Modélisation de trajectoires sémantiques et calcul de similarité intégrés à un ETL

Cécile Cayèré

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE EUCLIDE

LABORATOIRE L3i

THÈSE présentée par :

Cécile CAYÈRÉ

soutenue le : **17 novembre 2022**
pour obtenir le grade de : **Docteur de l'université de La Rochelle**

Discipline : **Informatique et Applications**

Modélisation de trajectoires sémantiques et calcul de similarité intégrés à un ETL

JURY :

Marlène VILLANOVA	Maître de conférences HDR	Université Grenoble Alpes	Rapporteure
Thomas DEVOGÈLE	Professeur des universités	Université de Tours	Rapporteur
Karine ZEITOUNI	Professeure des universités	Université de Versailles St-Quentin	Examinatrice
Antoine DOUCET	Professeur des universités	La Rochelle Université	Examinateur
Christian SALLABERRY	Maître de conférences HDR	Université de Pau et des Pays de l'Adour	Co-directeur
Cyril FAUCHER	Maître de conférences	La Rochelle Université	Co-directeur
Marie-Noëlle BESSAGNET	Maître de conférences	Université de Pau et des Pays de l'Adour	Encadrante (invitée)
Philippe ROOSE	Maître de conférences HDR	Université de Pau et des Pays de l'Adour	Encadrant (invité)

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens tout d'abord à remercier mes deux directeurs de thèse, Cyril Faucher et Christian Sallaberry, ainsi que mes deux encadrants, Marie-Noëlle Bassagnet et Philippe Roose. Vous avez, tous les quatre, su être patients et bienveillants à mon égard. Que ce soit à La Rochelle ou à Pau, en présentiel ou en distantiel, en période confinement ou pas, vous étiez présents pour répondre à toutes mes questions et avez été d'une grande aide durant ces trois ans. Grâce à vous, cette thèse s'est déroulée dans de très bonnes conditions.

Je n'oublierai jamais que c'est bel et bien grâce à vous, Christian et Marie-Noëlle que je me suis lancée dans la recherche après mon stage recherche de Master. En première année de Master, vous étiez déjà mes encadrants de TER (Travail d'étude et de recherche). Je me rappelle encore de cette question que vous avez posée à la fin de la dernière réunion de TER : *"Qui d'entre vous veut continuer dans la recherche ?"* à laquelle j'ai évidemment répondu : *"Moi !"*. Le stage a suivi, puis la thèse.

Je remercie également toutes les personnes du projet DA3T ainsi que Maxime qui a réalisé une grande partie du développement présenté dans ce mémoire.

Je voudrais remercier toutes les personnes qui prendront le temps de lire ce mémoire, mais également toutes les personnes qui se sont intéressées de près ou de loin à mon travail et enfin, toutes celles que j'ai eu le plaisir et la chance de rencontrer durant les conférences auxquelles j'ai eu l'occasion de participer.

Je remercie Esther, ma meilleure amie. Même si nous ne nous voyons plus aussi souvent qu'avant, les quelques moments que nous passons ensemble sont si précieux à mes yeux. Je remercie également tous mes amis. Passer du temps avec vous m'a permis d'évacuer et de souffler un peu.

Je souhaite remercier toute ma famille : mes parents, mon frère, Mamie Mané, Mamie Jeannette et Papi René, Virginie, Jean-Paul, Alexandra, Jean-Pierre, Joseph, Léopold, Albert, Louis et Catalina. Je vous remercie tous très fort car vous avez tous été un soutien incroyable pour moi. Ma famille a été mon carburant dans cette aventure. Je remercie particulièrement mes parents qui m'ont soutenu du début à la fin de cette expérience, je n'y serai jamais arrivée sans vous. Merci papa pour la relecture. Merci maman pour tous ces moments passés ensemble à discuter de la thèse et d'autres choses. Je remercie également mon frère qui m'a toujours poussé à faire ce que je voulais de ma vie, sans jamais juger mes choix. Même si tu

es mon petit frère, tu es depuis longtemps un exemple de force, d'intelligence et de maturité pour moi. Notre objectif de traverser les Pyrénées à la fin de ma thèse a été une grande source de motivation. J'espère que nous trouverons finalement l'occasion de le faire ensemble.
Pour finir, j'aimerais remercier du plus profond de mon coeur mon amoureux, Bastien. Je pourrais t'écrire un paragraphe de remerciement aussi long que cette thèse mais ce n'est pas nécessaire car tu sais déjà tout. Tu as toujours été là dans mes moments de doutes et de remises en question. Tu m'as vue dans mes bons comme dans mes mauvais jours et tu as toujours été compréhensif. Je ne compte plus le nombre de fois où je t'ai parlé de mon avancement dans cette thèse, où je t'ai demandé si mes figures étaient assez claires, mes tournures de phrases assez compréhensibles. Maintenant cette étape de ma vie est terminée, nous allons pouvoir nous concentrer sur la suite...

Je dédie cette thèse à ma famille et à mon amoureux.

Publications

Publications nationales

— Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose, Maxime Masson. *Mesure de similarité pour les trajectoires sémantiques : prise en compte de trois niveaux de granularité*. INFORSID : 5-20. Dijon, France, 2022.

— Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose. *Proposition d'un modèle de trajectoires multi-aspects et multi-niveaux appliqué au tourisme*. Ingénierie des Connaissances : 56-64. Bordeaux, France, 2021.

— Cécile Cayèré, Jérémy Richard, Mélanie Mondo. *Enrichir les traces GPS des visiteurs : enjeux et propositions*. INFORSID, Atelier Évolution des SI : vers des SI pervasifs ? : 40-48. Dijon, France, 2021.

— Cécile Cayèré. *Plateforme ETL dédiée à l'analyse de la mobilité touristique dans une ville*. INFORSID, Forum Jeunes Chercheuses - Jeunes Chercheurs : 13-16. Dijon, France, 2020.

— Maxime Masson, Cécile Cayèré, Marie-Noëlle Bessagnet, Christian Sallaberry, Philippe Roose, Cyril Faucher. *Visualisation spatio-temporelle de données de mobilité touristique extérieures*. EGC, Atelier GAST : 26-40. Blois, France, 2022.

Publications internationales

— Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose, Maxime Masson, Jérémy Richard. *Multi-Level and Multiple Aspect Semantic Trajectory Model : Application to the Tourism Domain*. ISPRS International Journal of Geo-Information, MDPI, 10(9) : 592. 2021.

— Cécile Cayèré, Cyril Faucher, Christian Sallaberry, Marie-Noëlle Bessagnet, Philippe Roose. *Tools for processing digital trajectories of tourists*. 21st IEEE International Conference on Mobile Data Management, Demo Paper : 232-233. Versailles, France, 2020.

— Maxime Masson, Cécile Cayèré, Marie-Noëlle Bessagnet, Christian Sallaberry, Philippe Roose, Cyril Faucher. *An ETL-like platform for the processing of mobility data*. ACM SAC : 547-555. Brno, Czech Republic, 2022.

— Salah Eddine Boukhetta, Christophe Demko, Karell Bertet, Jérémy Richard, Cécile Cayèré. *Temporal Sequence Mining Using FCA and GALACTIC*. 26th International Conference on Conceptual Structures : 185-199. Bolzano, Italy, 2021.

Résumé

Cette dernière décennie, nous avons pu constater une montée en popularité des applications touristiques, patrimoniales, sportives basées sur la localisation des téléphones. Ces applications collectent des données spatio-temporelles, données géolocalisées et horodatées, qui permettent de retracer plus ou moins précisément le déplacement des utilisateurs au cours du temps. La trace de mobilité correspond à l'ensemble des données spatio-temporelles d'un utilisateur.

Par ailleurs, dans le domaine de la recherche, le concept de trajectoire sémantique prend de plus en plus de place quand il s'agit de représenter et d'étudier la mobilité. Il s'agit d'une sous-partie d'intérêt de la trace enrichie avec des données externes afin de donner plus de sens au déplacement.

Dans le projet régional DA3T, servant de cadre à cette thèse, nous faisons l'hypothèse que l'analyse des traces de mobilité de touristes visitant une ville peut aider les aménageurs dans la gestion et la valorisation des territoires touristiques. Pluridisciplinaire, ce projet fait collaborer des géographes et des informaticiens dont l'objectif commun est de concevoir des méthodes et développer des outils logiciels d'aide à l'analyse de ces traces.

Cette thèse s'intéresse au traitement des traces de mobilité, notamment touristiques, et propose une plateforme modulaire, de type ETL, permettant de créer et d'exécuter des chaînes de traitement sur ces données. Au fil d'une chaîne de traitement, la trace de mobilité brute se mue en trajectoires sémantiques. Les modules, constituant les chaînes de traitement, effectuent un traitement bas niveau et appartiennent à différentes catégories (p. ex. pré-traitement, enrichissement, visualisation, etc.). Les contributions de cette thèse sont : (i) un modèle de trajectoire sémantique multi-niveau et multi-aspect et (ii) deux mesures calculant la similarité entre deux trajectoires sémantiques s'intéressant aux dimensions spatiale, temporelle et thématique. Notre modèle (i) est utilisé comme modèle de transition entre les modules d'une chaîne de traitement. Nous l'avons mis à l'épreuve en instanciant des trajectoires sémantiques issues de différents jeux de données de domaines variés. Nos deux mesures (ii) sont intégrées à notre plateforme comme modules de traitement. Ces mesures présentent des originalités : l'une est la combinaison de sous-mesures, chacune permettant d'évaluer la similarité des trajectoires sur les trois dimensions (c.-à-d. spatiale, temporelle et thématique) et selon trois niveaux de granularité différents (c.-à-d. micro, méso et macro), l'autre est la combinaison de deux sous-mesures bidimensionnelles (c.-à-d. spatio-temporelle et tempo-thématique) centrées autour d'une dimension en particulier (ici, temporelle). Nous

avons évalué nos deux mesures en les comparant à d'autres mesures et à l'avis de géographes experts en tourisme.

Ainsi, nous avons proposé une plateforme générique et extensible pouvant prendre en entrée n'importe quel type de traces de mobilité (p. ex. traces précises ou imprécises, traces étalées dans le temps ou pas, traces d'humains, d'animaux, de véhicules, etc.) appartenant à n'importe quel domaine (p. ex. géographie du tourisme, zoologie, sociologie, etc.). Deux contributions (un modèle et deux mesures) découlent de cette plateforme et ont, toutes les deux, donné des résultats satisfaisants lors des expérimentations mises en oeuvre.

Chapitre 1

Introduction

Le monde est en perpétuel changement. Chaque observation faite sur un phénomène spécifique à un instant donné (p. ex. localisation d'un être vivant, température de l'air constituant la météo, activité en ligne d'un utilisateur, etc.) est unique et est le résultat de quantité de facteurs différents. Chaque observation étant associée à un moment précis, la dimension temporelle joue un rôle majeur dans l'étude de l'évolution d'un phénomène. Une telle évolution peut être représentée sous la forme d'une série temporelle, c.-à-d. une séquence de valeurs relevées à intervalles de temps constants ou variables (Nakamura et al. [2013b], Magdy et al. [2015]). Lorsque le phénomène observé est de nature physique (p. ex. être vivant, météo, parcelle de terrain, etc.), il est courant d'observer l'évolution de sa géométrie ou de son déplacement dans l'espace au cours du temps. Ainsi, la dimension spatiale prend de l'importance. Nous parlons de **données spatio-temporelles** que Flouvat [2019] classe en quatre grandes catégories : événement, région ou champ continu, réseau et mobilité. Dans ce mémoire, nous nous intéressons aux données de mobilité qui décrivent le déplacement d'un objet mobile (p. ex. animal, humain, véhicule, etc.). Ces données sont détaillées dans le chapitre 2, partie 2.1.

Le déplacement d'un objet mobile est de nature continue observable partout dans notre environnement physique. Pour faciliter la capture et le stockage d'un déplacement, il est discrétisé ; c.-à-d. qu'il est simplifié en une suite de **positions** géolocalisées et horodatées, appelée **trace de mobilité**. Selon Parent et al. [2013], une **trajectoire brute** d'un objet mobile est une partie de la trace de mobilité qui a de l'intérêt pour une application donnée. Dans le contexte de la mobilité touristique, les trajectoires peuvent être construites sur des critères spatiaux et/ou temporels (p. ex. la trace d'une semaine d'un touriste pourrait résulter en un ensemble de trajectoires journalières). Souvent, la trajectoire brute est limitée pour expliquer un déplacement. Cette constatation a donné lieu à la création du concept de **trajectoire sémantique** [Parent et al., 2013] qui consiste à lier des données d'enrichissement à la trajectoire, à des parties de la trajectoire et/ou à des positions de la trajectoire. Une donnée d'enrichissement peut être simple ou complexe. Une donnée d'enrichissement simple décrit un phénomène particulier sous la forme d'un **label** textuel (p. ex. un point d'intérêt peut être décrit par son nom). Une donnée d'enrichissement complexe, quant à elle, décrit un objet du monde réel sous la forme d'un agrégat de caractéristiques (c.-à-d. un ensemble de données de différents types). Dans ce dernier cas, nous parlons d'**aspect** [Mello et al., 2019] (p. ex. un point d'intérêt peut être décrit par son nom, sa localisation, ses heures d'ouverture, etc.). Un

intervalle temporel de la trajectoire associé à des données d'enrichissement s'appelle communément un **épisode** [Yan et al., 2010].

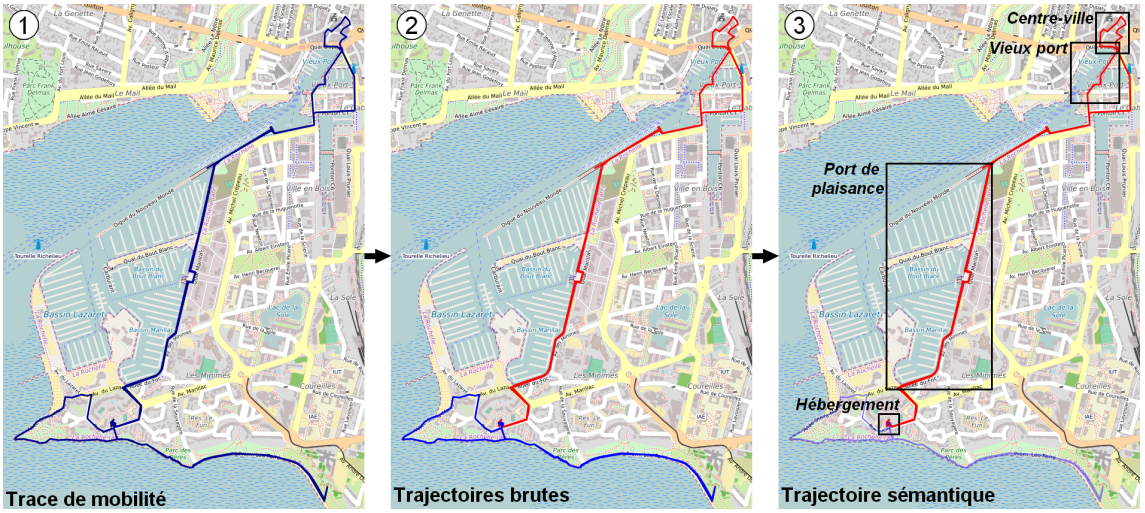


Figure 1.1 – De la trace de mobilité à la trajectoire sémantique en passant par la trajectoire brute

La figure 1.1 schématise les étapes du processus de traitement d'une trace de mobilité. La partie 1 de la figure 1.1 représente la trace de mobilité brute (en bleu foncé) d'un touriste durant un séjour de deux jours à La Rochelle. Il s'agit de l'intégralité du déplacement capturé qui n'a pour l'instant subi aucun traitement. La partie 2 de la figure 1.1 montre deux trajectoires brutes journalières (en bleu et en rouge) construites à partir de la trace de mobilité. Dans cet exemple, les trajectoires ont été construites sur le critère temporel d'une trajectoire par journée mais elles auraient aussi pu être construites sur d'autres critères. Enfin, la partie 3 de la figure 1.1 montre l'une des deux trajectoires brutes précédentes. Elle a été enrichie et est devenue une trajectoire sémantique (en rouge). Les données d'enrichissement consistent ici en l'ensemble des lieux d'intérêt traversés par le touriste durant une journée (c.-à-d. l'hébergement du touriste, le port de plaisance, le vieux-port et le centre-ville).

Pour passer d'une trajectoire brute à une trajectoire sémantique, plusieurs traitements dits d'enrichissement peuvent être mis en œuvre, à savoir l'annotation et la segmentation. Ainsi, l'annotation est le fait d'attacher des données d'enrichissement à la trajectoire entière, des parties de la trajectoire et/ou des positions de la trajectoire. Par exemple, certaines positions de la trajectoire sémantique présentée sur la figure 1.1 sont annotées avec les lieux d'intérêt traversés par le touriste. La segmentation est le fait de diviser la trajectoire en épisodes selon un certain critère. Une **interprétation** de la trajectoire est une séquence d'épisodes obtenue après segmentation de la trajectoire [Yan et al., 2011]. Par exemple, la trajectoire sémantique présentée sur la figure 1.1 peut être segmentée selon les annotations de type lieu. Cela donne une interprétation de la trajectoire telle que :

$$\{([t_0, t_1], \text{"Hébergement"}), ([t_2, t_3], \text{"Port de plaisance"}), ([t_4, t_5], \text{"Vieux port"}), ...$$
$$([t_{n-1}, t_n], \text{"Hébergement"})\}$$

Les données de mobilité sont plus amplement détaillées et formalisées dans le chapitre 2, partie 2.1. Dans cette introduction, nous présentons, d'abord, le contexte de cette thèse (cf. partie 1.1) en décrivant le projet qui lui sert de cadre, les motivations, l'objectif, les verrous de recherche ciblés et les hypothèses de travail mises en place. En suivant, nous introduisons les deux contributions de cette thèse (cf. partie 1.2). Enfin, nous résumons l'organisation de la suite du mémoire (cf. partie 1.3).

1.1 Contexte

Dans cette partie, nous définissons le contexte de la thèse en présentant le projet qui lui sert de cadre. Puis, nous décrivons le scénario de motivation servant de fil directeur à nos travaux. Nous définissons les objectifs de la thèse. Ensuite, nous exposons les verrous scientifiques soulevés par nos travaux et enfin, nous présentons nos hypothèses de travail.

1.1.1 Projet DA3T

Cette thèse est menée dans le cadre du projet régional Nouvelle-Aquitaine **DA3T** (c.-à-d. **D**ispositif d'**A**nalyse des **T**races numériques pour la valorisation des **T**erritoires **T**ouristiques) [1]. Comme l'indique son nom, l'objectif du projet est de proposer un dispositif d'analyse des traces de mobilité dans le but d'aider les aménageurs et décideurs locaux dans la gestion et la valorisation des territoires touristiques en Nouvelle-Aquitaine. Il s'agit d'un projet pluridisciplinaire réunissant informaticiens et géographes dans le but de produire des outils et des méthodes de traitement et d'analyse de traces de mobilité. Le projet s'articule autour de trois thèses, à savoir :

— une thèse en géographie, soutenue par Mélanie Mondo, sur l'apport des traces de mobilité couplées à des entretiens de touristes volontaires dans l'étude et la compréhension des comportements touristiques [Mondo, 2022] ;
— une thèse en informatique, menée par Cécile Cayèré, sur la conception d'outils dédiés au traitement des traces de mobilité (que nous allons développer dans ce mémoire) ;
— une seconde thèse en informatique, menée par Jérémy Richard, sur le couplage des interactions avec un compagnon de visite (p. ex application mobile du musée, tablette tactile prêtée à l'accueil du musée, etc.) et des traces de mobilité *indoor* de visiteurs pour résumer leur expérience de visite au sein d'un musée. Cette thèse n'est pas encore soutenue mais une partie de ces travaux est présentée dans des articles [Jérémy et al., 2021].

Travaillant à la fois sur les traces de mobilité *indoor* (c.-à-d. en intérieur, comme dans un bâtiment, un musée, etc.) et *outdoor* (c.-à-d. en extérieur, comme dans une région touristique, une ville balnéaire, etc.) et souhaitant garder le contrôle sur nos jeux de données, nous avons mis en place différents processus de collecte. Dans cette thèse, nous nous intéressons principalement aux traces de mobilité *outdoor* et nous ne détaillons pas les processus de collecte

1. Site DA3T : `https://lienss.univ-larochelle.fr/DA3T`

des traces de mobilité *indoor* faisant l'objet de la thèse de Jérémy Richard.

Afin de collecter un jeu de traces de mobilité *outdoor* appartenant à des touristes, nous avons développé une application mobile appelée Géoluciole [2]. Cette dernière capture, à intervalles de temps réguliers, la position d'un téléphone mobile identifié ainsi qu'un ensemble de données supplémentaires (p. ex. la vitesse de déplacement au moment de la capture, l'orientation du déplacement au moment de la capture, etc.). La prise en charge de la capture des données de touristes volontaires nous permet de garder le contrôle sur leur sélection, leur sauvegarde et leur pré-traitement au sein de l'application et de minimiser les effets de boîtes noires qui pourraient survenir avec l'utilisation d'autres applications de collecte ou des jeux de traces de mobilité librement disponibles sur l'*Open Data*. Le lieu de collecte principal que nous avons choisi est la ville de La Rochelle, une ville côtière située à l'ouest de la France dans le département de la Charente-Maritime. La campagne de collecte a eu lieu durant l'été 2020 près de l'office du tourisme de La Rochelle ; campagne au cours de laquelle Mélanie Mondo invita un nombre important de touristes à installer et lancer Géoluciole durant leur séjour. L'originalité du projet se trouve dans sa façon de compléter la trace de mobilité brute d'un touriste volontaire à l'aide d'un entretien semi-directif lui permettant d'apporter des précisions sur son séjour. Dû aux contraintes induites par le passage d'un tel entretien en fin de séjour, seule une minorité de touristes a accepté d'y participer. Ainsi, notre jeu de données Géoluciole comprend 92 traces de mobilité appartenant à des touristes volontaires, parmi lesquels 15 ont passé un entretien. Cette campagne de collecte est décrite plus en détails dans la thèse de Mélanie Mondo [Mondo, 2022].

Présentons maintenant le scénario de motivation basé sur nos données servant de fil directeur tout au long de ce mémoire.

2. Lien GooglePlay Géoluciole : `https://play.google.com/store/apps/details?id=fr.univ_lr.geoluciole`

1.1.2 Scénario de motivation et identification des besoins

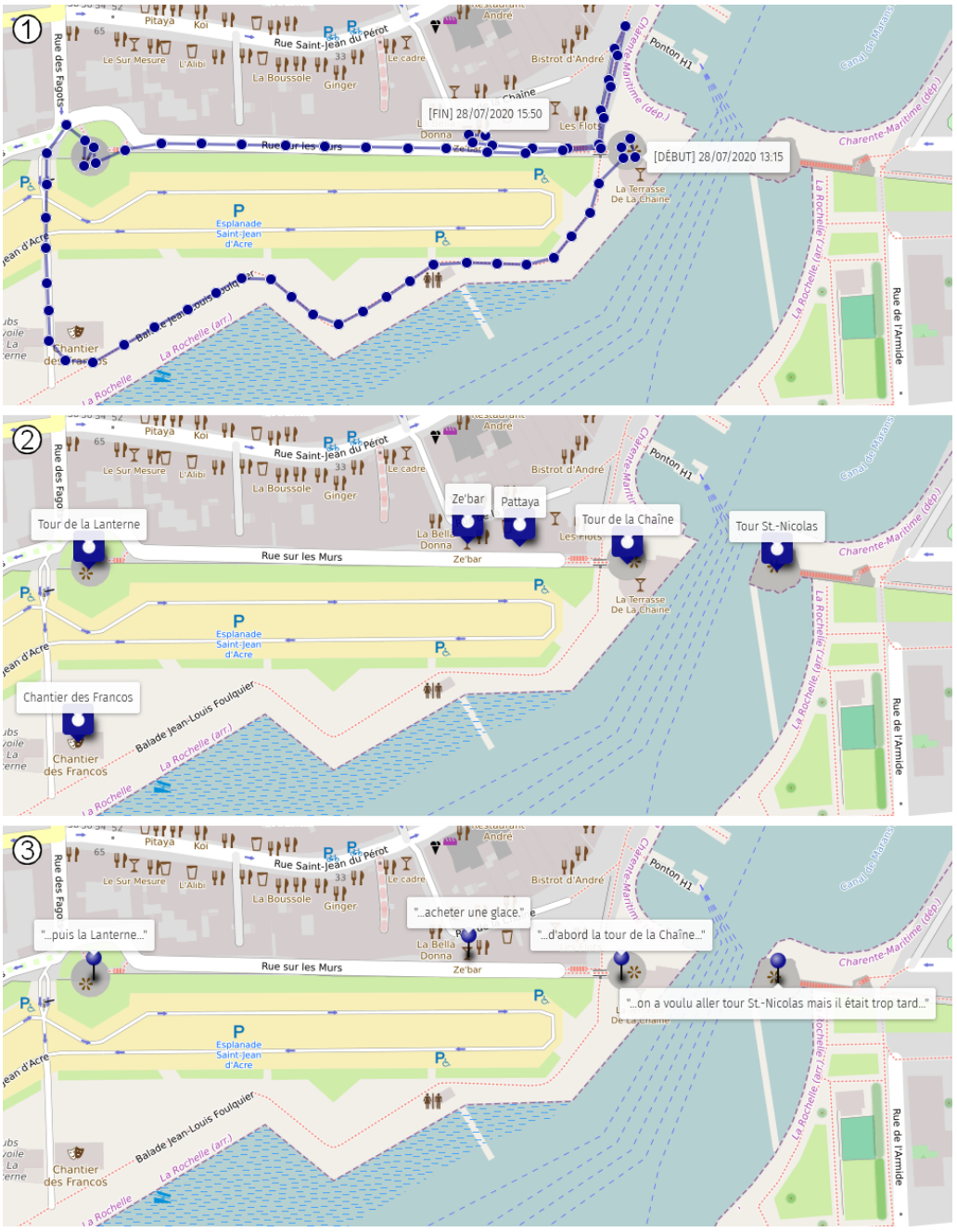


Figure 1.2 – Données manipulées dans le projet DA3T

La figure 1.2 illustre les trois types de données manipulées dans le projet DA3T (c.-à-d. les traces de mobilité brutes, les données d'enrichissement et les données issues des entretiens). Prenons la trace de mobilité brute d'un touriste visitant La Rochelle. Cette trace représente le déplacement du touriste grâce à une séquence de positions géolocalisées et horodatées. Il est possible qu'elle présente des valeurs aberrantes dû à la qualité, à la version et à l'âge du téléphone mobile utilisé mais aussi dû à l'état de la réception du signal GPS. Il est donc nécessaire de faire des traitements de nettoyage et de correction de cette trace pour la rendre

plus fidèle à la réalité du déplacement. La trace de mobilité d'un touriste représente son déplacement complet commençant à la première position capturée et se terminant à la dernière. Dans cet exemple, nous nous intéressons à une trajectoire brute nettoyée (cf. figure 1.2, 1) extraite de la trace de mobilité initiale. Nous pouvons émettre quelques premières observations sur cette trajectoire brute. Elle est constituée d'un enchaînement de phases de mouvements et de phases d'arrêt (reconnaissables par les amas de points) qui se succèdent de manière alternée. En effet, elle commence par un arrêt qui, en connaissant La Rochelle, s'avère être la tour de la Chaîne ; puis, elle enchaîne sur une phase de déplacement, longeant la côte ; elle s'arrête de nouveau, à proximité immédiate ou à l'intérieur de la tour de la Lanterne ; elle se dirige vers le Vieux Port puis rebrousse chemin jusqu'à s'arrêter à nouveau. La raison du demi-tour reste incertaine : nous pourrions, par exemple, l'expliquer par la recherche d'un point d'intérêt particulier ou envisager que le touriste ait perdu son chemin. Tout ce discours peut être plus ou moins précis selon la connaissance de la ville et de ses points d'intérêt. Ainsi, enrichir la trace avec les points d'intérêt de la ville (cf. figure 1.2, 2), ou toutes sortes de données permettant d'apporter des informations supplémentaires, permettrait de mieux comprendre le déplacement du touriste (p. ex. la météo évoluant au fil du déplacement, les quartiers traversés, les événements en cours concordant avec le lieu et le moment d'une partie du déplacement, etc.). Ces données d'enrichissement proviennent de sources diverses et hétérogènes telles que l'*Open Data*, de bases de données spécialisées, d'avis d'expert ou, comme nous allons le voir maintenant, d'entretiens. Cet enrichissement indique les points d'intérêt traversés par la trace. Nous pouvons valider que la trace s'est arrêtée à la tour de la Chaîne, à la tour de la Lanterne et à Ze'bar, et nous constatons qu'elle est passée à proximité du chantier des Francos et de Pattaya. L'extrait d'entretien (cf. figure 1.2, 3) correspondant à la trace est le suivant :

> *"Le 28/07, visite des tours : d'abord la tour de la Chaîne, puis la Lanterne, puis on a voulu aller tour Saint-Nicolas mais il était trop tard alors nous avons fait demi-tour et sommes allés acheter une glace."*

Cet extrait d'entretien permet de confirmer certaines déductions (p. ex. la visite de la tour de la Chaîne) ou d'en corriger d'autres (p. ex. le demi-tour au Vieux-Port car il n'y a plus assez de temps pour aller visiter la tour Saint-Nicolas). Il permet également de découvrir d'autres activités telles que l'achat d'une glace. Ainsi, agréger ces trois types de données, issues de sources distinctes, nous permet d'être plus précis dans nos analyses. En somme, l'entretien peut compléter les silences de la trace de mobilité et vice-versa.

