart

Hailstone size	Measurement	
	in.	cm.
bb	< 1/4	< 0.64
pea	1/4	0.64
dime	7/10	1.8
penny	3/4	1.9
nickel	7/8	2.2
quarter	1	2.5
half dollar	1 1/4	3.2
golf ball	1 3/4	4.4
billiard ball	2 1/8	5.4
tennis ball	2 1/2	6.4
baseball	2 3/4	7.0
softball	3.8	9.7
Compact disc / DVD	4 3/4	12.1

Note: Hail size refers to the diameter of the hailstone.