Cet exemple illustre la motivation principale de notre projet qui est d'enrichir les traces de mobilité avec différents types de données d'enrichissement pour arriver à mieux comprendre les comportements touristiques. Ainsi, notre premier besoin est un moyen de représenter des données de mobilité à n'importe quelle étape de leur traitement, c.-à-d. de la trace de mobilité brute à la trajectoire sémantique enrichie avec des données hétérogènes.

Une fois pré-traitées et enrichies, des analyses peuvent être réalisées par des experts sur les trajectoires pour essayer de comprendre les déplacements touristiques. L'étude des com-

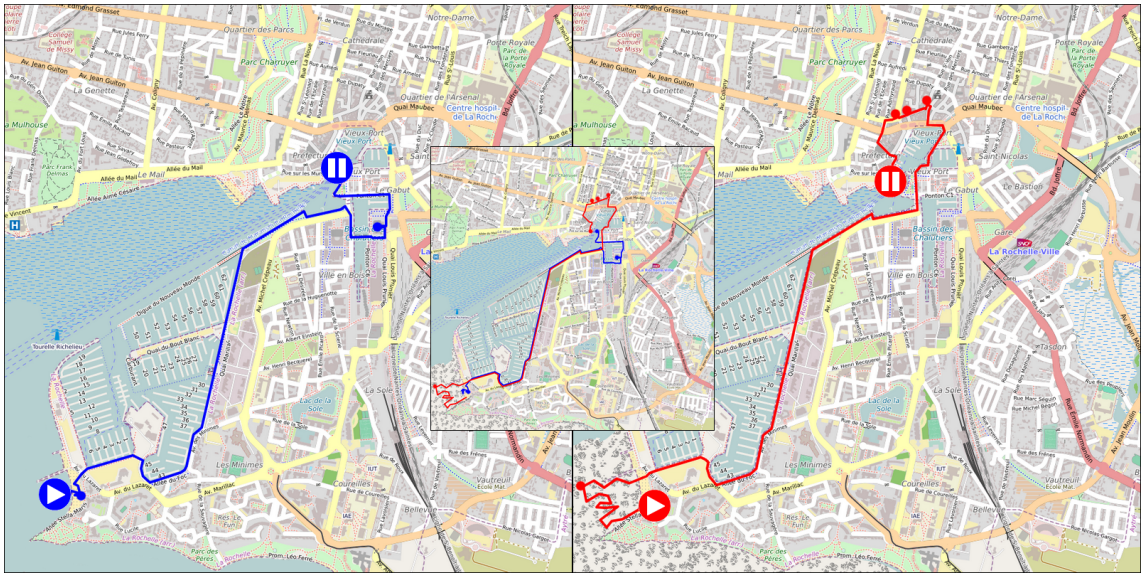


Figure 1.3 – Deux trajectoires sémantiques appartenant à deux touristes différents

portements touristiques via la comparaison des trajectoires deux à deux selon chacune de leurs dimensions (c.-à-d. dimensions spatiale, temporelle et thématique) intéresse particulièrement les géographes du projet. Par exemple, ils souhaitent pouvoir comparer deux trajectoires représentatives appartenant à deux catégories de touristes différentes (p. ex. visiteurs orientés culture, sport, etc.) pour évaluer si ces catégories présentent des similitudes sur certaines dimensions ou comparer une trajectoire de touriste avec un parcours type de l'office du tourisme pour évaluer à quel point les touristes suivent le parcours proposé.

La figure 1.3 illustre la dimension spatiale de deux trajectoires sémantiques construites à partir de traces collectées. La figure 1.4 met en évidence la dimension temporelle à travers un axe du temps ainsi que la dimension thématique grâce à la représentation des différentes interprétations des trajectoires. Les aspects considérés ici sont les points d'intérêt, la météo, la marée et les activités touristiques issues des entretiens. Comparons manuellement ces deux trajectoires :

— **Dimension spatiale (cf. figure 1.3) :** Des similitudes spatiales sont clairement visibles entre les trajectoires 1 et 2 (respectivement, bleue et rouge). Les deux se situent au centre-ville de La Rochelle et ont un point de départ et d'arrivée plus ou moins similaires (dans les mêmes zones). Les deux trajectoires comportent deux phases plutôt stationnaires ou de visite (c.-à-d. les zones de départ et d'arrivée où les positions sont plus proches les unes des autres) séparées par une phase de déplacement (c.-à-d. les longues lignes sans détour où les positions sont plus éloignées les unes des autres). De plus, il est à noter que les deux touristes traversent la ville en suivant le même chemin. Cependant, les arrêts (c.-à-d. amas de points aux mêmes endroits) que nous pouvons identifier à l'oeil nu ne sont pas les mêmes. Pour améliorer la comparaison, il faudrait pouvoir comparer les deux séquences de coordonnées géographiques. Nous pouvons déduire de toutes ces analyses que, malgré quelques légères différences, les deux trajectoires sont très similaires sur le plan spatial.

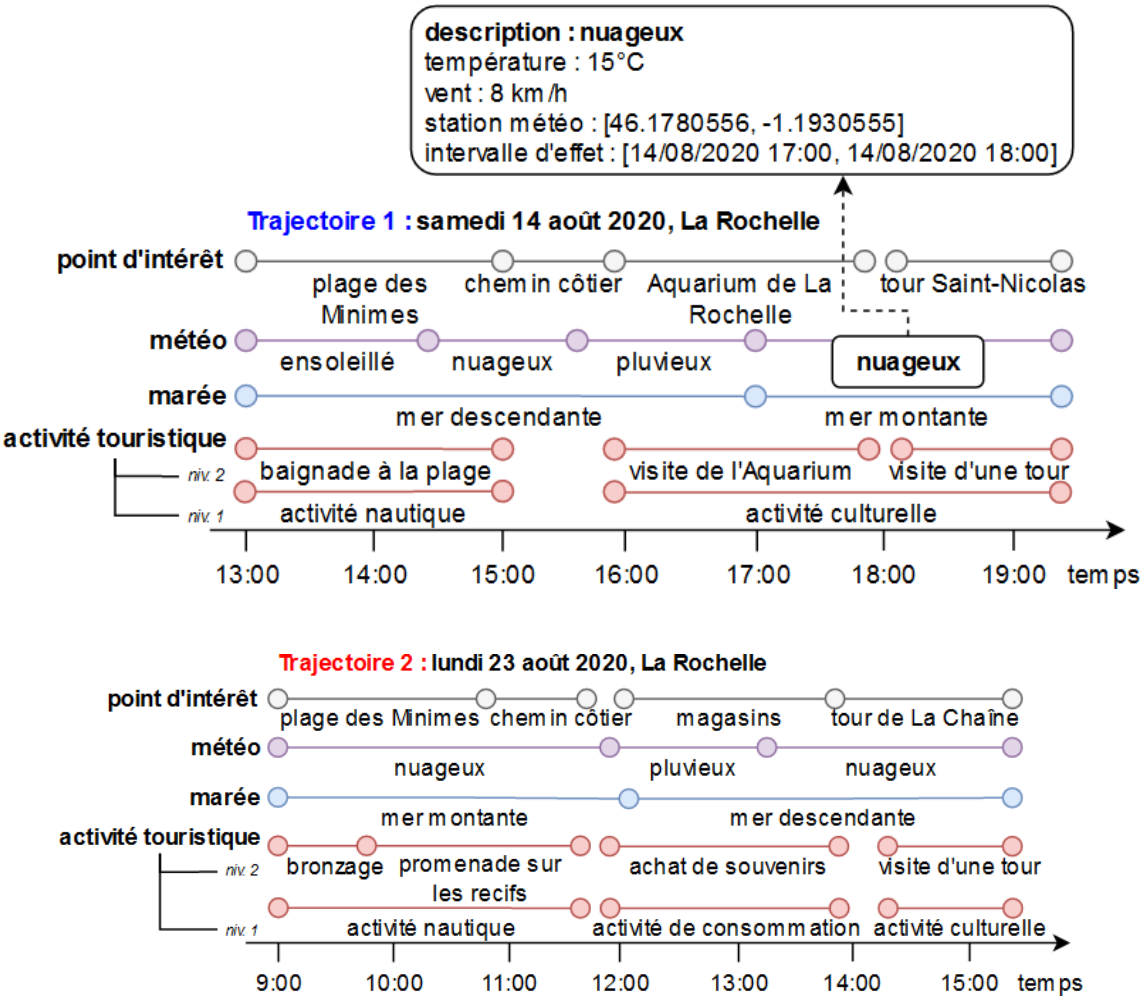


Figure 1.4 – Enrichissement des deux trajectoires sémantiques

— **Dimension temporelle (cf. figure 1.4) :** Côté dimension temporelle, les deux trajectoires se passent au mois d'août (c.-à-d. l'été). Ensuite, nous pouvons remarquer que l'une des trajectoires se passe le week-end (c.-à-d. un samedi) et l'autre en semaine (c.-à-d. un lundi). La durée des deux déplacements est de 6 heures environ mais ils ne se déroulent pas aux mêmes moments de la journée (l'un se déroule l'après-midi de 13h00 à 19h00, l'autre le matin et en début d'après-midi de 9h00 à 15h00). Ainsi, nous pouvons conclure que mis-à-part l'année, le mois et leur durée, les trajectoires sont plutôt différentes sur le plan temporel.

— **Dimension thématique (cf. figure 1.4) :** Pour finir, concernant la dimension thématique, quatre interprétations enrichissent la trajectoire (à savoir, les points d'intérêt traversés par le touriste, la météo, la marée et les activités touristiques mentionnées dans l'entretien). L'interprétation concernant les activités touristiques se décompose en plusieurs niveaux de détail. Chaque épisode d'une interprétation (p. ex. l'épisode "nuageux" de la trajectoire 1) est un aspect décrit par un ensemble d'attributs (p. ex. sa description, sa température, etc.) qui ne sont pas tous détaillés ici pour ne pas surcharger la figure. La plus longue sous-séquence partagée par deux trajectoires est appelée

plus longue séquence commune [Vlachos et al., 2002]. En considérant uniquement les points d'intérêt traversés par les touristes, la plus longue séquence commune aux deux trajectoires est {plage des Minimes, chemin côtier}. Nous pouvons aller plus loin grâce aux types (c.-à-d. l'attribut "type") des points d'intérêt, ce qui donne la plus longue séquence commune {plage, chemin, tour}. Nous pouvons également nous intéresser à toutes les interprétations en même temps, nous obtenons la plus longue séquence commune suivante : {plage des Minimes, (plage des Minimes, nuageux),
(chemin côtier, nuageux), chemin côtier, nuageux}. En observant, cette séquence, on se rend compte qu'il n'y a aucune correspondance quant à l'interprétation des marées. Mis-à-part cette différence, les trajectoires sont plutôt similaires sur le plan thématique.

Nous pouvons conclure que, même dans des contextes différents (c.-à-d. période de la journée et de la semaine, météo, marée, etc.), les profils de ces deux touristes sont assez similaires. En effet, les deux ont pratiqué des activités nautiques puis culturelles et, qui plus est, dans des lieux relativement proches (c.-à-d. plage des Minimes puis centre-ville). Ils ont aussi suivi le même itinéraire (c.-à-d. le long du port de plaisance) pendant une grande partie de leurs déplacements.

Cette comparaison manuelle est un travail fastidieux pour les géographes du projet car beaucoup de paramètres sont à prendre en compte. Un second besoin, qui ressort de nos discussions avec eux, est un outil automatique de comparaison de deux trajectoires sémantiques imitant au mieux le travail des experts.

En nous appuyant sur ce scénario de motivation, nous présentons l'objectif de la thèse.

1.1.3 Objectif de la thèse

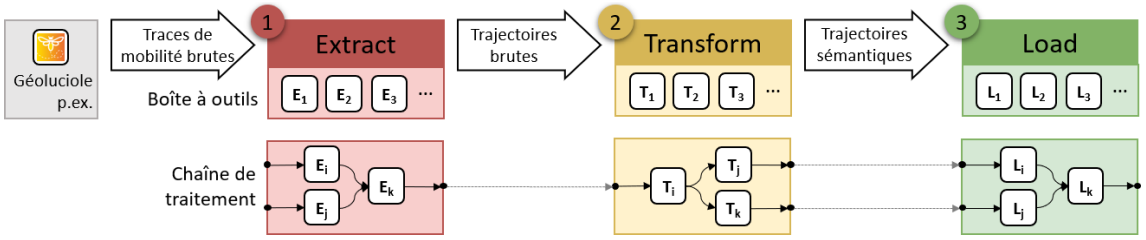


Figure 1.5 – Plateforme modulaire, type ETL

L'objectif de cette thèse est de fournir aux géographes du projet des outils logiciels et des méthodes pour les aider à analyser les traces de mobilité touristiques et à mieux comprendre les comportements des touristes en ville. Certains de ces outils existent déjà dans la littérature, d'autres nécessitent d'être adaptés à nos données et, enfin, certains n'existent pas à ce jour. Tous ces outils doivent être intégrés à une même plateforme modulaire de type **ETL** (c.-à-d. ***E**xtract, **T**ransform, **L**oad*). La figure 1.5 schématise le fonctionnement de cette plateforme. Elle doit permettre à un utilisateur de concevoir et d'exécuter ses propres chaînes de traitement à partir de modules de bas niveaux (sélectionnés depuis une boîte à

outils de modules) en vue de répondre à un questionnement de plus haut niveau sur des jeux de traces de mobilité donnés. Par l'enchaînement personnalisé et paramétré de ces modules de traitement, l'utilisateur doit pouvoir nettoyer, enrichir avec des données externes ou locales, filtrer et visualiser de différentes manières ses données de mobilité. Au fil d'une chaîne de traitement, les trajectoires brutes sont construites à partir des traces de mobilité brutes, grâce à des modules d'extraction, de nettoyage et de construction (c.-à-d. modules *Extract*), puis les trajectoires brutes sont enrichies et deviennent des trajectoires sémantiques, grâce à des modules d'enrichissement et de segmentation (c.-à-d. modules *Transform*). Les trajectoires sémantiques peuvent ensuite être visualisées et sauvegardées (c.-à-d. modules *Load*). Les modules de la plateforme ont été spécifiés dans un cahier des charges consultable en annexe (cf. annexe C). Le jeu de traces principal que nous utilisons est celui issu de l'application Géoluciole. Cependant, nous souhaitons que notre approche soit générique et puisse accepter en entrée n'importe quels autres jeux de traces de mobilité pour être utilisable dans différents contextes d'application (p. ex. traces de mobilité d'animaux, de véhicules, etc.).

Cette proposition vise plusieurs questionnements et verrous scientifiques, que ce soit sur la modélisation ou le traitement des données, présentés ci-après.

1.1.4 Verrous scientifiques

Les verrous de recherche généraux énoncés dans l'atelier Mobilités et Trajectoires du groupe de recherche MAGIS [3] ont servi de point de départ à la caractérisation de nos propres verrous. Ainsi, nous avons relevé trois verrous de l'atelier qui s'appliquent dans notre contexte :

1. l' "extraction de sémantique à partir de données issues de capteur", en lien avec les modules de traitement sur les données brutes (p. ex. module de détection des arrêts) ;
2. la conception de "méthode de confrontation et de traitement conjoint de ces données avec des données issues d'enquêtes et d'analyse de l'environnement", en lien avec la mise en relation des entretiens et des trajectoires des touristes ;
3. la "meilleure compréhension des comportements de mobilité.", en lien avec l'objectif principal du projet qui est d'arriver à mieux comprendre la mobilité touristique par l'utilisation des traces de mobilité.

En nous appuyant sur ces trois verrous de haut niveau et sur les besoins des géographes décrits dans la partie 1.1.2, nous constatons qu'il existe des verrous plus précis à lever pour arriver à notre objectif de plateforme modulaire. Rappelons et précisons les besoins pour introduire les verrous de la thèse.

Nous avons besoin d'un modèle de trajectoire sémantique pouvant représenter des traces de mobilité tout au long de leur processus de traitement (c.-à-d. de la trace de mobilité à la trajectoire sémantique). Même si notre projet est centré sur le domaine du tourisme, ce modèle doit être assez générique pour accepter tout type de traces de mobilité (p. ex. traces d'animaux, d'humains, de véhicules, etc.). La construction des trajectoires brutes à partir des traces de mobilité ne doit pas être contrainte par le modèle (que se soit temporellement et/ou

3. Site (atelier Mobilités et Trajectoires) : `https://gdr-magis.imag.fr/?page_id=83`

spatialement). Comme nous le montre la figure 1.4, l'enrichissement des trajectoires doit pouvoir se faire avec des données complexes, définies par un certain nombre d'attributs, que nous distinguons de l'enrichissement avec de simples labels. Le modèle doit aussi permettre à l'utilisateur d'enrichir la trajectoire sur plusieurs thématiques et selon un ou plusieurs niveaux de détail. La conception de ce modèle induit le premier verrou de cette thèse.

(V1) Bâtir un modèle de représentation de traces de mobilité *indoor* et *outdoor* prenant en compte plusieurs niveaux d'enrichissement avec des données complexes.

Ce verrou **(V1)** est en lien avec le verrou (2) du groupe de recherche MAGIS car un tel modèle permet de mettre en relation des données de mobilité avec des données d'enrichissement, et notamment avec des données issues d'entretiens avec les touristes.

Concernant le second besoin, nous avons comparé dans la partie 1.1.2 deux trajectoires sémantiques manuellement. Ce travail s'avère fastidieux pour les experts. Aussi, nous devons développer une mesure de similarité entre trajectoires sémantiques intégrant les trois dimensions de la trajectoire sémantique (c.-à-d. spatiale, temporelle et thématique) et imitant la manière de comparer des experts. Ainsi, nous définissons notre deuxième verrou scientifique.

(V2) Définir une mesure de similarité entre trajectoires sémantiques intégrant les dimensions spatiale, temporelle et thématique et dont les résultats s'approchent de l'avis d'experts.

Ce verrou **(V2)** est en lien avec les verrous (2) et (3) du groupe de recherche MAGIS car une telle mesure permet, d'une part, de comparer deux trajectoires en s'appuyant sur les données de mobilité et sur les données d'enrichissement de manière conjointe afin, d'autre part, de déceler des comportements de mobilité récurrents.

La plateforme modulaire est un livrable du projet DA3T et est présentée au chapitre 4. Ici, nous centrons notre mémoire sur la description des deux principales contributions de ce travail de thèse qui sont liées aux verrous **(V1)** et **(V2)**. La partie suivante énonce les hypothèses de travail sur lesquelles nous nous appuyons et dont certaines sont basées sur les verrous précédemment identifiés.

1.1.5 Hypothèses de travail

Afin de lever ces deux verrous, nous émettons des hypothèses qui guident les travaux de cette thèse. Nous partons du constat que la trajectoire sémantique, l'objet central de notre recherche, a trois dimensions (c.-à-d. dimensions spatiale, temporelle et thématique). Elles sont représentées sur la figure 1.6.

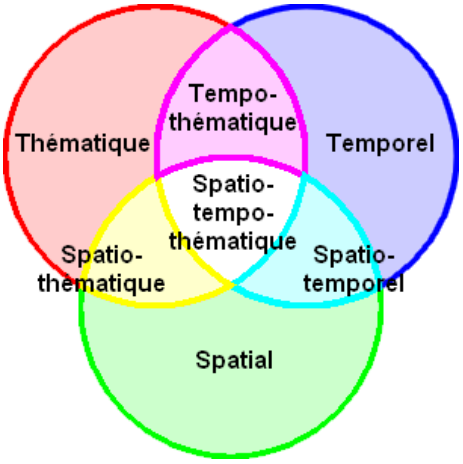


Figure 1.6 – Dimensions d'une trajectoire sémantique

La dimension spatiale d'une trajectoire GPS est une suite de coordonnées GPS, c.-à-d. une suite de paires $(longitude, latitude)$ qui représente plus ou moins fidèlement l'itinéraire emprunté par l'objet mobile. La dimension temporelle d'une trajectoire GPS est une suite d'horodatages. Chaque horodatage est lié à un point de la trajectoire ; le tout représente le déplacement de l'objet mobile observé. Enfin, la dimension thématique d'une trajectoire sémantique est un ensemble de séquences de données d'enrichissement appartenant à une certaine thématique (p. ex. météo, points d'intérêt, etc.). Nous considérons que ces trois dimensions ont la même importance et centrons toutes nos contributions autour de ces trois dimensions.

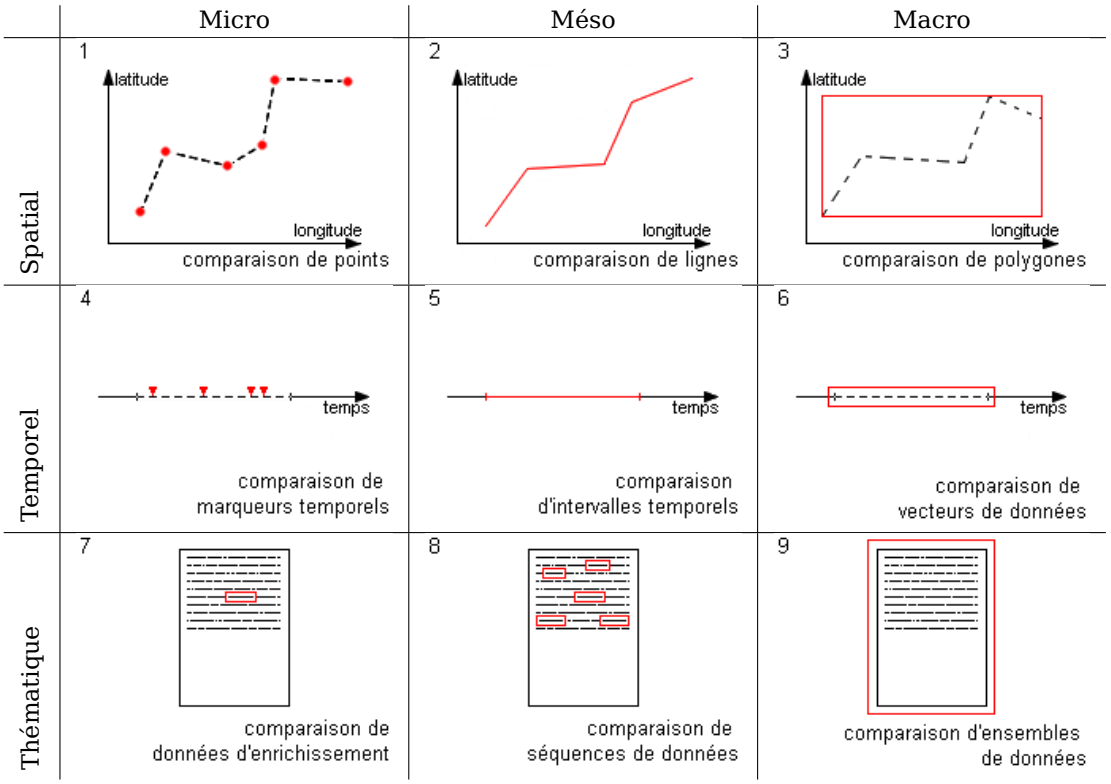
Il existe beaucoup de modèles de trajectoires sémantiques et nous les détaillons dans la partie 2.2 de l'état de l'art. Pour satisfaire les besoins des géographes et du processus d'enchaînement de modules de traitement au sein de la plateforme modulaire, nous avons combiné des caractéristiques de modèles existants dans un même modèle de trajectoire sémantique permettant d'enrichir la trajectoire grâce à des séquences de données complexes. Chaque séquence (c.-à-d. interprétation de la trajectoire) décrit un axe thématique particulier (p. ex. la météo au cours de la trajectoire, les quartiers traversés, etc.) et peut être approfondie sur plusieurs niveaux de détail (p. ex. l'interprétation des activités touristiques sur deux niveaux de la figure 1.4). La complexité des données composant les séquences se caractérise par leur définition à l'aide d'un certain nombre d'attributs et d'un type (p. ex. la donnée "nuageux" de type météo de la figure 1.4). Ainsi, il s'agit de modéliser tous les types de traces de mobilité (qu'elles soient *indoor* ou *outdoor*) et de données d'enrichissement. L'hypothèse suivante fait écho au verrou **(V1)**.

(H1) Combiner certaines caractéristiques de modèles existants dans un même modèle de trajectoire sémantique permet d'enrichir les trajectoires avec plusieurs interprétations composées de données complexes, sur plusieurs niveaux de détail.

Dans leurs travaux d'analyse, les géographes sont souvent amenés à comparer deux trajectoires entre elles. La prise en compte de toutes les dimensions des deux trajectoires dans la

comparaison est une tâche fastidieuse. Il existe des mesures de similarité qui comparent une ou plusieurs de ces dimensions mais rarement les trois simultanément. Nous détaillons ces mesures dans la partie 2.3 de l'état de l'art. Afin de comparer les trois dimensions des trajectoires sémantiques simultanément, nous avons conçu et développé deux mesures de similarité basées sur deux hypothèses.

Table 1.1 – Tableau récapitulatif des dimensions et niveaux de granularité d'une trajectoire sémantique



La première mesure consiste à combiner à l'aide de coefficients de pondération des sous-mesures de similarité spatiale, temporelle et thématique sur trois niveaux de granularité (c.-à-d. grain micro, grain méso et grain macro) chacune. Le tableau 1.1 illustre ces niveaux de granularité. Ainsi, par exemple, deux trajectoires sont comparées sur leur dimension spatiale, d'abord, au niveau de chacun des points relevés par le capteur (grain micro, cf. tableau 1.1, 1), puis, au niveau des segments correspondants (grain méso, cf. tableau 1.1, 2) et enfin, au niveau des boites englobantes (grain macro, cf. tableau 1.1, 3).

Nous supposons que cette première approche de comparaison, du grain le plus précis vers le grain le plus large, permettra d'imiter l'analyse d'un géographe. L'hypothèse suivante fait écho au verrou **(V2)**.

(H2.1) Combiner des mesures de similarité spatiale, temporelle et thématique permet de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes. L'originalité consiste à proposer différentes granularités d'observation de ces dimensions.

La seconde mesure consiste à combiner à l'aide de coefficient de pondération deux sous-mesures de similarité bidimensionnelles. Nous supposons que cette seconde approche de comparaison mettant en valeur une dimension comme dimension pivot permettra d'imiter l'analyse d'un géographe. L'hypothèse suivante fait écho au verrou **(V2)**.

(H2.2) Combiner des mesures de similarité bidimensionnelles (p. ex. spatio-temporelle et tempo-thématique) permet de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes.

Pour répondre à l'objectif d'aider des géographes, spécialistes du tourisme, à analyser un jeu de traces de mobilité brutes, nous mettons en place une plateforme modulaire de création de chaînes de traitement. Les deux hypothèses suivantes font référence à la mise en oeuvre de cette plateforme.

(H3) Dans un traitement modulaire des traces de mobilité, un enchaînement de modules de bas niveau peut permettre de répondre à un questionnement de haut niveau.

(H4) L'enrichissement d'une trajectoire touristique brute aide les géographes, experts du domaine, à mieux comprendre le déplacement d'un touriste.

Pour résumer, l'hypothèse **(H1)** implique de créer un nouveau modèle de trajectoire sémantique intégrant les caractéristiques de modèles existants. Les hypothèses **(H2.1)** et **(H2.2)** s'intéressent aux mesures de similarité entre trajectoires sémantiques. La première suppose qu'en utilisant trois niveaux de granularité pour comparer chaque dimension d'une paire de trajectoires sémantiques, les résultats de la mesure globale sont meilleurs. La seconde suppose qu'en utilisant deux sous-mesures de similarité bidimensionnelles, les résultats de la mesure globale sont meilleurs. L'hypothèse **(H3)** s'intéresse au travail de développement qui est fait dans cette thèse. Enfin, l'hypothèse **(H4)** s'intéresse à l'intégralité de ce travail et même du projet DA3T dans son ensemble.

Nous avons identifié deux contributions permettant de répondre aux deux premières hypothèses et aux verrous scientifiques mentionnés précédemment.

1.2 Contributions

Deux contributions scientifiques résultent de ce travail. La première est un modèle de trajectoire sémantique et la seconde comprend deux mesures de similarité permettant de calculer les proximités spatiale, temporelle et thématique de deux trajectoires sémantiques. Ces deux contributions sont détaillées dans le chapitre 3.

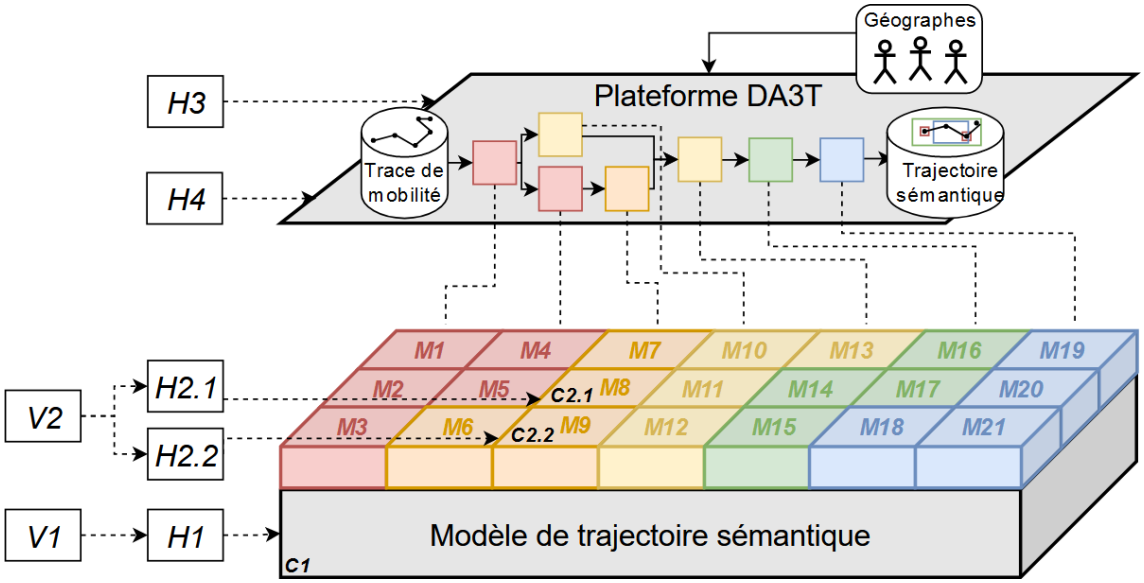


Figure 1.7 – Organisation des verrous, des hypothèses de travail et des contributions du mémoire

La figure 1.7 illustre les travaux présentés dans ce mémoire et les liens qu'ils entretiennent avec les verrous et les hypothèses exposés précédemment. L'objectif global de la thèse est la plateforme modulaire. Elle se retrouve en haut de la figure car tous les travaux décrits dans ce mémoire en découlent. Elle est la partie visible de notre travail par l'utilisateur final (c.-à-d. les géographes et les aménageurs du territoire). Le cahier des charges de la plateforme requiert des fonctionnalités qui, pour la majorité, existent déjà dans la littérature et que nous adaptons et intégrons à la plateforme sous la forme de modules de traitement de bas niveau. La conception de la plateforme s'appuie sur les hypothèses **(H3)** et **(H4)**. Notre première contribution **(C1)** porte sur un modèle de représentation des trajectoires sémantiques qui sert de socle à l'ensemble des modules. En effet, il s'agit du modèle de transition qui permet de faire circuler les données dans une chaîne de traitement conçue avec la plateforme. Cette contribution découle du verrou **(V1)** et de l'hypothèse **(H1)**. Nos deux autres contributions **(C2.1)** et **(C2.2)** proposent deux mesures de similarité intégrées à deux modules de calcul de similarité dans la plateforme. Ces contributions découlent du verrou **(V2)** et des hypothèses **(H2.2)** et **(H2.2)**.

1.2.1 (C1) Modèle de trajectoire sémantique

Notre première contribution est un modèle de trajectoire sémantique. Il permet, d'abord, de représenter des traces de mobilité. Elles peuvent être issues de différents domaines (p. ex. tourisme, animalier, sportif, etc.) et elles peuvent être *indoor* ou *outdoor*. Ensuite, il permet de représenter des données d'enrichissement qui peuvent décrire n'importe quels phénomènes du monde réel (p. ex. la météo, les points d'intérêt, les événements, etc.) dans toute leur complexité. Chaque donnée d'enrichissement à un type et un certain nombre d'attributs relatifs à ce type. Le modèle permet de représenter toutes les étapes de traitement des traces de mobilité, c.-à-d. la création des trajectoires brutes et l'enrichissement les transformant ainsi en trajectoires sémantiques. Les données d'enrichissement sont liées à des parties de

la trajectoire. Notre modèle est générique, car il accepte n'importe quel type de traces et de données d'enrichissement, et extensible, car il accepte l'ajout de nouvelles classes pour représenter plus fidèlement un nouveau domaine d'application. Nous nous sommes inspirés de caractéristiques de modèles existants pour créer notre modèle. Pour représenter les données d'enrichissement, nous nous sommes basés sur la notion d'aspect, à laquelle nous avons ajouté des attributs dimensionnels et intégré la gestion de versions. Nous avons également utilisé la notion de niveau d'épisodes pour permettre une meilleure caractérisation des épisodes et un enrichissement plus détaillé des trajectoires.

Ce modèle a fait l'objet de deux publications : la première a été acceptée à la conférence nationale IC 2021 [Cayèré et al., 2021a] et la seconde est parue dans la revue IJGI en 2021 également [Cayèré et al., 2021b]. La mise en place d'un tel modèle permet d'apporter notre première contribution au premier verrou de recherche **(V1)** en nous appuyant sur l'hypothèse de travail **(H1)**.

1.2.2 (C2) Mesures de similarité multidimensionnelles entre trajectoires sémantiques

Notre seconde contribution en comporte deux. En effet, il s'agit de deux mesures permettant d'évaluer la similarité entre deux trajectoires sémantiques de la même manière qu'un expert. La première mesure combine des sous-mesures pour les trois dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) selon trois niveaux de granularité (c.-à-d. micro, méso et macro) et les agrège grâce à des coefficients de pondération. Les valeurs des nombreux coefficients sont optimisées grâce à notre expérimentation mais l'expert peut les changer selon ses besoins. La seconde mesure combine des sous-mesures de similarité bidimensionnelles centrées sur la dimension temporelle (c.-à-d. une mesure spatio-temporelle et une mesure tempo-thématique). Comme pour la première proposition, ces mesures sont pondérées grâce à des coefficients dont les valeurs sont optimisées grâce à notre expérimentation.

Ces mesures ont fait l'objet d'une publication à la conférence nationale INFORSID 2022 [Cayèré et al., 2022]. Nous apportons ainsi deux contributions à notre second verrou de recherche **(V2)** en nous appuyant sur les hypothèses **(H2.1)** et **(H2.2)**, respectivement.

1.3 Organisation du mémoire

Ce mémoire s'organise en plusieurs chapitres. Cette partie clôture le chapitre d'introduction 1. Le chapitre suivant 2 fait l'état de l'art des modèles de trajectoires sémantiques et des mesures de similarité pouvant être appliquées aux trajectoires sémantiques. Le chapitre 3 fait part des deux contributions de la thèse, à savoir (1) un modèle de trajectoire sémantique répondant aux besoins des géographes et des aménageurs du territoire et (2) deux mesures de similarité entre trajectoires sémantiques dont l'une permet de comparer des déplacements selon trois dimensions (c.-à-d. spatiale, temporelle et thématique) et leurs trois niveaux de granularité (c.-à-d. micro, méso et macro) et la seconde permet de comparer des déplacements à travers la combinaison de deux sous-mesures de similarité bidimensionnelles.

Ensuite, le chapitre 4 détaille le processus de conception, décrit l'architecture et commente le développement de la plateforme modulaire. Le chapitre 5 présente les expérimentations mises en place pour valider nos propositions et commente les résultats. Finalement, le chapitre 6 conclut ce mémoire.

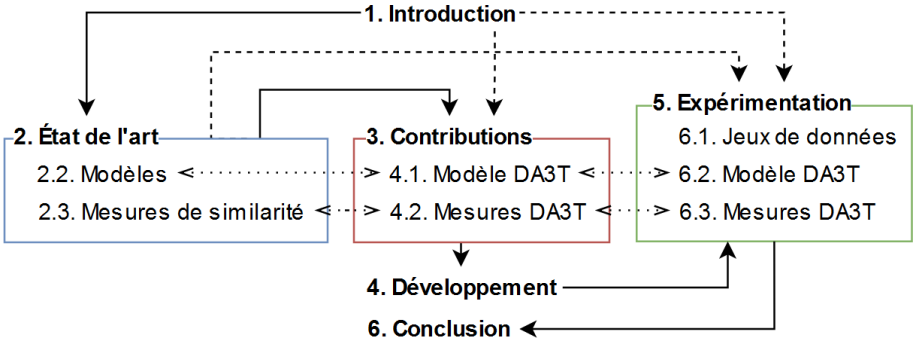


Figure 1.8 – Plan du mémoire de thèse

La figure 1.8 montre l'enchaînement linéaire des chapitres du mémoire par les flèches pleines, la possibilité de faire une lecture moins séquentielle est donnée en suivant les flèches en pointillés longs et les liens relatifs au contenu thématique des sections par des flèches bidirectionnelles en pointillés courts.

Chapitre 2

État de l'art

Ce chapitre est divisé en trois parties. La première partie (cf. partie 2.1) recense et commente, parmi les définitions formelles de la littérature, celles qui sont en lien avec notre contexte d'application et nos objectifs, telles que les traces de mobilité, les trajectoires brutes, les trajectoires sémantiques, les données d'enrichissement, etc. Les deuxième et troisième parties (cf. parties 2.2 et 2.3) font respectivement l'état de l'art des modèles de trajectoires sémantiques et des mesures de similarité applicables aux trajectoires sémantiques.

2.1 Définitions formelles sur les données de mobilité

L'évolution d'un phénomène dans un espace donné est de nature continue. Par exemple, un touriste visitant une ville ne se téléporte pas d'une position à l'autre mais il marche entre ces deux positions. Les **données spatio-temporelles** cherchent à représenter cette évolution mais due aux contraintes physiques du matériel de capture et de stockage, elle doit être discrétisée. Flouvat [2019] classe les données spatio-temporelles en quatre grandes catégories :

— Les **données d'évènements** cherchent à représenter l'historique d'évènements géolocalisés et horodatés. Ces données sont caractérisées par un ensemble de types d'évènement ainsi qu'un ensemble d'occurrences de ces types. Par exemple, le type d'évènement "festival Les Francofolies" a une occurrence géolocalisée et horodatée chaque année.

— Les **données de régions** cherchent à représenter des attributs variant de manière continue dans un espace. Plusieurs méthodes de découpage permettent de diviser l'espace en régions régulières ou irrégulières (p. ex. *rasters*, courbes de niveau, pavages de polygones, etc.). Afin de discrétiser ces données, nous considérons qu'au sein d'une même région les attributs évoluent de la même manière. Par exemple, la température de l'air est légèrement différente dans chaque point de l'espace mais, quand nous regardons la météo, nous l'obtenons pour une ville donnée.

— Les **données de réseaux** cherchent à représenter des entités géolocalisées en lien les unes avec les autres. Ces données peuvent être représentées grâce à un graphe dyna-

mique (ou graphe temporel) où les entités correspondent à des nœuds liés entre eux par des arcs pondérés ou pas. L'évolution du graphe dynamique dans le temps est décrit par une série de *snapshots* du graphe à des temps donnés [Andriamampianina et al., 2021]. Par exemple, un réseau de transport urbain peut être représenté grâce à un graphe où les arrêts sont des noeuds et les trajets sont des arcs. Si ce réseau change, alors un nouveau *snapshot* du graphe sera créé.

— Enfin, les **données de mobilité**, auxquelles nous nous intéressons dans cette thèse, décrivent le déplacement d'un objet mobile (p. ex. humain, véhicule, animal, etc.) au cours du temps. Nous en détaillons les particularités dans la suite de cette partie.

Généralement, les données de mobilité d'un objet mobile se présentent sous la forme de **trace de mobilité** (ou de mouvement). Une trace de mobilité peut s'apparenter à une **série temporelle**, c'est-à-dire une séquence de valeurs échantillonnées à des temps spécifiques Magdy et al. [2015]. Nous nous appuyons sur la définition de Parent et al. [2013] pour définir la trace de mobilité telle que :

Définition 2.1.1 (Trace de mobilité)**.** Une trace de mobilité représente l'intégralité du déplacement capturé pour un objet mobile donné. Une trace de mobilité Tm de taille n est définie par un tuple tel que :

$$Tm = (o, \{p_0, p_1, ...p_n\})$$

Avec o l'identifiant de l'objet mobile, suivi par une liste de n positions spatio-temporelles où p_i est défini par un tuple tel que :

$$p_i = (t_i, l_i, d_i)$$

Avec t_i une donnée temporelle (p. ex. instant ou intervalle temporel), l_i une donnée spatiale (p. ex. point, ligne ou polygone) et d_i un ensemble de données additionnelles capturées en même temps que la localisation (p. ex. la vitesse de déplacement, l'altitude, l'orientation, etc.).

Une trace de mobilité n'est pas toujours intéressante dans son intégralité pour une application donnée. Par exemple, si nous souhaitons observer le déplacement des touristes en centre-ville, une trace représentant le déplacement d'un touriste sur son séjour complet n'est pas entièrement utile. Seules les parties de la trace où il est effectivement au centre-ville sont intéressantes. Nous qualifions une partie intéressante de la trace de mobilité de **trajectoire brute**. La définition de trajectoire brute suivante se fonde également sur la définition proposée par Parent et al. [2013] :

Définition 2.1.2 (Trajectoire brute)**.** Une trajectoire brute est une partie de la trace de mobilité qui a de l'intérêt pour une application donnée. Une trajectoire brute Tb d'une trace de mobilité de taille n est définie par un tuple tel que :

$$Tb = (o, \{p_{start}, ...p_{end}\})$$

Avec $start \geq 0$ et $end \leq n$.

Une trajectoire brute ne représente que le déplacement brut d'un objet mobile. Afin de rendre ce déplacement plus compréhensible et explicable, la notion de **trajectoire sémantique** apparaît. Il s'agit d'utiliser des données provenant de l'*Open Data*, de bases de données spécialisées, d'avis d'experts, d'interviews, de résultats de calcul, etc. pour enrichir sémantiquement la trajectoire brute. Cet enrichissement peut être fait grâce à l'**annotation** (manuelle ou automatique) et/ou à la **segmentation** de la trajectoire. Les définitions suivantes sont inspirées de celles données par Parent et al. [2013] :

Définition 2.1.3 (Annotation)**.** L'annotation de la trajectoire est le fait d'attacher des données d'enrichissement à la trajectoire, à une partie de la trajectoire ou à une position de la trajectoire. Une donnée d'enrichissement attachée de cette manière à la trajectoire s'appelle une annotation.

Définition 2.1.4 (Segmentation)**.** La segmentation de trajectoire est le fait de découper une trajectoire en épisodes selon un ou plusieurs critères de segmentation. Ce découpage aboutit à une liste d'épisodes appelée interprétation de la trajectoire.

Il est possible de segmenter une trajectoire avec un critère basé sur des annotations ou d'annoter des épisodes issu d'une segmentation donnée. Un certain nombre de travaux s'intéressent à formaliser le concept de trajectoire sémantique [Yan et al., 2011] [Parent et al., 2013] [Bogorny et al., 2014] [Flouvat, 2019] [Mello et al., 2019] [Noureddine et al., 2020]. Nous nous appuyons sur ces travaux pour définir ce concept tel que :

Définition 2.1.5 (Trajectoire sémantique)**.** Une trajectoire sémantique est une trajectoire brute enrichie à l'aide de données d'enrichissement par le biais d'annotations et/ou de segmentations. Une trajectoire sémantique Ts est définie par un tuple tel que :

$$Ts = (o, \{p_0, p_1, ...p_n\}, \{I_0, ...I_m\}, a)$$

Avec a l'ensemble des annotations enrichissant l'intégralité de la trajectoire, la liste des positions de la trajectoire $p_i = (t_i, l_i, d_i, a_i)$ où a_i est l'ensemble des annotations enrichissant la position d'index i et, enfin, la liste des interprétations de la trajectoire tel que $I_j = (t_0, a_0), ...(t_k, a_k)$ avec (t_j, a_j) un épisode où a_j est l'ensemble des annotations enrichissant la partie de la trajectoire délimitée par l'intervalle temporel t_j.

Dépendant du modèle et des approches, l'enrichissement peut se faire de différentes façons ou avec différents types de données d'enrichissement (p. ex. simples labels ou objets complexes). Nous détaillons les modèles de trajectoires sémantiques dans la partie suivante.

2.2 Modèles de trajectoire sémantique

Cette partie a pour but d'expliquer et de classifier les modèles de trajectoire sémantique issus de la littérature.

2.2.1 Introduction

Les travaux de Albanna et al. [2015] proposent une classification des modèles de trajectoire sémantique à travers quatre types d'approche, à savoir : (i) la modélisation basée sur les types de données qui propose de créer des types de données pour représenter les trajectoires sémantiques, (ii) la modélisation basée sur les patrons de conception (en anglais, *design pattern*) qui consiste à représenter les composants basiques communs à chaque trajectoire (p. ex. les arrêts, les déplacements, etc.) à l'aide d'objets liés entre eux avec des relations, et laisse au développeur la tâche d'ajouter les données d'enrichissement spécifiques à chaque application, (iii) la modélisation utilisant les ontologies qui propose de représenter les trajectoires sous la forme de concepts spatiaux, temporels et thématiques à l'aide d'ontologies spatiale, temporelle et de domaines et (iv) la modélisation hybride qui propose trois modèles différents représentant trois niveaux d'abstraction de la trajectoire sémantique (c.-à-d. le modèle de données brutes, le modèle conceptuel découpant la trajectoire en séquence d'épisodes et le modèle sémantique). Arslan et al. [2018] classe également les modèles de trajectoire sémantique selon cette catégorisation.

Une autre manière de catégoriser les modèles est de s'intéresser d'une part aux modèles ontologiques et d'autre part à tous les autres modèles conceptuels [Nogueira et al., 2018], ce qui donne deux grands groupes qui correspondent aux classes (ii) et (iii) de la classification de Albanna et al. [2015].

Contrairement à ces deux classifications, nous n'allons pas classifier les modèles selon leur type (p. ex. ontologique, conceptuel, etc.) car ce n'est pas un critère propre à la modélisation des trajectoires sémantiques et il peut s'appliquer à la classification de modèles appartenant à des domaines variés. Nous allons nous appuyer sur la manière, qu'ont ces modèles, de segmenter et d'enrichir les trajectoires. Dans un premier temps, nous parlons des modèles de trajectoire sémantique segmentant la trajectoire en une séquence d'épisodes d'arrêt et de déplacement avant de les enrichir (cf. partie 2.2.2). Ensuite, nous discutons des modèles, plus génériques, découpant la trajectoire en épisodes selon un prédicat donné (cf. partie 2.2.3). Nous parlons ensuite des modèles permettant de réaliser plusieurs segmentations de la trajectoire pour l'enrichir, que nous appelons modèles multi-interprétation (cf. partie 2.2.4). Nous abordons un autre cas de modèles multi-interprétation qui propose d'introduire une hiérarchie entre les épisodes (cf. partie 2.2.5). Enfin, nous terminons sur le cas particulier d'un modèle multi-aspect qui ne rentre dans aucune des précédentes catégories mais qui présente des caractéristiques intéressantes (cf. partie 2.2.6). Afin de présenter notre classification, pour chaque catégorie, nous décrivons par un schéma le modèle de trajectoire sémantique par un schéma et analysons les divers travaux scientifiques s'y rapportant.

2.2.2 Modèles basés sur les arrêts et les déplacements

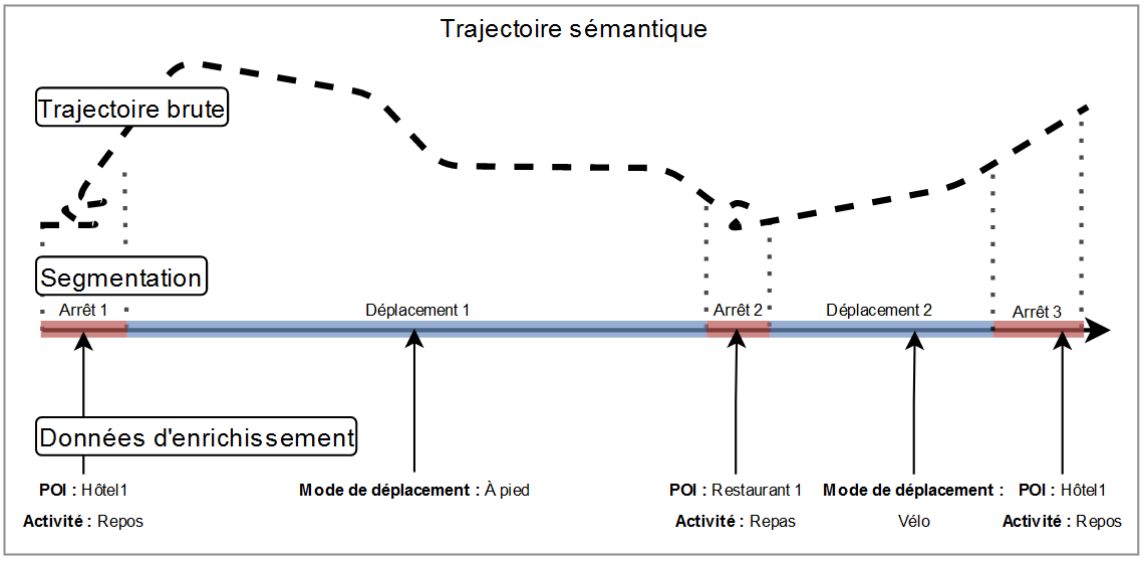


Figure 2.1 – Modèle basé sur une segmentation arrêt/déplacement

La figure 2.1 montre une trajectoire segmentée selon les arrêts et les déplacements de l'objet mobile et enrichie avec des données relatives à ces arrêts et déplacements. Les premiers modèles de données représentant les trajectoires sémantiques étaient basés sur cette segmentation.

Spaccapietra et al. [2008] présentent un modèle conceptuel de données pour représenter les trajectoires sémantiques. Ce modèle segmente la trajectoire en temps d'arrêt et temps de déplacement au sein de cette trajectoire (en anglais, *stops* et *moves* de la trajectoire). Les définitions d'arrêt et de déplacement proposées dans Spaccapietra et al. [2008] présentent les caractéristiques génériques des arrêts et des déplacements mais restent assez larges pour correspondre à des applications diverses. Nous les utilisons pour donner les définitions suivantes :

Définition 2.2.1 (Arrêt)**.** Un arrêt est une partie de la trajectoire dont la temporalité est définie par un intervalle temporel non-vide durant lequel l'objet mobile reste statique selon l'application en question. Les intervalles temporels de deux arrêts consécutifs sont toujours disjoints.

Définition 2.2.2 (Déplacement)**.** Un déplacement est une partie de la trajectoire dont la temporalité est définie par un intervalle temporel non-vide situé entre deux arrêts consécutifs, le début de la trajectoire et le premier arrêt ou la fin de la trajectoire et le dernier arrêt.

Prenons l'exemple d'un touriste visitant une ville. Lorsqu'il arrive à la place du marché, il décide de regarder les étals. Selon l'échelle d'étude et ce que nous souhaitons observer, la place du marché entière peut être considérée comme un arrêt, ou bien chacun des étals

du marché auprès desquels s'arrête le touriste peuvent être considérés comme des arrêts (entrecoupés de phases de déplacement pour circuler d'un étal à l'autre). De plus, selon l'application, certains arrêts peuvent être pertinents alors que d'autres non. Si l'intérêt de l'application porte sur la détection des points d'intérêt où le touriste s'est arrêté, alors un arrêt contraint par une foule n'est pas pertinent pour l'analyse de ces trajectoires.

Une fois la trajectoire segmentée, Spaccapietra et al. [2008] proposent aussi d'annoter les épisodes d'arrêt et de déplacement avec des propriétés variables ou non-variables au cours du temps. Dans ce modèle, il est possible d'ajouter des informations sémantiques à l'objet mobile, à la trajectoire globale, à chaque arrêt et déplacement mais il est impossible d'annoter les points car ce modèle s'abstrait totalement de la suite de points brute. Le modèle a été éprouvé sur des données concernant la migration de cigognes.

Alvares et al. [2007] se basent sur le modèle de Spaccapietra et al. [2008] et y ajoutent la notion d'arrêts candidats. Une application possède une liste d'arrêts candidats, chacun étant défini par une géométrie (i.e. un polygone spatial) et par un temps d'arrêt minimum. Un algorithme de détection d'arrêts est proposé avec ce modèle. Lorsqu'une trajectoire croise l'un des polygones représentant un arrêt candidat et y reste au moins pendant le temps d'arrêt minimum alors la section de la trajectoire qui traverse ce polygone est considéré comme un arrêt de la trajectoire. Cela nécessite, pour chaque application, de construire au préalable sa liste d'arrêts candidats. Le principal manque de cet algorithme réside dans la non détection des arrêts qui ne sont pas définis au préalable par l'application, ce qui ne permet pas de découvrir de l'information. Des données issues de touristes visitant des points d'intérêt ont servi d'exemples pour ce modèle.

Baglioni et al. [2008] utilisent la même définition de trajectoire que Spaccapietra et al. [2008] et proposent une ontologie décrivant une trajectoire sémantique à travers les arrêts et les déplacements de l'objet mobile. Cette ontologie a été mise en application sur des trajectoires de personnes dans le but de caractériser leur comportement de déplacement. Pour enrichir ces trajectoires, ils s'appuient sur des connaissances d'experts exprimées sous la forme d'axiomes permettant d'effectuer des raisonnements automatiques sur l'ontologie.

Yan et al. [2008] proposent un modèle ontologique pour représenter la trajectoire sémantique. Afin de faciliter sa maintenance et sa conception mais aussi d'optimiser les mécanismes de requête, l'ontologie est décomposée en trois modules différents : (i) l'ontologie de la trajectoire géométrique contient les concepts génériques décrivant une trajectoire (p. ex. les arrêts, les déplacements, les concepts temporels, les concepts spatiaux, etc.), (ii) l'ontologie géographique contient les concepts génériques et issus du domaine d'application décrivant l'environnement géographique (p. ex. la topographie et les réseaux du territoire) et (iii) l'ontologie du domaine d'application contient les concepts propres au domaine d'application (p. ex. les points d'intérêt touristiques pour une application liée au domaine du tourisme). Ce modèle a été mis en application dans le domaine de la gestion du trafic routier.

Dans le modèle présenté dans Frihida et al. [2009], la trajectoire sémantique est représentée grâce à un type abstrait de données. Ce modèle est dépendant de l'application car

la trajectoire sémantique se présente sous la forme de voyages vers des activités (c.-à-d. des déplacements et des arrêts). Les voyages peuvent soit être simples, quand l'intérêt est uniquement porté sur l'activité ciblée et qu'il sont représentés par des lignes allant d'une activité à une autre, soit riches, lorsque les voyages sont représentés par des sous-trajectoires. Ce modèle à été mis en pratique avec des requêtes dans un contexte d'activités journalières de personnes.

Pour conclure, le principal inconvénient des modèles basés sur une segmentation en épisodes d'arrêt et de déplacement est que l'enrichissement est contraint à ces épisodes. Pourtant, il est possible qu'une donnée d'enrichissement change au cours d'un épisode. Par exemple, un touriste peut pratiquer des activités différentes dans un même lieu d'arrêt ; il peut bronzer puis jouer au frisbee sur une plage. Pour remédier à cet inconvénient, il existe des modèles plus génériques que nous présentons dans les parties suivantes.

2.2.3 Modèles basés sur des épisodes

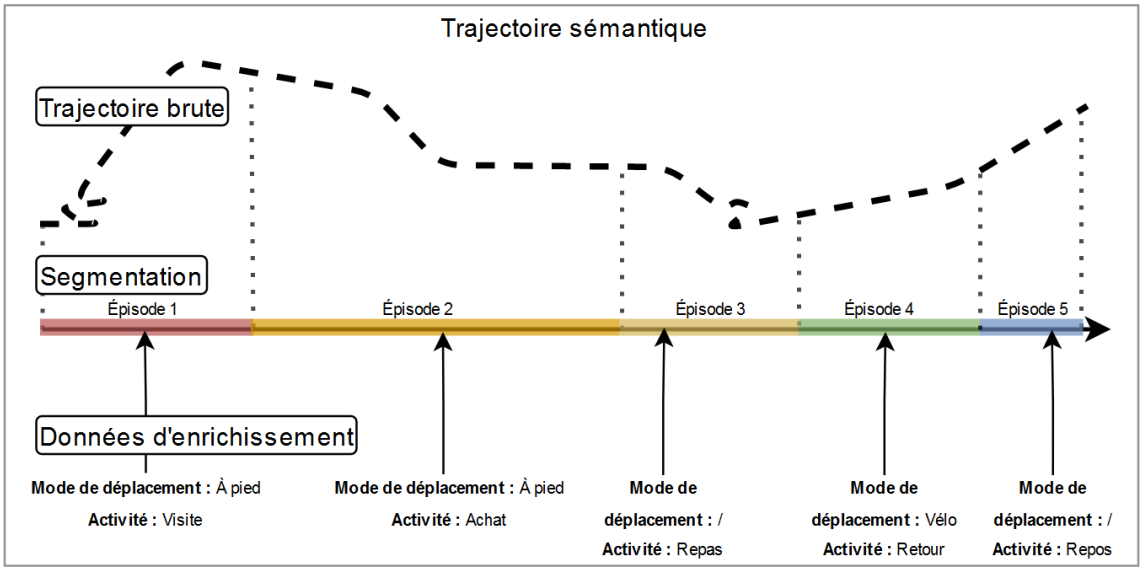


Figure 2.2 – Modèle basé sur une segmentation en épisodes

La figure 2.2 montre une trajectoire segmentée en épisodes et enrichie avec des données relatives à ces épisodes. Les modèles basés sur les épisodes sont une généralisation de ceux basés sur les arrêts et les déplacements.

Yan et al. [2011] proposent un modèle de trajectoire sémantique dans le contexte d'un *framework* appelé **SeMiTri** (***Se****mantic* ***Mi****ddleware for* ***Tr****ajectori****es***). Ce modèle consiste à représenter les trajectoires sémantiques à travers des séquences d'épisodes, appelée interprétation. Nous nous appuyons sur les définitions d'épisode et d'interprétation données par Yan et al. [2011] pour donner les définitions suivantes :

Définition 2.2.3 (Épisode)**.** Un épisode est une partie de la trajectoire dont la temporalité est définie par l'ensemble des positions spatio-temporelles consécutives répondant à un prédicat donné.

Définition 2.2.4 (Interprétation)**.** Une interprétation de la trajectoire est une liste d'épisodes basée sur un ou plusieurs prédicats. À chaque changement de valeurs des prédicats, un nouvel épisode est créé.

Un prédicat peut porter sur les positions spatio-temporelles (p. ex. construction d'épisodes d'arrêt et de déplacement basée sur la vitesse et le rayon de déplacement) ou des données d'enrichissement (p. ex. construction d'épisodes portant sur les activités touristiques pratiquées basée sur les données collectées pendant un entretien). Chaque ensemble d'annotations d'un même type (p. ex. annotations météorologiques, relatives aux points d'intérêt, relatives aux activités, etc.) permet de construire un ensemble d'épisodes en segmentant la trajectoire à chaque fois que la valeur de l'annotation change. Une liste d'épisodes basée sur des annotations d'un même type est une interprétation de la trajectoire. Dans leur chaîne de traitement, Yan et al. [2011] proposent de passer de la trajectoire brute (c.-à-d. séquence de positions géocalisées et horodatées représentant le déplacement d'un objet mobile au cours du temps) à la trajectoire sémantique (c.-à-d. séquence de positions géocalisées et horodatées enrichies avec des annotations). Puis, il proposent de transformer la trajectoire sémantique en **trajectoire sémantique structurée**, définie comme une représentation de la trajectoire sémantique uniquement décrite par une séquence d'épisodes annotés et localisés dans l'espace grâce à des objets spatiaux (c.-à-d. polygones, lignes ou points). Cette notion s'abstrait complètement de la définition initiale de la trajectoire brute car elle ne considère plus les positions géolocalisées et horodatées. L'application de ce modèle se fait grâce au *framework* SeMiTri. Cependant, malgré la généricité du modèle, SeMiTri se base sur une segmentation standard en épisodes d'arrêt et de déplacement. Selon la définition de trajectoire sémantique structurée, il relie ces épisodes à des objets géographiques de types point (pour les arrêts) et ligne (pour les déplacements) et les annote avec des données relatives aux activités pratiquées (pour les arrêts) et aux modes de transport (pour les déplacements). Finalement, le modèle de trajectoire sémantique, basé sur une segmentation en épisodes, présenté dans ce travail, propose une vision de la trajectoire sémantique plus générique que les modèles basés uniquement sur les arrêts et déplacements. Il a pour objectif d'être indépendant des applications et d'accepter des trajectoires hétérogènes. Il a été éprouvé sur des trajectoires de véhicules et de personnes.

Bogorny et al. [2014] présente un modèle conceptuel de représentation des trajectoires sémantiques, appelé **CONSTAnT** (***CON****ceptual model of* ***S****emantic* ***T****r****A****jec****T****ories*), qui propose de segmenter la trajectoire en **sous-trajectoires sémantiques** (c.-à-d. en épisodes). Dans ce travail, chaque trajectoire sémantique a un but général (p. ex. visiter une ville) et chaque sous-trajectoire a un but plus spécifique participant au but général (p. ex. aller au musée). En plus de son but, une sous-trajectoire peut être décrite par le comportement de l'objet mobile lors du déplacement et son moyen de transport. Chaque point des sous-trajectoires peut être décrit par des informations sur l'environnement ainsi que sur sa localisation et les évé-

nements s'y déroulant. Une limitation de ce modèle concerne les données qui enrichissent la trajectoire. Elles sont limitées aux types de données décrits dans le modèle, à savoir, le but du déplacement, le comportement de l'objet mobile, le moyen de transport, les informations relatives à l'environnement, au lieu et aux évènements (liés aux lieux). Par exemple, nous disposons de la retranscription des entretiens avec les touristes et il est courant qu'ils émettent des commentaires sur ce qu'ils ont pensé de tel ou tel endroit de la ville. Il s'agit d'une information importante pour le géographe qui souhaite comprendre le comportement des touristes. Or, ce type de données ne trouve pas sa place dans ce modèle. De plus, un autre problème concerne les déplacements qui n'ont aucun but général. Dans notre cas d'usage, un touriste peut visiter une ville pendant un temps puis aller faire des courses dans un même déplacement. Dans cet exemple, il faudrait segmenter le déplacement en deux trajectoires de manière arbitraire car il est impossible de connaître le moment exact où le touriste a pris la décision d'aller faire ses courses. Les notions de but spécifique et de but général sont très subjectives. Ce modèle a été testé sur une application touristique et une application concernant la migration des oiseaux pour montrer sa généricité avec deux jeux de données très différents.

Le modèle présenté dans Moreau et al. [2018] s'appuie sur une partie du modèle CONSTAnT de Bogorny et al. [2014]. Il représente la trajectoire sémantique avec une séquence d'activité issue d'une ontologie. Les activités sont soit statiques, soit mobiles, mais contrairement aux modèles basés sur les arrêts et les déplacements, l'alternance entre les deux n'est pas obligatoire. Une activité mobile est décrite par un mode de déplacement (p. ex. à pied, en vélo, en voiture, etc.) également issue d'une ontologie dédiée. Une activité peut éventuellement avoir lieu dans un endroit spécifique avec un nom, une géométrie et un type. Les types de lieu sont eux aussi issus d'une ontologie. Enfin, le modèle permet également de représenter la météo au cours du déplacement dont la description est elle aussi issue d'une ontologie spécialisée. Ainsi, utiliser des ontologies dédiées pour décrire des informations sémantiques permet de faciliter la comparaison entre deux trajectoires sémantiques.

Noureddine et al. [2020] présente un modèle permettant de représenter à la fois les trajectoires sémantiques *indoor* et *outdoor*. Leur approche consiste à segmenter les trajectoires selon des critères spatiaux et thématiques simultanément (pour former une seule séquence d'épisodes). Un critère de segmentation spatial est basé sur les points d'intérêt traversés par les trajectoires. Ils sont géolocalisés et organisés dans une hiérarchie spatiale qui unifie les espaces intérieurs et extérieurs. Cette hiérarchie va des plus petits niveaux de granularité d'un espace intérieur (p. ex. un tableau dans une pièce, elle-même dans un musée, etc.) aux plus hauts niveaux de granularité en extérieur (p. ex. un quartier dans une ville, elle-même dans un pays, etc.). Un critère de segmentation thématique est basé sur les valeurs prises par un élément thématique donné (p. ex. la valeur "nuageux" pour décrire la météo).

Pour conclure, l'inconvénient des modèles basés sur une segmentation en épisodes est qu'ils ne proposent qu'une seule perspective d'étude de la trajectoire avec une segmentation basé sur des types de données souvent contraints.

2.2.4 Modèles multi-interprétation

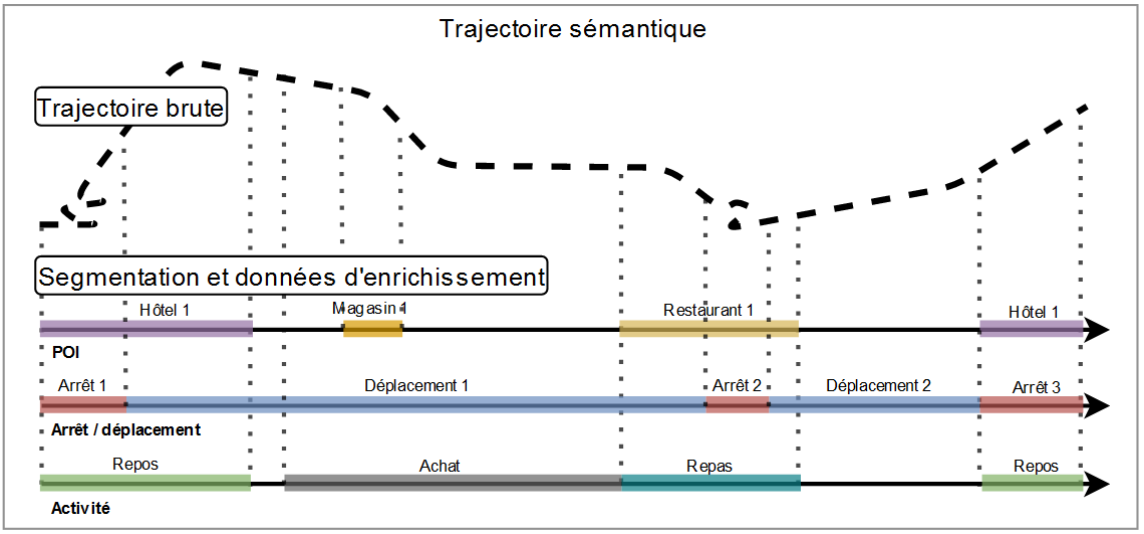


Figure 2.3 – Modèle basé sur plusieurs segmentations

La figure 2.3 montre une trajectoire segmentée en plusieurs séquences d'épisodes, chacune basée sur un prédicat donné (p. ex. POI pour point d'intérêt, arrêt et déplacement, activité, mode de déplacement, etc.).

Nogueira and Martin [2015] proposent le modèle ontologique **STEP** (***S**emantic **T**rajectory **EP**isodes*), dérivé de l'ontologie QualiTraj [Nogueira and Martin, 2014][Nogueira et al., 2014], pour décrire la trajectoire sémantique à l'aide d'un certain nombre de séquences d'épisodes, chacune s'intéressant à une **variable d'intérêt** (en anglais, *feature of interest*) particulière (p. ex. la vitesse de déplacement, la température ambiante, etc.). La valeur d'une variable d'intérêt peut être quantitative (c-à-d. une valeur d'un type de base accompagnée d'une unité) ou qualitative (c.-à-d. une valeur issue d'une autre ontologie ou d'une énumération). Cette valeur évolue au fil de la trajectoire et chaque changement donne lieu à un nouvel épisode sémantique. Un épisode peut posséder une portée spatiale, une portée temporelle ou une portée spatio-temporelle (en anglais, *spatial, temporal and spatio-temporal extent*) qui permet de spécifier des limites spatiales et/ou temporelles (p. ex. un épisode décrivant une température ambiante de 18°C a une portée temporelle égale à la durée durant laquelle il faisait cette température). Il s'agit d'un modèle qui se veut générique dans lequel les données d'enrichissement ne sont pas contraintes. Ce modèle a été mis en pratique dans un contexte de trajectoires de coureurs à pieds et des requêtes, permettant de répondre à des questions spécifiques (p. ex. Quelle était la météo durant la course ? Comment la vitesse du coureur a évolué durant la course ?), ont été exécutées sur l'instanciation du modèle. Un inconvénient de ce modèle est qu'un épisode est forcément construit à partir d'une seule et unique valeur d'intérêt (p. ex. un unique épisode dont les caractéristiques sont "ensoleillé" et "aquarium de La Rochelle" est impossible à décrire avec ce modèle)

2.2.5 Modèles multi-interprétation et multi-niveau

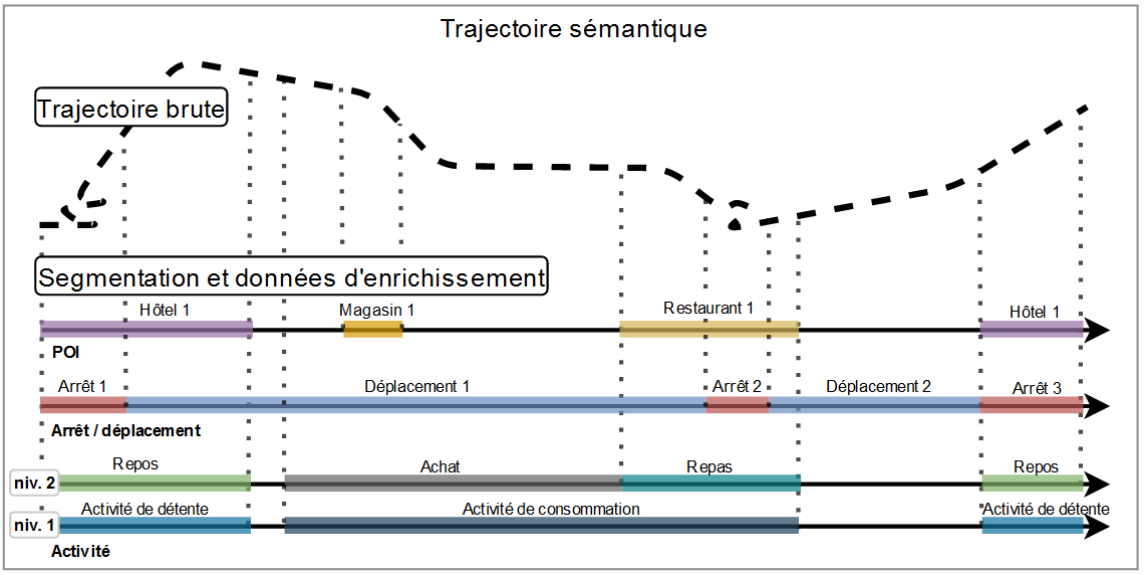


Figure 2.4 – Modèle basé sur plusieurs segmentations dont les épisodes peuvent être détaillés sur plusieurs niveaux de détail

La figure 2.4 montre une trajectoire segmentée en plusieurs séquences d'épisodes, chacune basée sur un type de données et pouvant être détaillée sur plusieurs niveaux.

Fileto et al. [2015] proposent un modèle ontologique, appelé **Baquara**[2], extension de **Baquara** [Fileto et al., 2013]. Ce travail redéfinit les notions de donnée de mobilité, de trace de mobilité et de trajectoire. Une **séquence de positions d'un objet mobile** (abrégé MOPS pour *Moving Object's Position Sequence*) représente l'historique de déplacement connu d'un objet mobile durant une certaine période de temps et est décrite par une séquence de positions spatio-temporelles. Cette description correspond à notre définition de trace de mobilité. Un **segment de mouvement** (abrégé MS pour *Movement Segment*) est une abstraction d'une sous-séquence continue d'une MOPS. Cette description généralise nos définitions de trajectoire et d'épisode. Chaque MS possède sa propre géométrie (qui est une abstraction du segment de MOPS initial), un certain nombre d'annotations et un ensemble de propriétés chronologiques (p. ex. previous indiquant le segment précédent, next indiquant le segment suivant, etc.) et hiérarchiques (p. ex. father indiquant le segment parent, level indiquant le niveau de profondeur du segment, etc.). Ce modèle permet d'ordonner mais surtout de hiérarchiser les MS : un MS peut être détaillé en plusieurs autres. Les MS peuvent être annotés par des valeurs numériques ou textuelles libres mais aussi avec des concepts (c.-à-d. des classes p. ex. "bar") ou des objets (c.-à-d des instances de classes p. ex. "Bar l'Aragon") tirés de différentes facettes d'analyse. Les facettes d'analyse couvertes par Baquara[2] sont les mêmes que celle du modèle CONSTAnT : Space, Time, Goal, Behavior, TransportationMeans, EnvironmentCondition, Activity, MovingObject et Event. Le modèle a été testé sur des données spatio-temporelles issues de Flickr et de Twitter et enrichies à l'aide des données de DBpedia et LinkedGeoData.

Nogueira et al. [2018] proposent une nouvelle version de l'ontologie **STEP** [Nogueira and Martin, 2015] dans laquelle l'une des évolutions les plus importantes est que les épisodes peuvent être organisés de manière hiérarchique afin de permettre différents niveaux de détail.

2.2.6 Cas particulier des modèles multi-aspect

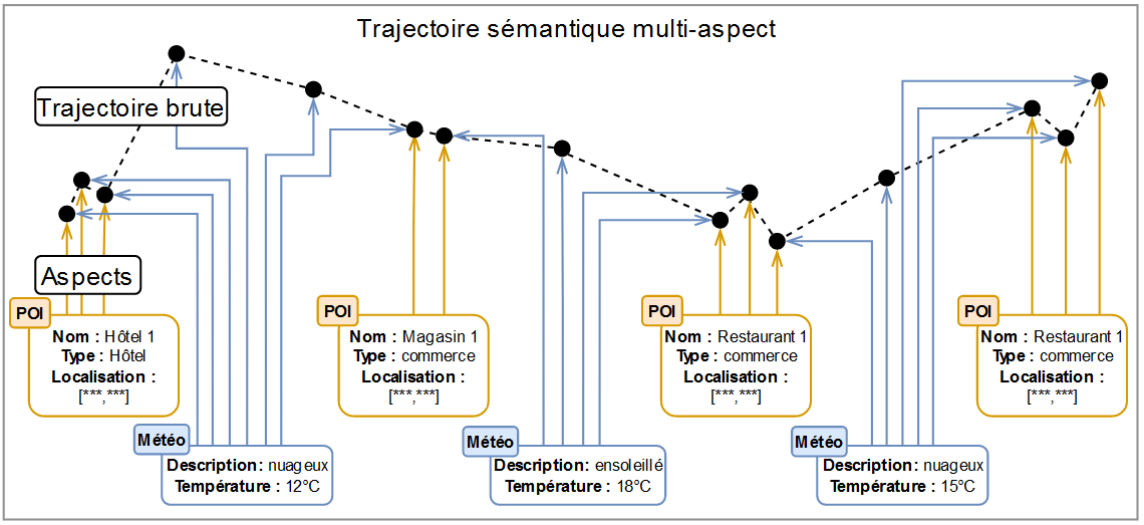


Figure 2.5 – Modèle basé sur un enrichissement avec des aspects

La figure 2.5 montre une trajectoire dont les points sont enrichis avec des aspects.

Mello et al. [2019] introduisent le modèle **MASTER** dans lequel le concept de **trajectoire multi-aspects** est une généralisation du concept de trajectoire sémantique basé sur la notion d'aspect. La définition suivante s'inspire de la définition d'aspect donnée par Mello et al. [2019] :

Définition 2.2.5 (Aspect)**.** Un aspect est un phénomène du monde réel qui est pertinent pour l'analyse des trajectoires. Il est caractérisé par un type et par un ensemble d'attributs reliés à ce type. De plus, le type peut être un sous-type d'un type plus général.

Un aspect peut être lié à l'ensemble de la trajectoire, à un point de la trajectoire, à l'objet mobile lui-même ou à une relation entre deux objets mobiles, et peut contenir tout type de données. Les concepts d'épisode ou de sous-trajectoire n'existent pas dans le modèle MASTER ce qui nécessite d'annoter chaque position lorsque l'aspect est valable pendant une partie de la trajectoire. Or, les annotations sur un épisode sont beaucoup plus efficaces que les annotations sur chaque point d'une trajectoire [Yan et al., 2011]. Les aspects possèdent un ou plusieurs types (p. ex. l'aspect avec pour nom « Hôtel La Marine » est de type « hôtel ») et il existe une notion de hiérarchie entre les types (p. ex. le type « hôtel » est un sous-type du type « hébergement » qui est lui-même un sous-type du type « point d'intérêt »). Chaque type possède un certain nombre d'attributs qui sont instanciés pour chaque aspect de ce type. Le modèle a subi une évaluation qualitative afin de tester les mécanismes de requête sur une

instanciation fictive d'une visite d'un touriste à Paris, et une évaluation quantitative afin de tester la solution de stockage des données adoptée.

2.2.7 Synthèse

Les modèles de représentation des trajectoires sémantiques ont été conçus afin de représenter au mieux la réalité d'un déplacement. Le but est de lier toutes les données d'enrichissement pertinentes selon l'application et les experts afin de faciliter l'explication d'un tel déplacement. Les premiers modèles de trajectoire sémantique les représentaient en matière de séquences d'épisodes d'arrêt et de déplacement, car c'est une segmentation qui semble, au premier abord, intuitive. Certains modèles proposent d'annoter librement ces épisodes d'arrêt et de déplacement, d'autres contraignent ces annotations (p. ex. des points d'intérêt pour les arrêts, des modes de déplacement pour les déplacements). Or, ce type de modèle est limité car il ne permet pas une segmentation personnalisée en fonction de l'application et une représentation des données d'enrichissement multiples qui peuvent expliquer un déplacement. De plus, au cours d'un arrêt donné, les données d'enrichissement associées peuvent changer, tout comme lors d'un déplacement [Noureddine et al., 2020]. Cela souligne un manque de généricité dans ce type de segmentation. Ces modèles ont évolué vers un type de modèles plus génériques et permissifs basés sur la segmentation en épisodes. La trajectoire est segmentée de n'importe quelle façon selon un prédicat donné. Certains modèles permettent d'effectuer plusieurs segmentations sur différents critères et d'autres permettent de hiérarchiser les épisodes. Enfin, nous nous sommes intéressés à un type de modèle qui enrichit la trajectoire avec des objets complexes, appelés aspects, qui permettent de se rapprocher un peu plus du réel qu'avec de simples annotations.

Référence	**Segmentation**	**Conservation des données brutes**	**Enrichissement libre**	**Multi-interprétation**	**Multi-niveau**	**Multi-aspect**	**Support des données *in-door* et *out-door***
Spaccapietra et al. [2008]	Arrêts et déplacements	✗	✓	✗	✗	✗	✗
Alvares et al. [2007]	Arrêts et déplacements	✗	✓	✗	✗	✗	✗
Baglioni et al. [2008]	Arrêts et déplacements	✗	✓	✗	✗	✗	✗
Yan et al. [2008]	Arrêts et déplacements	✗	✓	✗	✗	✗	✗
Frihida et al. [2009]	Arrêts et déplacements	✗	✗	✗	✗	✗	✗
Yan et al. [2011]	Épisodes	✓	✓	✗	✗	✗	✗
Bogorny et al. [2014]	Épisodes	✓	✗	✗	✗	✗	✗
Moreau et al. [2018]	Épisodes	✓	✗	✗	✗	✗	✗
Noureddine et al. [2020]	Épisodes	✗	✓	✗	✗	✗	✓
Nogueira and Martin [2015]	Épisodes	✓	✓	✓	✗	✗	✗
Fileto et al. [2015]	Épisodes	✓	✓	✓	✓	✗	✗
Nogueira et al. [2018]	Épisodes	✓	✓	✓	✓	✗	✗
Mello et al. [2019]	/	✓	✓	✓	✗	✓	✗

Table 2.1 – Synthèse des modèles de trajectoire sémantique

Le tableau 2.1 fait la synthèse de cette partie 2.2 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre modèle dans chaque modèle présenté précédemment. Chaque colonne correspond à une caractéristique d'intérêt, extraite de des besoins des géographes et de la plateforme, et chaque ligne à un modèle présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans un modèle. Les caractéristiques qui nous intéressent sont les suivantes :

— **Conservation des données brutes** : Les modèles ayant cette caractéristique ne s'abstraient pas entièrement des données brutes ; ils permettent de représenter les traces de mobilité brutes et leurs positions. Notre modèle servant de modèle de transition entre des modules de traitement, la trace de mobilité doit pouvoir être représentée à n'importe quelle étape de son traitement.

— **Enrichissement libre** : Ces modèles permettent d'enrichir les trajectoires avec n'importe quelles données. Ces données ne sont pas contraintes par des classes spécifiques. Selon le contexte de l'application, les données d'enrichissement peuvent être radicalement différentes et il est donc important de ne pas les contraindre à une application spécifique car nous souhaitons concevoir un modèle générique.

— **Multi-interprétation** : Les modèles multi-interprétation permettent plusieurs segmentations des trajectoires pour créer des interprétations sur différents axes thématiques (p. ex. météo, points d'intérêt, événement, etc.) laissés au choix de l'application. Pour comprendre le déplacement d'un touriste, les géographes veulent pouvoir faire plusieurs interprétations de la trajectoire en s'appuyant sur des données d'enrichissement différentes.

— **Multi-niveau** : Les modèles multi-niveau permettent un enrichissement sur plusieurs niveaux de détails. Ces niveaux sont intéressants dans notre contexte pour détailler certaines données d'enrichissement (p. ex. les activités touristiques).

— **Multi-aspect** : Les modèles multi-aspect utilisent des aspects pour enrichir les trajectoires, c.-à-d. des objets complexes pouvant représenter tout phénomène du monde réel. Nous souhaitons que notre modèle soit multi-aspect afin de représenter les données d'enrichissement le plus fidèlement possible.

— **Support des données *indoor* et *outdoor*** : Ces modèles peuvent représenter aussi bien les trajectoires *indoor* que *outdoor*. Dans notre projet, nous travaillons sur ces deux types de trajectoires (c.-à-d. trajectoires de touristes dans une ville et trajectoires de visiteurs dans un musée) qui doivent pouvoir être modélisés grâce à un modèle commun.

Nous constatons qu'aucun des modèles de notre état de l'art ne présente l'ensemble des caractéristiques d'intérêt. Aussi, notre contribution sera de concevoir un modèle de représentation de trajectoire sémantique qui présente toutes ces caractéristiques. Pour cela, nous nous inspirons de certains modèles présentés dans cette partie.

Dans la partie suivante, nous présentons les mesures de similarité pouvant être utilisées pour comparer des trajectoires sémantiques sur leurs dimensions spatiale, temporelle et/ou thématique.

2.3 Mesures de similarité entre trajectoires sémantiques

Cette partie a pour but d'expliquer et de classifier les mesures de similarité pouvant être utilisées pour comparer deux trajectoires sémantiques.

2.3.1 Introduction

Une mesure de similarité (à l'inverse, mesure de distance) permet d'attribuer un score de ressemblance (à l'inverse, de différence), appelé similarité (à l'inverse, différence), à une paire d'objets de même nature pour les comparer. Plus la similarité est élevée, plus les deux objets sont similaires (à l'inverse, plus la distance est élevée, plus les deux objets sont différents). Par exemple, la distance entre deux nombres peut être représentée par la valeur absolue de leur différence. Nous définissons la distance et similarité telles que :

Définition 2.3.1 (Distance/Similarité)**.** La distance (à l'inverse, la similarité) entre deux objets A et B représente l'éloignement (à l'inverse, la ressemblance) de ces deux objets. Elle est mesurée en utilisant une mesure de distance d (à l'inverse, de similarité s).

Un score produit, pour un couple d'objets donné, par une mesure de distance ou de similarité donnée est une valeur numérique. Cette valeur dépend, à la fois, de la façon dont fonctionne la mesure et de la ressemblance des deux objets comparés. Il est parfois nécessaire de normaliser les scores produits par différentes mesures sur un intervalle $[0, 1]$ afin de pouvoir les combiner, les comparer, etc. De plus, la normalisation permet de faciliter le passage d'un score de similarité normalisé à un score de distance normalisé (et inversement). Ainsi, dans ce contexte, il suffit de soustraire le résultat de l'une à 1 pour arriver au résultat de l'autre, tel que l'équation 2.1 le suggère.

$$s = 1 - d \text{ et } d = 1 - s \tag{2.1}$$

Nous distinguons trois dimensions comparables dans les trajectoires sémantiques. La dimension spatiale d'une trajectoire sémantique est définie par l'ensemble de coordonnées géographiques représentant l'itinéraire de l'objet mobile. Une autre dimension, presque indissociable de la dimension spatiale lorsqu'on parle de trajectoire, est la dimension temporelle car à chaque coordonnée géographique est associé un horodatage. Enfin, la dimension thématique de la trajectoire sémantique donne du sens au déplacement à l'aide de différents axes thématiques (c.-à-d. des interprétations) tels que la météo, les quartiers traversés, les activités pratiquées, etc. Deux trajectoires peuvent être comparées selon une (p. ex. elles peuvent être comparées seulement sur la dimension spatiale) ou plusieurs de ces dimensions, séparément (p. ex. elles peuvent être comparées dans un premier temps sur la dimension spatiale et dans

un second temps, sur la dimension temporelle) ou de manière combinée (p. ex. elles peuvent être comparées sur la dimension spatiale et sur la dimension temporelle simultanément).

Dans les exemples de mesures présentés dans cet état de l'art, nous utiliserons les lettres R et S pour parler de deux trajectoires sémantiques à comparer, respectivement de taille m et n. Pour plus de lisibilité dans les formules présentées, nous avons simplifié la représentation mathématique donnée dans la partie 2.1 en retirant l'identifiant de la trajectoire, la liste des interprétations et les annotations sur la trajectoire entière. Ainsi, ces trajectoires sont définies telles que :

$$R = \{r_1, r_2, ...r_m\} \text{ et } S = \{s_1, s_2, ...s_n\}$$

Où r_i et s_i sont les positions de R et S. Une position r_i est un tuple tel que :

$$r_i = (t_i, (x_i, y_i), d_i, a_i)$$

Où t_i est le moment de la capture, (x_i, y_i) représente les coordonnées GPS de la position, d_i est l'ensemble de données complémentaires enregistrées au moment de la capture (p. ex. la vitesse de déplacement, la précision de la capture, etc.) et, enfin, a_i est l'ensemble des annotations de la positions.

En mathématique, les mesures de distance peuvent être métriques ou non-métriques. Une mesure de distance métrique satisfait les conditions suivantes [Chen et al., 2009] :

— **Non-négativité :** $d(R, S) \geq O$;
— **Unicité :** $d(R, S) = 0 \Leftrightarrow R = S$;
— **Symétrie :** $d(R, S) = d(S, R)$;
— **Inégalité triangulaire :** $d(R, S) + d(S, T) \geq d(R, T)$.

Dans le cas de mesures normalisées, les mesures de similarité peuvent également être métriques ou non-métriques. Une mesure de similarité métrique satisfait les conditions suivantes [Chen et al., 2009] :

— **Non-négativité :** $s(R, S) \geq 0$;
— **Unicité :** $s(R, S) = 1 \Leftrightarrow R = S$;
— **Symétrie :** $s(R, S) = s(S, R)$;
— **Inégalité triangulaire :** $s(R, S) + s(S, T) \leq s(R, T) + 1$.

Les matrices de similarité (ou de distance) [Cleasby et al., 2019] sont parfois utilisées pour calculer le score de similarité (ou de distance) entre deux trajectoires ou, de manière plus abstraite, entre deux séries temporelles. Cette méthode propose de calculer les scores de similarité (ou de distance) des éléments des séquences avant de calculer le score final. Une telle matrice est de taille $m * n$ et chaque cellule contient le score de similarité (ou de distance) d'une paire d'éléments. Par exemple, la cellule C_{ij} contient le score de similarité (ou de distance) des éléments R_i et S_j. Le remplissage de la matrice dépend de la mesure utilisée.

Dans cette partie de notre état de l'art, nous allons revenir sur les mesures de similarité

ou de distance existantes pour comparer des objets spatiaux, temporels et/ou thématiques. Il existe plusieurs travaux s'intéressant à comparer et classifier les mesures de similarité de trajectoires [Wang et al., 2013, Magdy et al., 2015, Cleasby et al., 2019, Su et al., 2020, Tao et al., 2021], cependant la dimension thématique est souvent omise, n'étant pas la dimension centrale de description du déplacement d'un objet mobile. Prenons pour exemples les classifications de Magdy et al. [2015] présentée sur la figure A.1 en annexe et de Su et al. [2020] présentée sur la figure A.2 en annexe. Ces deux classifications catégorisent les mesures de similarité en deux grands groupes, à savoir, les mesures de similarité spatiale (c.-à-d. des mesures s'intéressant uniquement à la dimension spatiale, les trajectoires comparées sont traitées comme des séquences de positions) et les mesures de similarité spatio-temporelle (c.-à-d. des mesures s'intéressant aux dimensions spatiale et temporelle). Magdy et al. [2015] classent les mesures de similarité spatiale en trois sous-catégories, (1a) celle des mesures basées sur les données spatiales, (1b) celle des mesures basées sur les formes géométriques des trajectoires et (1c) celle des mesures basées sur la direction du mouvement, et classent les mesures de similarité spatio-temporelle en deux sous-catégories, (2a) celle des mesures basées sur la vitesse du mouvement et (2b) celle des mesures basées sur les séries temporelles. Su et al. [2020] classent les mesures de similarité spatiale et spatio-temporelle en deux catégories chacune, celle des mesures discrètes (c.-à-d. considérant les points seulement) et celles des mesures continues (c.-à-d. considérant les points ainsi que les mouvements entre les points). Premièrement, nous remarquons qu'il n'y a pas de catégorie concernant la dimension thématique des trajectoires. Ces travaux se concentre uniquement les mesures de similarité des trajectoire brutes. Nous constatons également une différence sur leur manière de classifier les mesures dédiées aux séries temporelles : Magdy et al. [2015] les classent dans les mesures spatio-temporelles, car l'ordre des éléments est considéré comme faisant partie de la dimension temporelle, alors que Su et al. [2020] les classent dans les mesures spatiales.

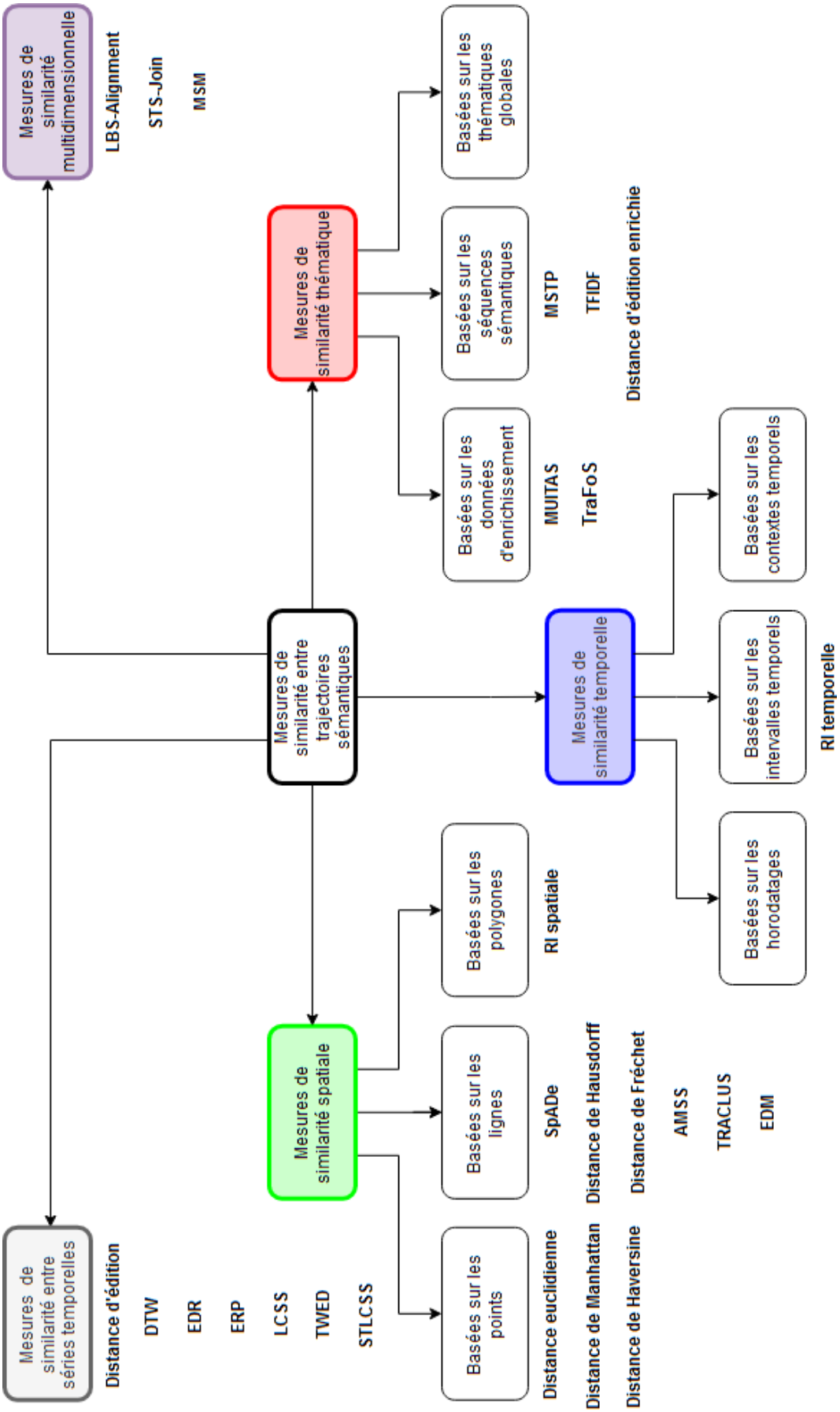


Figure 2.6 – Classification des mesures de similarité dimensionnelles

La figure 2.6 montre notre classification des mesures de similarité. À la différence des deux classifications présentées précédemment, nous avons une catégorie pour les mesures de similarité thématique ainsi qu'une catégorie pour les mesures de similarité multidimensionnelle. Il est important de prendre en compte ces catégories de mesures car nous souhaitons comparer des trajectoires sémantiques. De plus, plutôt que de classer les mesures de similarité entre séries temporelles avec les mesures spatiales ou spatio-temporelles, nous en avons fait une classe à part. En effet, ces mesures peuvent être utilisées pour comparer les trajectoires sémantiques sur différentes dimensions. Concernant les sous-catégories des mesures de similarité spatiale, temporelle et thématique, nous nous sommes appuyés sur les niveaux de granularité représentés dans le tableau 1.1. Certaines sous-catégories sont vides car nous n'avons trouvé aucune mesure y correspondant (c.-à-d. mesures de similarité temporelle basées sur les horodatages, mesures de similarité temporelle basées sur les contextes temporels et mesures de similarité thématique basées sur les thématiques globales).

Pour classifier les mesures de similarité de l'état de l'art, nous nous appuyons sur la ou les dimensions ciblées par les mesures et sur les classifications présentées plus haut (cf. figure 2.6). Premièrement, nous discutons des mesures de similarité spatiale (cf. partie 2.3.2). Ensuite, nous parlons des mesures de similarité temporelle (cf. partie 2.3.3). Puis, nous abordons les mesures de similarité thématique (cf. partie 2.3.4). Pour finir, nous abordons le sujet des mesures dédiées à la comparaison des séries temporelles (cf. partie 2.3.5) puis des mesures de similarité multidimensionnelle (cf. partie 2.3.6).

2.3.2 Mesures de similarité spatiale

Lorsqu'il est question d'attribuer un score de similarité à deux objets spatiaux, le calcul de distance physique entre ces objets est souvent utilisé. Nous souhaitons calculer la similarité spatiale entre deux trajectoires sémantiques GPS. La dimension spatiale d'une telle trajectoire est une suite de coordonnées GPS, c.-à-d. une suite de paires $(longitude, latitude)$ qui représente plus ou moins fidèlement l'itinéraire emprunté par l'objet mobile. La précision d'une trajectoire spatiale est influencée par plusieurs facteurs comme les spécificités et la qualité du matériel de capture, la qualité du signal GPS, la fréquence d'échantillonnage, etc.

Dans cette partie, nous allons étudier les différentes mesures de similarité (et de distance) de la littérature qui peuvent être utilisées ou adaptées pour comparer deux trajectoires sur leur dimension spatiale. Nous classons les mesures de similarité spatiale (cf. figure 2.6, case verte) en trois catégories. Premièrement, (i) celles qui s'appuient sur une comparaison point à point pour calculer le score global. Deuxièmement, (ii) celles qui s'appuient sur une comparaison ligne à ligne. Troisièmement, (iii) celles qui s'appuient sur une comparaison polygone à polygone.

(i) Mesures de similarité basées sur la comparaison des points

Le calcul de la distance spatiale entre deux trajectoires demande parfois de calculer les distances entre les points qui les composent. Pour calculer la distance entre deux points, il est possible d'utiliser la distance euclidienne (c.-à-d. L_2norm) ou la distance de Manhattan (c.-à-d. L_1norm). Soient deux points A et B tels que A est défini par les coordonnées (A_1, A_2) et B

est défini par les coordonnées (B_1, B_2). La distance entre ces deux points peut être calculée grâce à la formule 2.2.

$$L_p norm(A, B) = \sqrt[p]{\sum_{i=1}^{n} |A_i - B_i|^p} \tag{2.2}$$

Où p permet de spécifier quelle distance est utilisée (p. ex. $p = 1$ représente la distance de Manhattan et $p = 2$, la distance euclidienne) et n représente le nombre de coordonnées définissant un point.

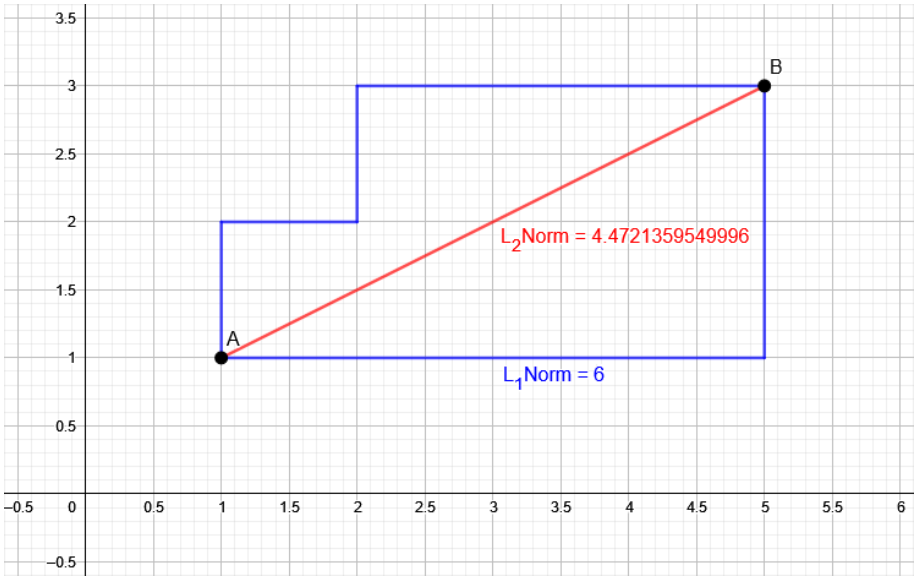


Figure 2.7 – Distance de Manhattan et distance euclidienne entre deux points

La figure 2.7 montre comment sont calculées la distance de Manhattan et la distance euclidienne entre deux points $A(1, 1)$ et $B(5, 3)$. La première correspond à la distance indirecte (c.-à-d. distance s'appuyant sur des déplacements verticaux et horizontaux comme dans les rues quadrillées de Manhattan) et la seconde correspond à la distance directe (c.-à-d. distance à vol d'oiseaux) qui sépare ces deux points.

La **distance euclidienne** [Faloutsos et al., 1994] est une mesure de distance métrique. Elle peut être appliquée sur les points d'un espace euclidien à une dimension (p. ex. sur des éléments de séries temporelles) ou plusieurs dimensions (p. ex. sur des points de trajectoires) en utilisant l'équation 2.3.

$$euclidiean(A, B) = L_2 norm(A, B) = \sqrt{\sum_{i=1}^{n} |A_i - B_i|^2} \tag{2.3}$$

Pour mesurer la distance entre deux trajectoires dans un espace euclidien, il est possible d'utiliser la distance euclidienne entre les points correspondants des deux trajectoires (distance entre le i-ème point d'une trajectoire avec le i-ème point de l'autre trajectoire) puis d'additionner toutes les distances calculées. C'est la distance euclidienne à étapes bloquées (*lock-step euclidean distance*) [Tao et al., 2021].

Cependant, la distance euclidienne fonctionne dans l'espace euclidien mais pas dans l'espace sphérique sur des points possédant des coordonnées GPS. Pour calculer la distance entre deux points GPS, il existe la formule de Haversine utilisée pour la première fois en 1805 dans les travaux de Andrew [1805]. Soient deux points A et B placés à la surface de la terre tels que A est défini par les coordonnées GPS (A_{lon}, A_{lat}) et B est défini par les coordonnées (B_{lon}, B_{lat}). La distance entre ces deux points peut être calculée grâce à la formule 2.4 :

$$haversine(A,B) = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{B_{lat}-A_{lat}}{2}\right) + \cos(A_{lat})\cos(B_{lat})\sin^2\left(\frac{B_{lon}-A_{lon}}{2}\right)}\right) \tag{2.4}$$

Où r est le rayon de la sphère (p. ex. le rayon de la Terre si nous travaillons sur des trajectoires terrestres).

Nous pouvons également ranger dans cette catégorie les mesures de similarité entre séries temporelles (cf. partie 2.3.5) qui peuvent s'adapter pour comparer la dimension spatiale de deux trajectoires.

(ii) Mesures de similarité basées sur la comparaison des lignes

D'autres mesures de similarité (ou de distance) spatiale se basent sur la division des trajectoires en segments ou en sous-trajectoires qu'elles comparent deux à deux.

Chen et al. [2007] proposent une nouvelle mesure de similarité entre deux séries temporelles continues, appelée **SpADe** (***S**patial **A**ssembling **D**istance*). Les deux séries temporelles comparées sont découpées en deux ensembles de motifs locaux lp de taille identique. La distance entre deux motifs locaux correspond à la somme pondérée des différences entre l'amplitude et la forme de deux motifs. Lorsque cette distance est inférieure à un certain seuil η, nous parlons de correspondance entre les deux motifs. Une matrice de correspondance est créée et représente les correspondances entre les motifs locaux des deux séries. Les éléments de la matrice sont des segments construits à partir de la projection des deux motifs locaux concernés. Mis les uns à la suite des autres, ces segments forment des chemins. L'ensemble des chemins reliant l'élément $(1,1)$ à l'élément (m,n) de la matrice est contenu dans un ensemble P. La mesure de distance SpADe est définie par l'équation 2.5.

$$SpADe(R,S) = min_{c_i \in C}\{cost(c_i)\} \tag{2.5}$$

Où $cost(c_i)$ est une fonction calculant la longueur d'un chemin, en l'occurrence le chemin c_i.

Dans les travaux de Alt [2009], les distances de Hausdorff et Fréchet sont présentées. Ces deux distances s'intéressent à évaluer la différence entre deux formes géométriques.

Intéressons-nous d'abord à la **distance de Hausdorff**. Cette distance attribue un score en fonction de la similarité de deux formes. La mesure de la distance de Hausdorff est définie

par l'équation 2.6.

$$hausdorff(R,S) = max\{hausdorff\_dir(R,S), hausdorff\_dir(S,R)\} \tag{2.6}$$

Où $hausdorff\_dir$ est une fonction appelée distance de Hausdorff dirigée qui permet d'évaluer à quel point une forme R est proche d'une partie d'une forme S. Le résultat est nul lorsque R est une sous-partie de S. Elle est définie par l'équation 2.7.

$$hausdorff\_dir(R,S) = max_{i=1}^{m}\{min_{j=1}^{n}\{||r_i - s_j||\}\} \tag{2.7}$$

La **distance de Fréchet**, quant à elle, permet de comparer deux courbes. Elle vise à minimiser la distance maximale qui relient les points des courbes. La mesure de distance de Fréchet est définie par l'équation 2.8.

$$frechet(R,S) = inf_{\sigma,\tau}\{max_{t\in[0,1]}\{||R_{\sigma(t)} - S_{\tau(t)}||\}\} \tag{2.8}$$

Où σ et τ englobent toutes les fonctions strictement monotones croissantes et continues.

Nakamura et al. [2013a] présentent la mesure de similarité **AMSS** (pour ***A**ngular **M**etric for **S**hape **S**imilarity*). Ces travaux proposent de transformer les séries temporelles à comparer en séquences de vecteurs. Ainsi, nos trajectoires R et S correspondent aux séquences de vecteurs Rv et Sv, de tailles respectives $m' = m - 1$ et $n' = n - 1$, définies telles que :

$$\begin{aligned} Rv &= ((r_2.x, r_2.y) - (r_1.x, r_1.y)), ((r_3.x, r_3.y) - (r_2.x, r_2.y)), ...((r_m.x, r_m.y)) - (r_{m'}.x, r_{m'}.y))) \\ &= (rv_1, rv_2, ..., rv_{m'}) \end{aligned}$$

$$\begin{aligned} Sv &= ((s_2.x, s_2.y) - (s_1.x, s_1.y)), ((s_3.x, s_3.y) - (s_2.x, r_2.y)), ...((s_n.x, s_n.y)) - (s_{n'}.x, s_{n'}.y))) \\ &= (sv_1, sv_2, ..., sv_{n'}) \end{aligned}$$

Où rv_i et sv_i sont des vecteurs particuliers. Deux vecteurs peuvent être comparés grâce à la distance euclidienne ou à la mesure de similarité cosinus. Ici, chaque paire de vecteurs est comparée grâce à la similarité cosinus. Cette dernière est utilisée lorsque les séquences sont comparées sur leurs formes géométriques plutôt que les tailles des vecteurs ou les positions exactes des points. Une matrice de similarité est construite de taille $m * n$. Le score de similarité entre les deux séries temporelles est défini par le chemin le moins coûteux allant de l'élément $(1,1)$ à l'élément (m,n) de la matrice, où le coût d'un chemin est égal à la somme des éléments consécutifs. La mesure AMSS est définie par l'équation 2.9.

$$AMSS(Rv,Sv) = max\left\{\begin{array}{l} 2sim(rv_{m'}, sv_{n'}) + AMSS(Rv_{m'-1}, Sv_{n'-1}) \\ 2sim(rv_{n'-1}, sv_{m'}) + sim(rv_{n'}, sv_{m'}) + AMSS(Rv_{m'-2}, Sv_{n'-1}) \\ 2sim(rv_{n'}, sv_{m'-1}) + sim(rv_{m'}, sv_{n'}) + AMSS(Rv_{m'-1}, Sv_{n'-2}) \end{array}\right\} \tag{2.9}$$

Où la similarité entre deux vecteurs est définie par l'équation 2.10 qui correspond à la

similarité cosinus.

$$sim(rv_i, sv_j) = cosinus\_similarity(rv_i, sv_j) = \begin{cases} 0 & \text{si } \theta > \frac{\pi}{2} \\ cos(\theta) \ = \frac{rv_i . sv_j}{|rv_i||sv_j|} & \text{sinon} \end{cases} \tag{2.10}$$

Où θ est l'angle entre les deux vecteurs r_i et s_j. Lorsque cet angle est trop grand, il n'est pas considéré afin d'éviter qu'il ait trop d'influence sur la valeur de similarité globale.

Chen et al. [2004] proposent de s'appuyer sur une nouvelle représentation des trajectoires pour les comparer. Une trajectoire est définie par sa séquence de mouvement, une séquence de paires $(movement\_direction, distance\_ratio)$ où $movement\_direction$ représente avec un angle la direction du mouvement entre deux points de la trajectoire (valeur comprise entre $-\pi$ et π) et $distance\_ratio$ représente le ratio de la distance du segment formé par les deux points par rapport à celles de tous les autres segments formés par les autres points de la trajectoire (valeur comprise entre 0 et 1). La représentation **MPS** (pour ***M**ovement **P**attern **S**tring*) s'appuie sur une carte de quantification (avec la direction du mouvement en abscisses et le ratio de distance en ordonnées). Cette carte est découpée en régions, chacune étant associée à un symbole particulier (p. ex. a, B, 1, +, etc.) . Ainsi, chaque élément de la séquence peut être associé à un symbole de la carte de quantification et la trajectoire entière peut être représentée grâce à une séquence de symboles. Nos trajectoires R et S correspondent respectivement aux MPS $Rm = (rm_1, rm_2, ..., rm_{m'})$ et $Sm = (sm_1, sm_2, ..., sm_{n'})$ de taille $m' = m - 1$ et $n' = n - 1$ et avec rm_i et sm_i des symboles particuliers. **EDM** (pour ***E**dit **D**istance on **MPS***) est une mesure de distance permettant de comparer deux MPS. EDM est définie par l'équation 2.11.

$$EDM(Rm, Sm) = \begin{cases} m' & \text{si n' = 0} \\ n' & \text{si m' = 0} \\ EDM(rm_{m'-1}, sm_{n'-1}) & \text{si } rm_{m'-1} \text{ et } sm_{n'-1} \text{ sont des symboles égaux ou voisins} \\ min \left\{ \begin{array}{l} EDM(rm_{m'-1}, sm_{n'-1}) + 1 \\ EDM(rm_{m'-1}, sm_{n'}) + 1 \\ EDM(rm_{m'}, sm_{n'-1}) + 1 \end{array} \right\} & \text{sinon} \end{cases} \tag{2.11}$$

Lee et al. [2007] présentent une mesure de similarité utilisée dans leur algorithme **TRACLUS** (pour ***TRA**jectory **CLUS**tering*). Tout d'abord, ils considèrent les trajectoires comme des séquences de segments (c.-à-d. paires de deux points consécutifs) et leur distance permet de comparer non pas deux trajectoires mais deux segments. La mesure de distance est composée de trois sous-mesures qui permettent chacune d'évaluer la distance entre les segments selon une caractéristique en particulier. Les caractéristiques de comparaison sont le parallélisme, la distance et l'angle des deux segments. La mesure de distance est définie par

l'équation 2.12.

$$TRACLUS\_distance(L_i, L_j) = \omega_\perp * distance_\perp(L_i, L_j) + \omega_\parallel * distance_\parallel(L_i, L_j) + \omega_\theta * distance_\theta(L_i, L_j) \quad (2.12)$$

Avec L_i et L_j les deux segments comparés et $\omega_\perp$, $\omega_\parallel$ et ω_θ les coefficients de pondération des différents composants de la distance globale. En ce qui concerne les trois sous-mesures de distance, elles sont définies par les équations 2.13, 2.14 et 2.15 dont les différentes variables sont représentées sur la figure 2.8.

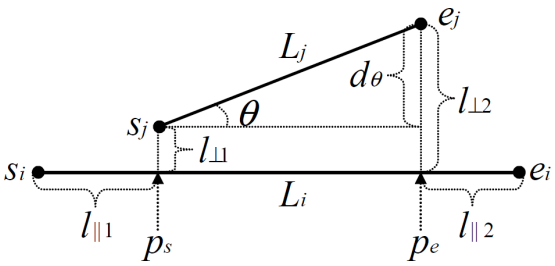


Figure 2.8 – (Figure issue de Lee et al. [2007]) Représentation des deux segments comparés

$$distance_\perp(L_i, L_j) = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}} \quad (2.13)$$

$$distance_\parallel(L_i, L_j) = min(l_{\parallel 1}, l_{\parallel 2}) \quad (2.14)$$

$$distance_\theta(L_i, L_j) = \begin{cases} ||L_j|| * sin(\theta) & \text{si } 0 \leq \theta < 90 \\ ||L_j|| & \text{si } 90 \leq \theta \leq 180 \end{cases} \quad (2.15)$$

(iii) Mesures de similarité basées sur la comparaison des polygones

Pour comparer deux trajectoires, nous pouvons nous intéresser à leurs boites ou polygones englobants. Certaines mesures permettent d'évaluer la similarité (ou la distance) de deux polygones.

RCC-8 (pour ***R**egion **C**onnection **C**alculus* 8) est un système de raisonnement qui étend les relations entre intervalles temporels d'Allen [Allen, 1983] aux régions spatiales [Aiello, 2002] [Sallaberry, 2013] (p. ex. deux polygones sont déconnectés, se superposent, etc.). Il est possible d'utiliser les relations topologiques décrites dans le RCC-8 [Sallaberry, 2013] pour comparer la position d'une région par rapport à une autre. Les relations entre deux régions R et S décrites dans le RCC-8 sont :

1. R et S **sont déconnectées** ;
2. R et S **sont connectées** ;
3. R et S **se superposent** ;
4. R **couvre** S ;
5. R **est couverte par** S ;
6. R **contient** S ;
7. R **est contenue dans** S ;
8. R et S **sont égales**.

Les 9-intersections sont utilisées dans les travaux de Egenhofer [1997] et décrivent les relations topologiques pouvant s'appliquer à des régions, des lignes et des points. Elle sont représentées sous la forme d'une matrice dont les lignes représentent la zone intérieure, la limite et la zone extérieure de R et les colonnes représentent la zone intérieure, la limite et la zone extérieure de S. Une fois remplie, la matrice décrit la relation topologique entre les deux objets spatiaux. Chaque relation est définie comme un noeud dans un graphe de voisinage créé à partir des représentations des relations dans la matrice. [Egenhofer, 1997] évalue la distance entre deux relations topologiques grâce à leur éloignement dans le graphe.

La mesure de similarité appliquée à la recherche d'information spatiale présentée dans Le Parc-Lacayrelle et al. [2007] s'appuie sur l'intersection de deux polygones pour évaluer leur similarité, avec un score nul lorsqu'il n'y a pas d'intersection. Soient R le polygone de requête, S le polygone évalué et I leur intersection. La mesure de similarité est définie par l'équation 2.16.

$$\text{RI\_spatiale}(R,S) = \frac{\text{precision}(R,S) + \text{signifiance}(R,S)}{2 + \text{distance}(R,S)} \tag{2.16}$$

Où $\text{precision}(R,S) = \frac{surface(I)}{surface(S)}$, $\text{signifiance}(R,S) = \frac{surface(I)}{surface(R)}$ et $\text{distance}(R,S) = \frac{d}{D}$ avec d la durée entre les centroïdes de R et I, et D la durée entre le début de R et le centroïde de R.

Synthèse

Nom	**Référence**	**Base**	**Comparaison des points**	**Comparaison des lignes**	**Comparaison des polygones**
Distance euclidienne	Faloutsos et al. [1994]	/	✓	✗	✗
Distance de Haversine	Andrew [1805]	/	✓	✗	✗
SpADe	Chen et al. [2007]	/	✗	✓	✗
Distance de Hausdorff	Alt [2009]	/	✗	✓	✗
Distance de Fréchet	Alt [2009]	/	✗	✓	✗
AMSS	Nakamura et al. [2013a]	/	✗	✓	✗
EDM	Chen et al. [2004]	ED	✗	✓	✗
TRACLUS	Lee et al. [2007]	/	✗	✓	✗
Degré de similarité temporelle appliqué à la RI spatiale	Le Parc-Lacayrelle et al. [2007]	/	✗	✗	✓

Table 2.2 – Résumé des caractéristiques des mesures de similarité spatiale présentées

Le tableau 2.2 fait la synthèse de cette partie 2.3.2 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre mesure de similarité. Chaque colonne correspond à une caractéristique d'intérêt, issue de discussions avec les géographes, et chaque ligne à une mesure présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans une mesure. Les caractéristiques qui nous intéressent sont les suivantes :

— **Comparaison des points** : Les mesures de similarité spatiale basées sur la comparaison des points évaluent la similarité des trajectoires en comparant les points des trajectoires deux à deux. Il s'agit de faire une comparaison fine des trajectoires où le moindre écart d'un déplacement à l'autre est pénalisé.

— **Comparaison des lignes** : Les mesures de similarité spatiale basées sur la comparaison des lignes évaluent la similarité des trajectoires en comparant des séquences de lignes ou de segments extraits des trajectoires. Cette échelle de comparaison intermédiaire convient pour comparer l'allure globale des trajectoires.

— **Comparaison des polygones** : Les mesures de similarité spatiale basées sur la comparaison des polygones évaluent la similarité des trajectoires en comparant les polygones englobants des trajectoires. Lors de leur comparaison manuelle, la première analyse des géographes se base sur la localisation globale des trajectoires. Cette échelle de mesure est parfaite pour imiter cette analyse initiale.

Aucune de ces mesures ne présente toutes les caractéristiques simultanément. Nous souhaitons mettre en place des mesures de similarité combinant des sous-mesures et nous pensons réutiliser certaines mesures présentées dans cette partie afin d'intégrer toutes les échelles de comparaison à notre mesure de similarité spatiale.

Dans la partie suivante, nous présentons des mesures de similarité temporelle.

2.3.3 Mesures de similarité temporelle

Nous souhaitons calculer la similarité temporelle entre deux trajectoires sémantiques. La dimension temporelle d'une telle trajectoire est une suite d'horodatages.

Dans cette partie, nous allons étudier les différentes mesures de similarité pouvant être utilisées ou adaptées pour comparer deux trajectoires sur leur dimension temporelle. Nous classons les mesures de similarité temporelle (cf. figure 2.6, case bleue) en trois catégories. Premièrement, (i) celles qui s'appuient sur une comparaison des horodatages pour calculer le score global. Deuxièmement, (ii) celles qui s'appuient sur une comparaison des intervalles temporels des trajectoires. Troisièmement, (iii) celles qui s'appuient sur une comparaison des contextes temporels. Lors de nos recherches, nous avons seulement trouvé des mesures rentrant dans la seconde catégorie.

(ii) Mesures de similarité basées sur la comparaison des intervalles temporels

Pour calculer la similarité de deux trajectoires, la dimension temporelle est souvent calculée de pair avec la dimension spatiale. Cependant, nous pouvons utiliser les relations d'Allen [Allen, 1983] pour comparer deux trajectoires sur leur dimension temporelle uniquement. Ces travaux définissent 13 relations possibles entre deux intervalles temporels de trajectoires R et S, telles que :

1. R **est égal à** S (p. ex. R = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 20 :00 :00"] et S = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 20 :00 :00"]) et S **est égal à** R ;
2. R **se passe avant** S (p. ex. R = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 12 :00 :00"] et S = ["06/06/2020 ; 15 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
3. et inversement, S **se passe avant** R ;
4. R **rencontre** S (p. ex. R = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 12 :00 :00"] et S = ["06/06/2020 ; 12 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
5. et inversement, S **rencontre** R ;
6. R **se superpose à** S (p. ex. R = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 15 :00 :00"] et S = ["06/06/2020 ; 12 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
7. et inversement, S **se superpose à** R ;
8. R **se passe durant** S (p. ex. R = ["06/06/2020 ; 12 :00 :00", "06/06/2020 ; 15 :00 :00"] et S = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
9. et inversement, S **se passe durant** R ;
10. R **commence** S (p. ex. R = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 12 :00 :00"] et S = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
11. et inversement, S **commence** R ;
12. R **finit** S (p. ex. R = ["06/06/2020 ; 15 :00 :00", "06/06/2020 ; 20 :00 :00"] et S = ["06/06/2020 ; 08 :00 :00", "06/06/2020 ; 20 :00 :00"]) ;
13. et inversement, S **finit** R.

Cependant, ces relations renvoient un résultat booléen et il est nécessaire de les adapter pour obtenir un score tenant compte des écarts entre les bornes de chaque intervalle.

Les travaux de Le Parc-Lacayrelle et al. [2007] présentent une mesure de similarité entre deux intervalles temporels dans un contexte de recherche d'information. Soient R l'intervalle de requête, S l'intervalle évalué et I leur intersection. La mesure de similarité est définie par l'équation 2.17.

$$\text{RI\_temporelle}(R,S) = \frac{\text{precision}(R,S) + \text{signifiance}(R,S)}{2 + \text{distance}(R,S)} \tag{2.17}$$

Où $\text{precision}(R,S) = \frac{duree(I)}{duree(S)}$, $\text{signifiance}(R,S) = \frac{duree(I)}{duree(R)}$ et $\text{distance}(R,S) = \frac{d}{D}$ avec d la durée entre les centroïdes de R et I, et D la durée entre le début de R et le centroïde de R.

Furtado et al. [2018] utilisent, dans la conception d'une mesure de similarité globale entre deux trajectoires sémantiques (cf. partie 2.3.6), une mesure de distance temporelle pour comparer deux intervalles temporels $[t1, t2]$ et $[u1, u2]$. Cette distance correspond au rapport de l'intersection des deux intervalles sur le plus grand intervalle construit à partir des bornes de ces deux intervalles. Elle est définie par l'équation 2.18.

$$\text{temporal\_distance}([t1, t2], [u1, u2]) = 1 - \frac{diameter([t1, t2] \cap [u1, u2])}{diameter([min(t1, u1), max(t2, u2])} \tag{2.18}$$

Avec $diameter([t1, t2]) = |t2 - t1|$ la fonction qui calcule le diamètre d'un intervalle temporel. Ainsi, si les deux intervalles sont égaux, la similarité temporelle est égale à 0 ; plus l'intersection diffère du plus grand intervalle, plus elle se rapproche de 1 ; et si leur intersection est vide, elle est égale à 1.

Synthèse

Nom	**Référence**	**Base**	**Comparaison des horodatages**	**Comparaison des intervalles temporels**	**Comparaison des contextes temporels**
Relations d'Allen	Allen [1983]	/	✗	✓	✗
Degré de similarité temporelle appliqué à la RI temporelle	Le Parc-Lacayrelle et al. [2007]	/	✗	✓	✗

Table 2.3 – Résumé des caractéristiques des mesures de similarité temporelle présentées

Le tableau 2.3 fait la synthèse de cette partie 2.3.3 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre mesure de similarité. Chaque colonne correspond à une caractéristique d'intérêt, issue de discussions avec les géographes, et chaque ligne à une mesure présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans une mesure. Les caractéristiques qui nous intéressent sont les suivantes :

— **Comparaison des horodatages** : Les mesures de similarité temporelle basées sur la comparaison des horodatages évaluent la similarité des trajectoires en comparant leurs horodatages deux à deux.

— **Comparaison des intervalles temporels** : Les mesures de similarité temporelle basées sur la comparaison de intervalles temporels évaluent la similarité des trajectoires en comparant les intervalles temporels délimités par les horodatages de début et de fin des trajectoires. Cette échelle de comparaison permet d'identifier si les deux trajectoires se passent au même moment et ont la même durée.

— **Comparaison des contextes temporels** : Les mesures de similarité temporelle basées sur la comparaison des contextes temporels évaluent la similarité des trajectoires en comparant les éléments de contextes temporels des trajectoires (p. ex. la saison, l'année, le mois, la période de la journée, etc.). De la même manière que pour la dimension spatiale, cette échelle est la plus intuitive pour un géographe qui cherche à comparer deux trajectoires.

Aucune de ces mesures ne présente toutes les caractéristiques simultanément.Comme pour la dimension spatiale, nous souhaitons mettre en place des mesures de similarité combinant des sous-mesures et nous pensons réutiliser certaines mesures présentées dans cette partie afin d'intégrer toutes les échelles de comparaison à notre mesure de similarité temporelle. Le manque de mesures dans deux des trois sous-catégories implique la création de nouvelles mesures ou l'adaptation de mesures à un nouveau contexte.

Dans la partie suivante, nous présentons des mesures de similarité thématique.

2.3.4 Mesures de similarité thématique

Nous souhaitons calculer la similarité thématique entre deux trajectoires sémantiques. La dimension thématique d'une telle trajectoire est un ensemble d'interprétations, c'est-à-dire des séquences d'aspects enrichissant la trajectoire sous différentes perspectives.

Dans cette partie, nous allons étudier les différentes mesures de similarité qui peuvent être utilisées ou adaptées pour comparer deux trajectoires sur leur dimension thématique. Nous classons les mesures de similarité thématique (cf. figure 2.6, case rouge) en trois sous-catégories. Premièrement, (i) celles qui s'appuient sur une comparaison des données d'enrichissement. Deuxièmement, (ii) celles qui s'appuient sur une comparaison des séquences sémantiques. Troisièmement, (iii) celles qui s'appuient sur une comparaison des thématiques globales. Lors de nos recherches, nous avons trouvé des mesures rentrant dans les deux premières catégories.

(i) Mesures de similarité basées sur la comparaison des données d'enrichissement

Les travaux de May Petry et al. [2019] présentent une mesure, appelée **MUITAS** (pour ***MU**lt**I**ple-aspect **TrA**jectory **S**imilarity*), qui permet de comparer des trajectoires multi-aspect. Elle s'intéresse aux attributs de chaque aspect pour une comparaison plus complète des trajectoires. De plus, elle utilise des seuils et des poids personnalisables pour chaque attribut. Les trajectoires sont considérées comme des ensembles de points, plutôt que des séquences. Cependant, la popularité des attributs partagés par plusieurs points ou la fréquence d'occurrence des points eux-mêmes ne sont pas explicitement prises en compte. En outre, le coût de calcul de l'algorithme est très élevé en raison de la mesure de similarité de trajectoire par paire, qui est quadratique par rapport au nombre de points et d'aspects. La mesure de

similarité MUITAS est définie par l'équation 2.19.

$$MUITAS(R,S) = \begin{cases} 0 & \text{si } m = 0 \text{ ou } n = 0 \\ \frac{MUITAS\_parity(R,S)+MUITAS\_parity(S,R}{m+n} & \text{sinon} \end{cases} \tag{2.19}$$

Le score est égal à 0 si au moins l'une des deux trajectoires n'a aucune position. Sinon, il utilise une fonction de parité pour prendre en compte d'éventuelles différences qui peuvent exister entre le score produit lorsque la première trajectoire est comparée avec la seconde et le score produit lorsque la seconde est comparée avec la première. La fonction $MUITAS\_parity$ de calcul du score de similarité d'une trajectoire par rapport à l'autre est définie par l'équation 2.20.

$$MUITAS\_parity(R,S) = \sum_{p1 \in R} max(\{MUITAS\_score(p1,p2) | p2 \in S\}) \tag{2.20}$$

Où $MUITAS\_score$, la fonction de calcul du score de similarité entre deux positions enrichies avec des aspects, est définie par l'équation 2.21.

$$MUITAS\_score(p1,p2) = \sum_{i=1} (MUITAS\_match_{asp\_type_i}(p1,p2) * w_i) \tag{2.21}$$

Où $MUITAS\_match_{asp\_type_i}$ la fonction de calcul du score de similarité entre deux positions selon un type d'aspect donné asp_type_i, est définie par l'équation 2.21.

$$MUITAS\_match_{asp\_type_i}(p1,p2) = \begin{cases} 1 & \text{si } att_j \in asp\_type_i,\ d_j(p1,p2) \leq \delta_j \\ 0 & \text{sinon} \end{cases} \tag{2.22}$$

Ici, pour le type d'aspect asp_type_i, le score de deux points vaut 1 lorsque tous les attributs ont une distance inférieure ou égale à un certain seuil δ_j et 0 sinon.

Récemment, Varlamis et al. [2021b] proposent une mesure de similarité des trajectoires multi-aspect comme composant principal d'un algorithme de clustering hiérarchique, appelé TraFoS. Dans un premier temps, les trajectoires sont transformées en vecteurs de fréquence représentant le nombre d'occurrences de chaque valeur d'un aspect donné. Lorsque les trajectoires comparées diffèrent sensiblement en longueur, il est possible de normaliser la fréquence absolue en la divisant par la longueur de la trajectoire. Pour chaque aspect, un arbre de partition est créé où toutes les trajectoires appartiennent à la racine de l'arbre et, au fur et à mesure que l'on se déplace vers le niveau des feuilles, de moins en moins de trajectoires se trouveront sur chaque nœud. La mesure de Wu et Palmer est ensuite utilisée pour calculer la similarité entre deux trajectoires par rapport à un seul aspect. Il suffit ensuite de pondérer chaque valeur de similarité pour aboutir à une valeur de similarité globale.

(ii) Mesures de similarité basées sur la comparaison des séquences sémantiques

Ying et al. [2010] présentent une mesure de similarité appelée Maximal Semantic Trajectory Pattern Similarity (MSTP-Similarity). Cette mesure permet de comparer deux utilisateurs à travers leur ensemble de trajectoires. MSTP est une valeur représentant un comportement

plus ou moins récurrent (dépendant du support dans la détection avec Prefix-Span) d'un utilisateur particulier. Lors du calcul de similarité entre deux MSTP, la séquence commune la plus longue (LCS) représentant la plus longue partie commune est utilisée. Deux valeurs mesurant les ratios de participation des parties communes par rapport aux deux motifs maximaux des trajectoires sémantiques sont calculées. La similarité MSTP est calculée à l'aide des deux valeurs ratios pondérées ou non.

Zheng et al. [2011] présentent TFIDF pour mesurer la similarité de deux utilisateurs. Certaines séquences communes, appelées séquences similaires, sont découvertes en faisant correspondre les séquences de leurs arrêts à chaque niveau du graphe hiérarchique. Ensuite, pour chaque arrêt d'une séquence similaire, la valeur TFIDF est calculée, où la valeur TF représente la fréquence minimale d'accès des deux utilisateurs à cette région de séjour dans la séquence similaire, tandis que la valeur IDF indique le nombre d'utilisateurs qui ont visité cette région de séjour. Enfin, la similarité entre deux utilisateurs est dérivée de la somme des valeurs TFIDF de toutes les régions de séjour dans les séquences similaires. L'ordre des arrêts dans la séquence n'est pas pris en compte.

Dans un objectif de partitionnement d'un ensemble de trajectoires sémantiques, les travaux de Moreau et al. [2018] présentent une mesure de distance multidimensionnelle. Ils utilisent une agrégation pondérée pour combiner une distance spatiale et une distance thématique mais n'intègrent pas de distance temporelle dans la distance globale. Concernant la distance spatiale, ils réutilisent une distance déjà existante à savoir DTW ou Fréchet. Concernant la distance thématique, l'article présente une nouvelle distance basée sur EDR, appelée distance d'édition enrichie appliquée à des séquences d'épisodes. Elle implémente les opérateurs classiques *Suppression*, *Insertion*, *Modification* de la mesure EDR et ajoute les opérateurs *Permutation*, *Scission* et *Rassemblement*. Toute entité qualitative doit être considérée au sein d'une ontologie ou hiérarchie de concepts afin de pouvoir être soumise à comparaison.

Synthèse

Nom	**Référence**	**Base**	**Comparaison des données d'enrichissement**	**Comparaison des séquences sémantiques**	**Comparaison des thématiques globales**	**Prise en compte des aspects**	**Pondération des axes thématiques**
MUITAS	May Petry et al. [2019]	/	✓	✗	✗	✓	✓
TraFoS	Varlamis et al. [2021b]	/	✓	✗	✗	✓	✓
MSTP	Lu and Tseng [2009]	LCSS	✗	✓	✗	✗	✗
TFIDF	Zheng et al. [2011]	Wu et Palmer	✗	✓	✗	✗	✗
Distance d'édition enrichie	Moreau et al. [2018]	ED	✗	✓	✗	✗	✗

Table 2.4 – Résumé des caractéristiques des mesures de similarité présentées

Le tableau 2.4 fait la synthèse de cette partie 2.3.4 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre mesure de similarité. Chaque colonne correspond à une caractéristique d'intérêt, issue de discussions avec les géographes, et chaque ligne à une mesure présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans une mesure. Les caractéristiques qui nous intéressent sont les suivantes :

— **Comparaison des données d'enrichissement** : Les mesures de similarité thématique basées sur la comparaison des données d'enrichissement évaluent la similarité des trajectoires en comparant les données d'enrichissement des positions deux à deux.

— **Comparaison des séquences sémantiques** : Les mesures de similarité thématique basées sur la comparaison des séquences sémantiques évaluent la similarité des trajectoires en comparant les séquences d'épisodes sémantiques (c.-à-d. les interprétations).

— **Comparaison des thématiques globales** : Les mesures de similarité thématique basées sur la comparaison des thématiques globales évaluent la similarité des trajectoires en comparant les données d'enrichissement qui ressortent le plus souvent dans chaque interprétation. Lorsqu'un géographe compare de trajectoires, il s'intéresse d'abord aux valeur qui ressorte le plus dans les différentes interprétations.

— **Prise en compte des aspects** : Ces mesures comparent des trajectoires multi-aspect. En effet, cette caractéristique est intéressante pour notre mesure car nous enrichissons nos trajectoires avec des aspects.

— **Pondération des axes thématiques** : Dans un contexte de comparaison de trajectoires multi-interprétation, il est intéressant de pouvoir attribuer un poids à chaque sous-score correspondant à chaque interprétation. Par exemple, dans un contexte de comparaison de trajectoires sémantiques touristiques, une interprétation basée sur les points d'intérêt patrimoniaux apporte beaucoup à la comparaison alors qu'une interprétation basée sur les localisations des bouches d'incendie apporte moins à la comparaison. La pondération permet de donner plus d'importance à chaque interprétation selon le contexte.

Aucune de ces mesures ne présente toutes les caractéristiques simultanément. Comme pour la dimension spatiale et temporelle, nous souhaitons mettre en place des mesures de similarité combinant des sous-mesures et nous pensons réutiliser certaines mesures présentées dans cette partie afin d'intégrer toutes les échelles de comparaison à notre mesure de similarité temporelle. Le manque de mesures dans une des trois sous-catégories implique la création de nouvelles mesures ou l'adaptation de mesures à un nouveau contexte.

Dans la partie suivante, nous présentons des mesures de similarité appliquées aux séries temporelles.

2.3.5 Mesures de similarité entre séries temporelles

Afin de comparer deux trajectoires sémantiques, nous pouvons utiliser les mesures de similarité s'appliquant aux séries temporelles (cf. figure 2.6, case grise). Une série temporelle est une suite de valeurs situées dans le temps que nous pouvons rapprocher de nos valeurs de longitude et de latitude horodatées. Une série temporelle classique n'a donc qu'une dimension alors qu'une trajectoire en a deux ou plus Su et al. [2020]. Pour les mesures présentées dans cette partie, nous admettons que le pas de temps entre chaque élément d'une série temporelle est régulier.

Vlachos et al. [2002] présentent une mesure de similarité entre deux trajectoires non-métrique basée sur LCSS (Longest Common Subsequence). Elle permet d'attribuer un score de similarité à une paire de trajectoires. Plus ce score est haut, plus les deux trajectoires se ressemblent. Elle utilise un seuil ϵ au delà duquel la distance entre deux éléments est trop différente pour qu'ils soient considérés comme similaires. La mesure de similarité LCSS est définie par l'équation 2.23.

$$LCSS(R,S) = \begin{cases} 0 & \text{si } \begin{array}{l} m = 0 \text{ ou} \\ n = 0 \end{array} \\ 1 + LCSS(Rest(R), Rest(S)) & \text{si } |r_1 - s_1| < \epsilon \\ max \left\{ \begin{array}{l} LCSS(Rest(R), S) \\ LCSS(R, Rest(S)) \end{array} \right\} & \text{sinon} \end{cases} \tag{2.23}$$

Vintsyuk [1968] présentent une mesure de distance, appelée DTW (*Dynamic Time Warping*). Cette mesure permet d'attribuer un score de distance à une paire de séries temporelles. Plus le score est élevé, moins les séries se ressemblent et inversement. Dans un premier temps, le but est d'aligner les deux séries de telle sorte que les éléments comparés soient les plus proches. Pour ce faire, une matrice de déformation (*warping matrix*) est construite. Prenons les deux séries temporelles R et S décrites plus haut. Un élément (i, j) de la matrice de déformation contiendra la distance $d(r_i, s_j)$ entre deux éléments r_i et s_j de R et S. Pour calculer la distance entre deux éléments, il est possible d'utiliser la distance euclidienne (c.-à-d. la norme L2) ou la distance de Manhattan (c.-à-d. la norme L1). Dans les travaux de Keogh and Ratanamahatana [2005], la distance entre deux éléments r_i et s_j est $d(r_i, s_j) = (r_i, s_j)^2$. La mesure DTW est la somme des éléments du chemin continu le moins coûteux pour aller de l'élément $(1, 1)$ à l'élément (m, n) de la matrice. Ainsi, la distance DTW est définie par l'équation 2.24.

$$DTW(R,S) = \begin{cases} 0 & \text{si } m = n = 0 \\ \infty & \text{si } \begin{array}{l} m = 0 \text{ ou} \\ n = 0 \end{array} \\ d(r_1, s_1) + min \left\{ \begin{array}{l} DTW(Rest(R), Rest(S)) \\ DTW(Rest(R), S) \\ DTW(R, Rest(S)) \end{array} \right\} & \text{sinon} \end{cases} \tag{2.24}$$

Ici, les éléments ne sont pas comparés deux à deux mais le même élément d'une série peut être comparé avec différents éléments de l'autre série ce qui permet la déformation de l'axe du temps. Ainsi, deux séries peuvent avoir un score élevé car leurs enchaînements d'éléments sont similaires mais ne se situent pas dans la même temporalité. Cette mesure est non-métrique, elle ne satisfait pas toujours la condition d'inégalité triangulaire Chen and Ng [2004]. La complexité de calcul de DTW est de $O(nm)$ Keogh and Ratanamahatana [2005].

Chen and Ng [2004] proposent une autre mesure de distance entre deux séries temporelles appelée ERP (*Edit distance with Real Penalty*). Cette distance est une combinaison de la distance EDR et de la distance euclidienne. Elle introduit la notion d'écart (*gap*) qui fait référence à une sous-partie de la série temporelle ou de la trajectoire située entre deux composantes similaires identifiées. La distance ERP est définie par l'équation 2.25.

$$ERP(R,S) = \begin{cases} \sum_{i=1}^{n} d_{ERP}(s_i, g) & \text{si } m = 0 \\ \sum_{j=1}^{m} d_{ERP}(r_j, g) & \text{si } n = 0 \\ min \left\{ \begin{array}{l} d_{ERP}(r_1, s_1) + ERP(Rest(R), Rest(S)) \\ d_{ERP}(r_1, g) + ERP(Rest(R), S) \\ d_{ERP}(g, s_1) + ERP(R, Rest(S)) \end{array} \right\} & \text{sinon} \end{cases} \tag{2.25}$$

Où $d_{ERP}(r_i, s_i) = |r_i - s_i|$, $d_{ERP}(r_i, g) = |r_i - g|$ et $d_{ERP}(s_i, g) = |s_i - g|$ et g une valeur constante.

Chen et al. [2005] proposent une mesure de similarité appelée EDR (*Edit Distance on Real sequences*). Cette mesure attribue un score de distance à un couple de trajectoires. La distance EDR est définie par l'équation 2.26.

$$EDR(R,S) = \begin{cases} n & \text{si } m = 0 \\ m & \text{si } n = 0 \\ min \left\{ \begin{array}{l} Subcost(r_1, s_1) + EDR(Rest(R), Rest(S)) \\ 1 + EDR(Rest(R), S) \\ 1 + EDR(R, Rest(S)) \end{array} \right\} & \text{sinon} \end{cases} \tag{2.26}$$

Les éléments r_i et s_j sont dits correspondants si $d(r_i, s_j) \leq \epsilon$ avec ϵ le seuil de correspondance. La fonction $Subcost(r_i, s_j) = 0$ si les éléments r_i et s_j correspondent et $Subcost(r_i, s_j) = 1$ sinon.

Marteau [2009] présentent la mesure de distance TWED (*Time Warp Edit Distance*). De la même manière que DTW, TWED permet de comparer deux séries temporelles en permettant la distorsion du temps. Cette mesure prend en compte les trois opérations d'édition, à savoir les opérations de suppression d'un élément de T1, de suppression d'un élément de T2 et de

modification (*match operation*).

$$TWED(R,S) = \begin{cases} 0 & \text{si } m = n = 0 \\ \sum_{i=1}^{n} d(s_i, s_{i+1}) & \text{si } m = 0 \\ \sum_{i=1}^{m} d(r_i, r_{i+1}) & \text{si } n = 0 \\ min \begin{cases} d(r_1, r_2) + \delta + TWED(Rest(R), S) \\ d(r_1, s_1) + TWED(Rest(R), Rest(S)) \\ d(s_1, s_2) + \delta + TWED(R, Rest(S)) \end{cases} & \text{sinon} \end{cases} \tag{2.27}$$

Nom	**Référence**	**Base**	**Possibilité de déformation locale du temps**	**Résistante au bruit**
LCSS	Vlachos et al. [2002]	/	✓	✓
DTW	Vintsyuk [1968]	/	✓	✗
EDR	Chen et al. [2005]	ED	✓	✓
ERP	Chen and Ng [2004]	Distance euclidienne et EDR	✓	✗
TWED	Marteau [2009]	/		✗

Table 2.5 – Résumé des caractéristiques des mesures de similarité entre séries temporelles présentées

Le tableau 2.5 fait la synthèse de cette partie 2.3.5 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre mesure de similarité. Chaque colonne correspond à une caractéristique d'intérêt, issue de discussions avec les géographes, et chaque ligne à une mesure présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans une mesure. Les caractéristiques qui nous intéressent sont les suivantes :

— **Possibilité de déformation locale du temps** : Les mesures disposant de cette caractéristique peuvent déformer le temps afin d'identifier les ressemblances entre deux séries temporelles (ou deux trajectoires). Seul l'ordre des éléments est pris en compte mais pas le temps entre deux éléments.

— **Résistance au bruit** : Parfois les trajectoires capturées ont des valeurs aberrantes qui peuvent empêcher une mesure de similarité de fonctionner correctement car elle s'éloigne trop de toutes les autres valeurs. Les mesures résistantes au bruit ne prennent pas en compte ces valeurs aberrantes dans le calcul du score final.

Les mesures de similarité entre séries temporelles peuvent être utilisées pour comparer

les trajectoires sémantiques sur différentes dimensions. En effet, la dimension spatiale peut être vue comme une série temporelle de positions spatiales, la dimension temporelle peut être vue comme une série temporelle d'horodatages et la dimension thématique peut être vue comme un ensemble de séries temporelles de données d'enrichissement. Nous envisageons d'adapter certaines de ces mesures pour les intégrer à notre mesures.

Dans la partie suivante, nous présentons des mesures de similarité multidimensionnelle.

2.3.6 Mesures de similarité multidimensionnelle

Certaines mesures de similarité permettent d'évaluer deux trajectoires selon plusieurs dimensions (cf. figure 2.6, case violette).

Les travaux de Little and Gu [2001] s'intéressent à l'observation d'objets mobiles sur une vidéo. L'approche présentée utilise la nature temporelle de la vidéo pour décrire le mouvement de l'objet mobile observé. Ainsi, la trajectoire d'un objet mobile est une séquence de localisations (x_i, yi) avec $i = 1..n$ où n est le nombre de frames dans la séquence. Ce travail propose de séparer l'information spatiale de l'information temporelle dans la trajectoire. Un déplacement est représenté par la courbe du chemin (en anglais, *path curve*) et par celle de la vitesse (en anglais, *speed curve*). Le chemin est construit selon une séquence d'images donnée et la vitesse peut être calculée si on connaît l'intervalle de temps entre deux images. La courbe de la vitesse est une courbe 2D avec x correspondant au temps et y à la vitesse. Ces travaux proposent une méthode de recherche spatio-temporelle qui, selon une trajectoire de requête, retourne les trajectoires similaires issues d'une base de données. Pour cela, la méthode s'intéresse successivement à la courbe du chemin et la courbe de la vitesse. Pour chacune de ces courbes, la sélection des trajectoires similaires à la trajectoire requête est faites grâce à un processus en trois étapes : (1) sélection des trajectoires dont les courbes ont des boîtes englobantes similaires à celle de la trajectoire requête, (2) sélection des trajectoires dont les angles des courbes sont similaires à ceux de la trajectoire requête et (3) sélection des trajectoires sont les tailles relatives des segments des courbes sont similaires à ceux de la trajectoire requête. Finalement, pour récupérer les trajectoires similaires à la trajectoire requête, il faut ensuite faire l'intersection des trajectoires issues des processus de sélection selon les deux courbes décrivant les trajectoires (c.-à-d. la courbe du chemin et la courbe de la vitesse).

Les travaux de Chen et al. [2020] introduisent un algorithme de jointure de trajectoires similaires appelé STS-Join (*Semantic Trajectory Similarity Join*) qui utilise une nouvelle mesure de similarité spatio-thématique pour grouper des paires de trajectoires sémantiques similaires. Ils définissent une trajectoire sémantique comme étant (1) une suite de points d'intérêt ayant une géolocalisation ρ et une description textuelle ψ telle que $R = \langle poi_1, poi_2, ..., poi_n \rangle$ où $poi_i = \{\rho, \psi\}$ (2) une suite de géolocalisations ρ accompagnée d'un unique document textuel ψ telle que $R = \{\rho, \psi\}$ où $\rho = \langle \rho_1, \rho_2, ..., \rho_n \rangle$. La mesure de similarité utilisée est appelée ST (*spatio-textual similarity*) et il s'agit d'une combinaison linéaire de la similarité spatiale des deux trajectoires et de leur similarité thématique (appelée similarité textuelle car n'utilisant

que des textes descriptifs). La distance ST est définie par l'équation 2.28.

$$ST(R,S) = \frac{\sum_{poi_i \in R} Rel(poi_i, S)}{|R|} + \frac{\sum_{poi_j \in S} Rel(poi_j, R)}{|S|} \tag{2.28}$$

Où Rel la fonction de pertinence d'un point d'intérêt par rapport à une trajectoire est définie par l'équation 2.29.

$$Rel(poi, R) = max_{poi_i \in R}\{\alpha * S(poi.\rho, poi_i.\rho) + (1-\alpha) * T(poi.\psi, poi_i.\psi)\} \tag{2.29}$$

Comme nous avons pu le constater, les trajectoires sémantiques comportent plusieurs dimensions. Les travaux de Lu and Tseng [2009] présentent la méthode **LBS-Alignment** (***L**ocation-**B**ased **S**ervice **Alignment***) pour mesurer la similarité spatio-tempo-thématique de deux séquences de transactions mobiles appartenant à deux utilisateurs de *smartphones*. Les transactions mobiles correspondent à des requêtes faites par un utilisateur à une station de base du réseau mobile terrestre. Les séquences de transactions mobiles s'apparentent à des trajectoires sémantiques car chaque transaction est composée du temps de requêtage, de l'identifiant de la station de base géolocalisée et de l'identifiant du ou des services demandés par l'utilisateur. La mesure de similarité LBS-Alignment de Lu and Tseng [2009] s'appuie sur ces trois dimensions (c.-à-d. spatiale, temporelle et thématique) pour calculer le score de similarité. Une matrice de similarité est construite pour mesurer la similarité de chaque paire d'éléments des séquences (c.-à-d. chaque paire de transactions) selon chaque dimension. Pour chaque paire de transactions, la dimension spatiale est évaluée. Des pénalités sont ajoutées au score lorsque les dimensions présentent des différences.

Les travaux de Furtado et al. [2018] proposent une mesure de similarité spatio-tempo-thématique, appelé **MSM** (pour ***M**ultidimensional **S**imilarity **M**easure*), qui permet de comparer deux trajectoires sémantiques sur leurs trois dimensions. Ils définissent une trajectoire sémantique S comme étant une séquence d'arrêts à des points d'intérêt particuliers (p. ex. hôtel, cinéma, musée, etc.) et la formalise de telle sorte que $S = \langle a_1, ..., a_n \rangle$ avec $a_i = ((x,y), [t1, t2], type)$ où (x,y) correspond au centroïde de l'arrêt, $[t1, t2]$ à l'intervalle de temps durant lequel l'arrêt est effectif et $type$ au type de l'arrêt. D'autres données d'enrichissement des arrêts peuvent être ajoutées à cette description à la suite du type. La distance MSM est calculée en deux temps : d'abord, c'est la première trajectoire qui est comparée à la seconde et ensuite, c'est la seconde qui est comparée à la première. MSM est définie par l'équation 2.30.

$$MSM(R,S) = \begin{cases} 0 & \text{si } m = 0 \text{ ou } n = 0 \\ \frac{MSM\_parity(R,S) + MSM\_parity(S,R)}{m+n} & \text{sinon} \end{cases} \tag{2.30}$$

Où MSM_parity, la fonction de calcul du score de similarité d'une trajectoire par rapport à l'autre, est définie par l'équation 2.31.

$$MSM\_parity(R,S) = \sum_{a \in R} max_{b \in S}\{MSM\_score(a,b)\} \tag{2.31}$$

Où MSM_score, la fonction de calcul du score de similarité entre deux éléments selon un ensemble de dimensions D (p. ex. spatiale, temporelle ou thématique), est définie par

l'équation 2.32.

$$MSM\_score(a,b) = \sum_{k=1}^{|D|}(MSM\_match_k(a,b) * \gamma_k) \tag{2.32}$$

Où MSM_match_k, la fonction de calcul de la correspondance entre deux éléments selon la dimension k, vaut 1 lorsque les deux arrêts a et b ont une distance inférieure ou égale à un certain seuil sur la dimension k et 0 sinon. Selon la dimension traitée, la fonction de distance utilisée peut être différente. Dans un exemple, Furtado et al. [2018] utilisent la distance euclidienne entre les localisations des arrêts pour la dimension spatiale, leur distance de similarité temporelle présentée précédemment (cf. partie 2.3.3) pour la dimension temporelle et une simple mesure égale à 0 si les deux types sont égaux et à 1 sinon pour la dimension thématique.

Nom	**Référence**	**Base**	**Prise en compte des trois dimensions**	**Pondération des dimensions**	**Comparaison de plusieurs axes thématiques**
	Little and Gu [2001]	/	✗	✗	✗
STS-Join	Chen et al. [2020]	/	✗	✓	✗
LBS-Alignment	Lu and Tseng [2009]	/	✓	✗	✗
MSM	Furtado et al. [2018]	/	✓	✓	✗

Table 2.6 – Résumé des caractéristiques des mesures de similarité multidimensionnelle présentées

Le tableau 2.6 fait la synthèse de cette partie 2.3.6 en rappelant la présence ou l'absence des caractéristiques que nous cherchons à intégrer à notre mesure de similarité. Chaque colonne correspond à une caractéristique d'intérêt, issue de discussions avec les géographes, et chaque ligne à une mesure présenté précédemment. Les intersections montrent la présence (avec ✓) ou l'absence (avec ✗) d'une caractéristique dans une mesure. Les caractéristiques qui nous intéressent sont les suivantes :

— **Prise en compte des trois dimensions** : Les mesures prennent les dimensions spatiale, temporelle et thématique dans leur comparaison des trajectoires.

— **Pondération des dimensions** : Dans un contexte de comparaison de trajectoires sur plusieurs dimensions, il est intéressant de pouvoir attribuer un poids à chaque sous-score correspondant à chaque dimension.

— **Comparaison de plusieurs axes thématiques** : Les trajectoires sémantiques sont souvent enrichies par plusieurs types de données d'enrichissement et il est donc important de pouvoir prendre en compte chacun de ces axes grâce à une mesure globale.

Aucune de ces mesures ne présente toutes les caractéristiques simultanément. Ainsi, dans la suite de ces travaux, nous allons mettre en place une mesure de similarité globale permettant de comparer deux trajectoires sémantiques selon les dimensions spatiale, temporelle et thématique.

2.4 Synthèse générale

Dans ce chapitre, nous avons, dans un premier temps, donné des définitions formelles sur les données de mobilité. La trace de mobilité, la trajectoire brute, la trajectoire sémantique, etc. sont définis pour la suite du mémoire.

Ensuite, nous avons fait l'état de l'art des modèles de trajectoires sémantiques. Nous avons classé les modèles dans plusieurs catégories : les modèles basés sur les arrêts et les déplacements, les modèles basés sur les épisodes, les modèles multi-interprétation, les modèles multi-interprétation et multi-niveau et les modèles multi-aspect. Nous souhaitons concevoir un modèle qui intègre des caractéristiques issues des modèles de l'état de l'art, telles que le multi-aspect, le multi-niveau, le multi-interprétation, etc.

Enfin, nous avons fait l'état de l'art des modèles de trajectoires et des mesures de similarité pouvant être utilisées pour comparer deux trajectoires sémantiques sur leurs dimensions spatiale, temporelle et thématique. Nous les avons classifié dans différentes catégories : les mesures de similarité spatiale, les mesures de similarité temporelle, les mesures de similarité thématique, les mesures de similarité entre séries temporelles et les mesures de similarité multidimensionnelle. Nous souhaitons concevoir des mesures de similarité qui combinent des sous-mesures s'intéressant à des dimensions différentes.

Le prochain chapitre présente nos deux contributions, à savoir, le modèle de trajectoire sémantique et les deux mesures de similarité entre trajectoires sémantiques.

Chapitre 3

Contributions à la modélisation et au traitement des traces de mobilité

Dans l'introduction de ce mémoire, nous avons établi notre objectif : celui de concevoir et développer une plateforme modulaire de type ETL (c.-à-d. ***E**xtract, **T**ransform, **L**oad*) permettant de construire et d'exécuter des chaînes de traitement dédiées aux traces de mobilité. Notre jeu de données principal est un ensemble de traces de mobilité touristiques et nous souhaitons utiliser cette plateforme pour faciliter l'analyse de ces données dans le but de mieux comprendre les déplacements touristiques. La motivation première des aménageurs du territoire est d'utiliser les analyses réalisées pour aménager et valoriser au mieux le territoire touristique.

La plateforme repose sur un modèle de représentation de trajectoires sémantiques qui sert de modèle de transition entre les modules d'une chaîne de traitement. Ainsi, les entrées et les sorties de chaque module sont des instanciations de ce modèle. Il possède certaines caractéristiques des modèles existants et d'autres qui lui sont propres. Chacune de ces particularités permet de répondre aux besoins des géographes. Ce modèle est la première contribution de cette thèse et est abordé plus en détail dans la suite de ce chapitre (cf. partie 3.1).

La plateforme intègre des modules de traitement de bas niveau, qui, enchaînés les uns à la suite des autres, permettent de répondre à des questionnements de plus haut niveau. Certains modules existent déjà dans la littérature, d'autres doivent être adaptés à notre plateforme et notre contexte et, enfin, les derniers sont conçus et développés dans le cadre de cette thèse. La seconde contribution de cette thèse regroupe deux mesures de calcul de similarité permettant d'évaluer la ressemblance de deux trajectoires sémantiques en utilisant les dimensions spatiale, temporelle et thématique. Nous détaillons ces mesures dans la suite de ce chapitre (cf. partie 3.2).

3.1 Modèle de trajectoire sémantique DA3T

Notre première contribution est un modèle de représentation des trajectoires sémantiques qui est utilisé comme modèle de transition dans la plateforme. Nous rappelons, d'abord, les besoins de géographes et aménageurs locaux, les hypothèses et les verrous de recherche relatifs à notre représentation des trajectoires sémantiques. Puis, nous présentons notre modèle de trajectoire sémantique, d'abord de manière générale, puis à travers des caractéristiques spécifiques. Enfin, nous présentons la structuration JSON que nous utilisons lorsque nous instancions le modèle.

3.1.1 Rappel des besoins, hypothèses et verrous de recherche

Notre objectif final étant de construire une plateforme modulaire, nous avons besoin d'un modèle de transition permettant de formaliser les données entrantes et sortantes de chaque module. Nous avons choisi de proposer un nouveau modèle car tous nos besoins n'étaient pas satisfaits en même temps par les modèles existants. Cependant, certaines caractéristiques de ces modèles semblent être intéressantes à intégrer au nôtre afin de combler certains besoins.

Comme énoncé dans la partie 1.2.1, notre modèle doit représenter d'une part les traces de mobilité brutes, d'autre part les données d'enrichissement et doit proposer une manière de lier ces deux types de données ensemble.

L'objet central de nos travaux est la trace de mobilité. Nous souhaitons pouvoir représenter tout type de traces de mobilité, qu'elles appartiennent à n'importe quel objet mobile (p. ex. traces d'animaux, d'humains, de véhicules, etc.) et qu'elles soient dans n'importe quel contexte (p. ex. *indoor* et *outdoor*). Cela implique de pouvoir décrire les spécificités des objets mobiles observés (p. ex. âge et sexe du touriste dans notre contexte) et des positions capturées (p. ex. positions GPS, positions *indoor*, etc.).

Pour permettre une meilleure compréhension du déplacement, la trace de mobilité est enrichie avec des données d'enrichissement pouvant représenter n'importe quel phénomène du monde réel. Il s'agit de données complexes, chacune caractérisée par un ensemble d'attributs évoluant au cours du temps. Prenons l'exemple d'un point d'intérêt tel que le musée maritime de La Rochelle : nous pouvons le décrire, à un temps donné, grâce à son nom "Musée Maritime de La Rochelle", grâce à sa localisation précise sous la forme d'un polygone spatial, grâce à ses horaires d'ouverture, etc. Chacune de ces données est un attribut du musée qui peut changer au cours du temps ; en 2015, le musée s'est agrandi, ceci a impliqué un changement du polygone spatial représentant sa localisation. Une donnée d'enrichissement est aussi caractérisée par un type pouvant appartenir à une hiérarchie de type. Reprenons l'exemple du musée maritime de La Rochelle : cette donnée est de type "musée" qui est un sous-type du type "point d'intérêt".

L'enrichissement d'une trajectoire est réalisée grâce à (1) l'annotation des positions de la trajectoire grâce aux données d'enrichissement et (2) la ou les segmentations de la trajectoire en épisodes selon un ou plusieurs critères donnés. L'annotation des positions avec des

données d'enrichissement peut se faire automatiquement en s'appuyant sur les attributs des données :

— Si la donnée d'enrichissement a un attribut spatial et un attribut temporel identifiés comme étant les critères d'annotation, alors il y a enrichissement de la position avec cette donnée si sa localisation et son marqueur temporel correspondent partiellement aux attributs spatial et temporel de la donnée d'enrichissement. Par exemple, nous voulons enrichir une trajectoire Géoluciole avec la donnée d'enrichissement décrivant le musée maritime de La Rochelle, si une ou plusieurs positions de la trajectoire sont localisées a l'intérieur du polygone spatial décrivant le musée et si leurs horodatages sont compris dans les horaires d'ouverture du musée, alors ces positions sont enrichies avec cette donnée et nous pouvons conclure que le touriste a visité ou est passé par le musée pendant ses horaires d'ouverture.

— Si la donnée d'enrichissement a un attribut spatial identifié comme étant le critère d'annotation, alors il y a enrichissement de la position avec cette donnée si sa localisation correspond partiellement à l'attribut spatial de la donnée d'enrichissement. Par exemple, nous voulons enrichir une trajectoire Géoluciole avec la donnée d'enrichissement décrivant la plage des Minimes de La Rochelle, si une ou plusieurs positions de la trajectoire sont localisées a l'intérieur du polygone spatial décrivant la plage, alors ces positions sont enrichies avec cette donnée et nous pouvons conclure que le touriste est allé à la plage.

— Si la donnée d'enrichissement a un attribut temporel identifié comme étant le critère d'annotation, alors il y a enrichissement de la position avec cette donnée si son marqueur temporel correspond partiellement à l'attribut temporel de la donnée d'enrichissement. Par exemple, nous voulons enrichir une trajectoire Géoluciole avec la donnée d'enrichissement décrivant la fête nationale, si une ou plusieurs positions de la trajectoire ont leurs horodatages qui correspondent à la date de l'événement, alors ces positions sont enrichies avec cette donnée et nous pouvons conclure que le touriste était à La Rochelle le jour de la fête nationale.

— Si la donnée d'enrichissement n'a aucun attribut identifié comme critère d'annotation, alors l'enrichissement peut être fait manuellement. Par exemple, nous voulons enrichir une trajectoire Géoluciole avec un extrait d'entretien, ce dernier peut être manuellement attaché à des positions de la trajectoire.

— Certains enrichissements peuvent être faits par calcul. Par exemple, nous voulons enrichir une trajectoire Géoluciole avec les phases d'arrêts et de déplacements, ces phases peuvent être calculées et attachées aux positions de la trajectoire.

La segmentation est faite selon un ou plusieurs critères de segmentation. Dans notre modèle, les critères sont des types d'aspect. Par exemple, prenons une trajectoire enrichie avec la météo et des points d'intérêt, il est possible de segmenter la trajectoire en s'appuyant sur

l'un des deux ou les deux types d'enrichissement simultanément. Notre modèle permet d'effectuer plusieurs segmentations sur une même trajectoire, ce qui donne plusieurs interprétations (c.-à-d. séquence d'épisodes) de la trajectoire. Chaque interprétation peut être détaillée sur plusieurs niveaux. Par exemple, une interprétation de la trajectoire construite à partir des activités touristiques peut avoir un premier niveau avec les activités générales comme "activité culturelle" et un second niveau avec des activités plus précises comme "visite d'un monument".

Le verrou **(V1)** décrit le besoin d'un tel modèle de représentation de traces de mobilité *indoor* et *outdoor* prenant en compte plusieurs niveaux d'enrichissement avec des données complexes.

Pour résoudre le verrou **(V1)**, nous émettons l'hypothèse **(H1)** de combiner certaines caractéristiques de modèles existants, dans un même modèle de trajectoire sémantique dans le but de représenter plus fidèlement des traces de mobilité brutes, de sources différentes, et leurs enrichissements complexes. D'après notre état de l'art, les caractéristiques multi-interprétation, multi-niveau et multi-aspect permettent de répondre à plusieurs de nos besoins. Ainsi, nous allons adapter et intégrer ces caractéristiques dans un modèle de représentation des trajectoires sémantiques.

3.1.2 Présentation du modèle DA3T

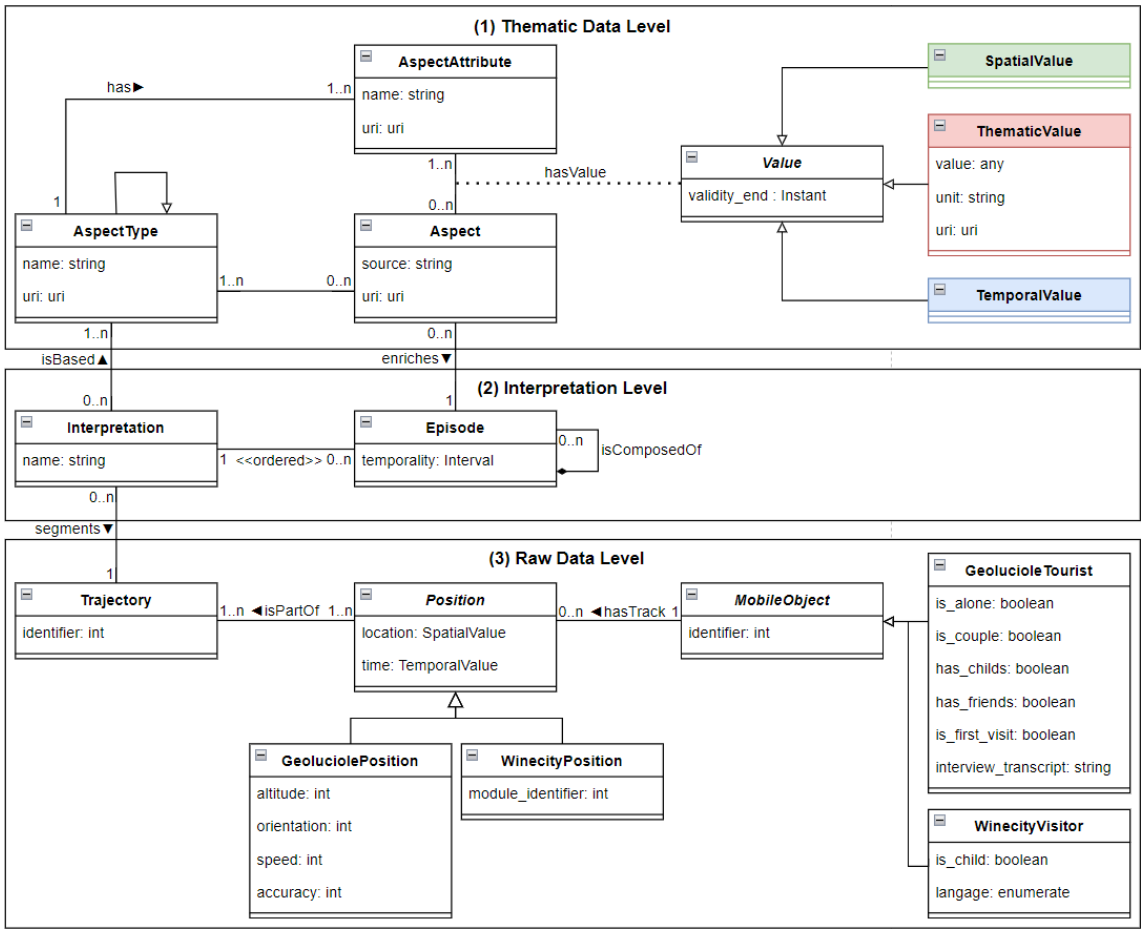


Figure 3.1 – Modèle de trajectoire sémantique

La figure 3.1 décrit notre modèle de trajectoire sémantique sous la forme d'un diagramme de classes. Il est décomposé en trois parties distinctes :

La partie *Raw data level* (cf. figure 3.1, bloc 3) rassemble les classes représentant les données brutes collectées. Les données générales relatives aux objets mobiles sont dans la classe *MobileObject* et les données plus spécifiques relatives à une catégorie d'objets mobiles en particulier (p. ex. les touristes volontaires de notre projet) sont dans les classes correspondantes qui en héritent (p. ex. la classe *GeolucioleVisitor*). La trace de mobilité d'un objet mobile est représentée par les instanciations de la classe *Position*. Il peut y avoir plusieurs types de positions (p. ex. dans notre projet, *IndoorPosition* pour les positions de visiteurs dans les musées et *OutdoorPosition* pour les positions collectées avec Géoluciole) qui ont différents attributs, mais qui héritent toutes de la classe générique *Position*. Un objet mobile possède une suite de positions qui décrit le déplacement capturé au complet, c.-à-d. sa trace de mobilité. Les trajectoires sont des sous-parties de cette trace de mobilité qui présentent un intérêt pour une application donnée et sont représentées grâce à la classe *Trajectory*. Ce modèle est **générique** et **extensible** en fonction du contexte applicatif. Les classes de base du modèle (p. ex. *MobileObject* et *Position* sont assez génériques pour accepter n'importe quel type de traces

et de données d'enrichissement. Les classes qui peuvent être étendues sont *MobileObject*, à laquelle il est possible d'ajouter des classes enfants représentant de nouveaux types d'objets mobiles (p. ex. une classe d'objets mobiles *Vehicle* héritant de *MobileObject*) et *Position* à laquelle il est possible d'ajouter des classes enfants représentant de nouveaux types de positions (p. ex. une classe de positions *GPSPosition* héritant de *Position*).

La partie *Semantic data level* (cf. figure 3.1, bloc 1) regroupe les données sémantiques. Comme dans le modèle MASTER Mello et al. [2019], nous souhaitons représenter les données sémantiques sous la forme d'aspects sémantiques. Notre modèle est donc qualifié de modèle **multi-aspect**. Quatre classes principales représentent ces aspects (c.-à-d. *Aspect*, *AspectType*, *Attribute* et *Value*). L'aspect représente un phénomène du monde réel identifié comme ayant de l'intérêt pour une application en question. Le type d'aspect représente la catégorie de ce phénomène (p. ex. *activité touristique*, *point d'intérêt*, *moyen de déplacement*, *météo*, etc.) et possède des attributs spécifiques (p. ex. les *points d'intérêt* sont chacun caractérisés par un *nom*, une *localisation*, un *type*, etc.). Un type d'aspect peut avoir des sous-types (p. ex. un type *mode de transport* peut avoir comme sous-type *voiture*, *vélo*, *bus*, etc.). Lors de la création d'un aspect, au minimum un type d'aspects lui est associé et chaque attribut associé à ce type est instancié grâce à la classe d'association *Value* (p. ex. l'aspect *tour de la lanterne* est un *point d'intérêt* qui a pour *nom* : *Tour de la Lanterne*, pour *localisation* : *[46.1558333,-1.1569444]*, pour *type* : *tour*, etc.). Dans le modèle original, les attributs sont uniquement instanciés sous la forme de chaînes de caractères. Dans notre modèle, nous avons choisi d'ajouter des classes pour distinguer les attributs temporels, spatiaux et thématiques afin, entre autres, d'optimiser l'automatisation de la phase d'enrichissement (p. ex. en faisant correspondre les attributs spatiaux d'un aspect avec les positions spatiales de la trajectoire). Les classes représentant les attributs spatiaux et temporels sont développées dans la figure 3.2

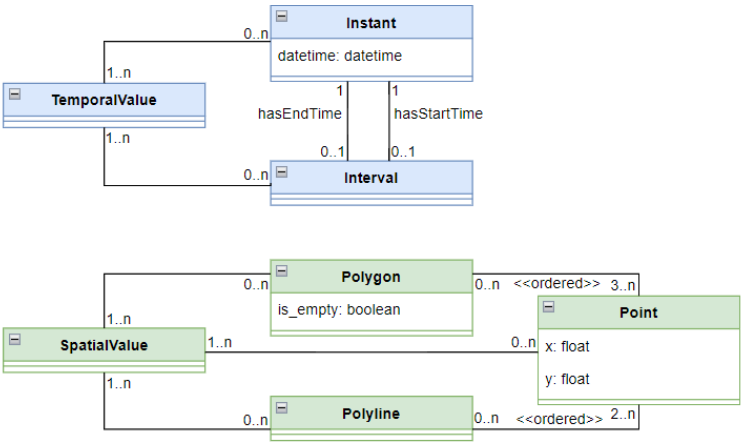


Figure 3.2 – Approfondissement des classes *SpatialValue* et *TemporalValue*

Chaque valeur d'attribut peut avoir une validité temporelle (p. ex. les horaires d'ouverture d'un point d'intérêt peuvent varier en fonction du jour de la semaine, la date ou le lieu d'un festival peuvent être différents d'une année à l'autre, etc.). Certaines classes du modèle peuvent être reliées à des concepts provenant d'ontologies grâce à un attribut *uri* (p. ex. pour décrire des aspects de type *point d'intérêt*, on peut s'appuyer sur des ontologies externes

comme celle de DATAtourisme [DATAtourisme, b]. Une URI peut soit décrire un type d'aspect, soit un attribut d'aspect, soit une valeur d'un aspect particulier.

La partie *Interpretation level* (cf. figure 3.1, bloc 2) sert de lien entre les données brutes et les données sémantiques. Un épisode de la classe *Episode* est un intervalle temporel auquel des données d'enrichissement sont liées. Une trajectoire spécifique peut être liée à une ou plusieurs interprétations de la classe *Interpretation*. Une interprétation est une séquence d'épisodes particulière (p. ex. la trajectoire d'un touriste peut avoir une interprétation pour décrire la *météo* au cours du déplacement, une autre interprétation pour décrire les *pratiques touristiques*, etc.) [Yan et al., 2011]. Il est possible de détailler un épisode grâce au lien de composition récursif signifiant qu'un épisode peut être précisé par d'autres épisodes (p. ex. une interprétation pour décrire les *pratiques touristiques* peut être décrite sur plusieurs niveaux allant des activités génériques comme "activité nautique" aux activités très précises "baignade à la plage"). Ainsi, notre modèle est qualifié de modèle **multi-niveau** [Fileto et al., 2015].

Nous avons présenté un modèle multi-aspect et multi-niveau qui est générique et extensible.

3.1.3 Caractéristiques du modèle

Nous souhaitons présenter les caractéristiques importantes du modèle, à savoir, la représentation des données d'enrichissement sous la forme d'aspect avec des attributs dimensionnels (qui est la combinaison de deux caractéristiques de modèles existants), l'enrichissement multi-interprétation et multi-niveau (qui sont des caractéristiques existantes dans certains modèles) et la gestion de versions des attributs (caractéristique qui n'existe pas dans les modèles de la littérature, à notre connaissance). Pour chacune de ces caractéristiques, nous donnons un exemple d'instanciation du modèle dans cette partie. Ces exemples sont basés sur le scénario de motivation présenté en partie 1.1.2.

Aspects avec attributs dimensionnels

Notre modèle utilise la notion d'aspect, introduite dans le modèle MASTER [Mello et al., 2019], pour représenter les données d'enrichissement. Quatre classes ont été reprises de ce modèle, les classes *Aspect*, *AspectType*, *AspectAttribute* et *Value*. Nous avons étendu la classe *Value* avec les classes *SpatialValue*, *TemporalValue* et *ThematicValue* qui permettent de définir des attributs avec des dimensions spécifiques. L'idée de décrire les aspects à l'aide d'attributs dimensionnels vient de l'ontologie STEP [Nogueira and Martin, 2015] dans lequel les épisodes peuvent avoir une portée spatiale, une portée temporelle ou une portée spatio-temporelle. Cependant, contrairement à l'ontologie STEP, nous utilisons les dimensions pour décrire les aspects afin d'améliorer la représentation des phénomènes du monde réel et de donner plus de détails aux enrichissements.

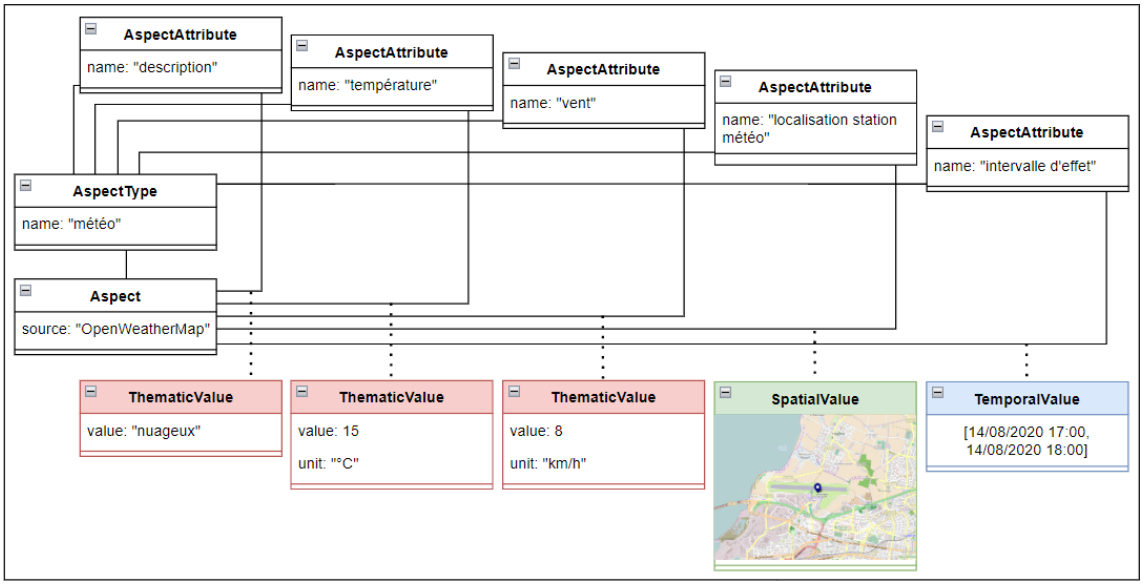


Figure 3.3 – Instanciation du modèle décrivant une donnée météo

La figure 3.3 montre une instanciation du modèle décrivant la météo de la Rochelle le 14/08/2020 entre 17h00 et 18h00. Il s'agit, ici, de l'aspect *nuageux* détaillé dans la figure 1.4, trajectoire 1. La classe *Aspect* est la classe autour de laquelle est construite une donnée d'enrichissement. C'est elle qui fait le lien entre la donnée d'enrichissement et un épisode de la trajectoire. Ici, cette classe nous apprend que la donnée d'enrichissement provient de l'API OpenWeatherMap [OpenWeather]. La classe *AspectType* définit le type de la donnée d'enrichissement qui a pour nom "météo". Cette donnée a cinq attributs *AspectAttribute* différents : trois d'entre eux sont des attributs thématiques *ThematicValue*, un autre est un attribut spatial *SpatialValue* et le dernier est un attribut temporel *TemporalValue*. Les attributs nommés "description", "température" et "vent" correspondent respectivement à un label textuel décrivant la météo, la température de l'air et la vitesse du vent et ont respectivement pour valeurs "nuageux", "15°C" et "8km/h". L'attribut "localisation station météo" correspond, comme son nom l'indique, à la localisation de la station météo d'où provient le relevé météo. Il s'agit d'une valeur spatiale, et plus précisément du point [46.1780556, -1,1930555]. Pour finir, l'attribut "intervalle d'effet" correspond à la durée pendant laquelle cette météo était effective et il a pour valeur [14/08/2020 17 :00, 14/08/2020 18 :00].

Nous avons choisi de représenter les données d'enrichissement sous la forme d'aspects car c'est un outil souple qui permet de représenter n'importe quel phénomène du monde réel. Cela s'avère utile pour enrichir les trajectoires avec des données hétérogènes et complexes. Nous avons étendu la notion d'aspect pour y intégrer des attributs dimensionnels. Cet ajout permet de décrire plus précisément un aspect, en lui ajoutant, par exemple, un lieu ou des horaires d'ouverture. De plus, cela permet de faciliter l'enrichissement en comparant les attributs spatiaux et/ou temporels des aspects avec les positions et/ou les horodatages des trajectoires.

Enrichissement multi-interprétation et multi-niveau

Le processus pour enrichir une trajectoire, c.-à-d. créer une interprétation liée à la trajectoire, se déroule en trois étapes. Tout d'abord, un ou plusieurs types d'aspect sont choisis pour créer une interprétation afin d'enrichir la trajectoire. Ensuite, le module d'enrichissement vérifie si chaque position de la trajectoire peut être liée aux aspects de ce ou ces types, soit grâce à un critère temporel (c.-à-d. correspondance entre un attribut temporel de l'aspect et l'horodatage de la position), soit spatial (c.-à-d. correspondance entre un attribut spatial de l'aspect et la localisation de la position), soit les deux. Puis, les positions consécutives correspondant aux mêmes aspects sont utilisées pour construire un épisode. Enfin, l'interprétation de la trajectoire (c.-à-d. séquences d'épisodes) relative aux types choisis est créée. Toutes les interprétations construites de cette façon permettent de donner plusieurs perspectives d'analyse des trajectoires, comme cela est fait dans les trajectoires de vie [Gensel et al., 2020]. Les interprétations peuvent être construites sur plusieurs niveaux de détail, c'est-à-dire que les épisodes peuvent être imbriqués les uns dans les autres. Nous nous sommes inspirés de l'ontologie STEP [Nogueira et al., 2018] pour introduire cette caractéristique multi-niveau à notre modèle. La hiérarchisation des épisodes (et des données de manière générale) permet de mieux appréhender les informations en allant du grain le plus général au grain le plus fin.

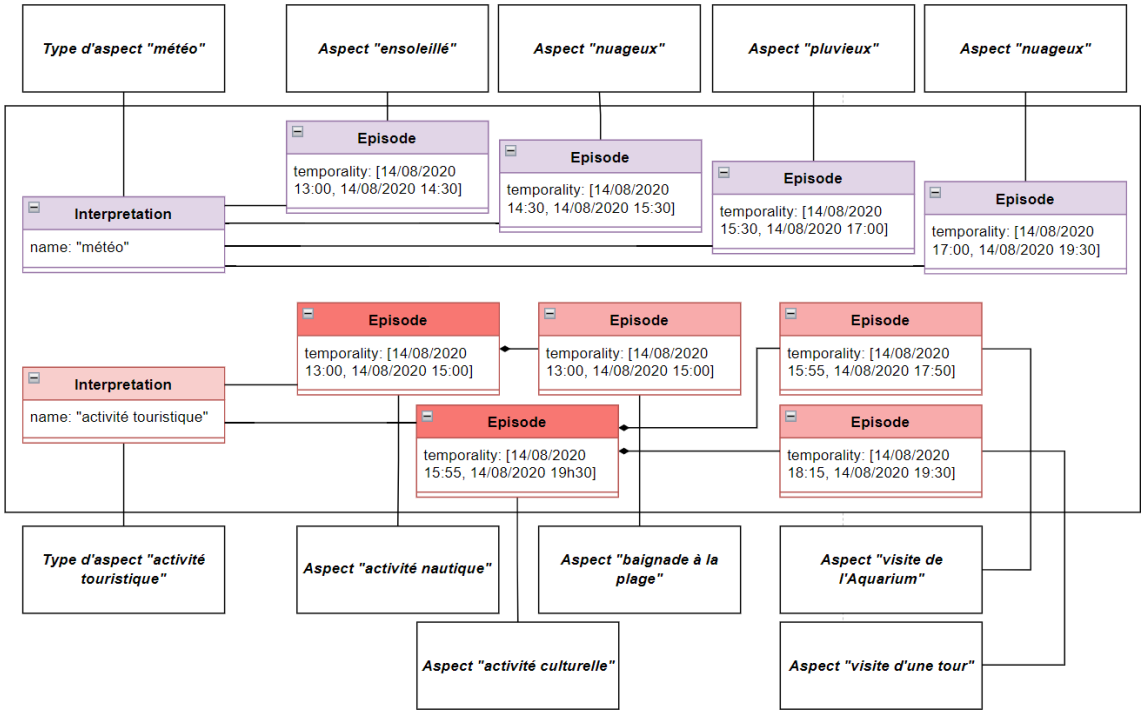


Figure 3.4 – Instanciation du modèle décrivant deux interprétations

La figure 3.4 montre une instanciation du modèle représentant deux interprétations, l'une basée sur la météo et l'autre sur les activités touristiques. Il s'agit, ici, des interprétations *météo* et *activité touristique* détaillées dans la figure 1.4, trajectoire 1. La classe *Interpretation* représente une interprétation de la trajectoire, c.-à-d. une séquence d'épisodes basée sur un ou plusieurs types d'aspects choisis. La classe *Episode* est une classe définissant l'intervalle temporel, au sein de la trajectoire, qui correspond à un aspect donné. L'interprétation *météo*

n'a qu'un seul niveau de détail, les épisodes s'enchaînent séquentiellement (p. ex. l'épisode lié à l'aspect "ensoleillé" est suivi par l'épisode lié à l'aspect "nuageux", etc.), alors que l'interprétation *activité touristique* est multi-niveau, certains épisodes en détaillent d'autres (p. ex. l'épisode lié à l'aspect "activité culturelle" est composé des épisodes liés aux aspects "visite de l'Aquarium" et "visite d'une tour"). Dans cet exemple, les deux interprétations montrées sont basées sur un type d'aspect chacune mais il est possible de construire une interprétation sur un ensemble de types d'aspect.

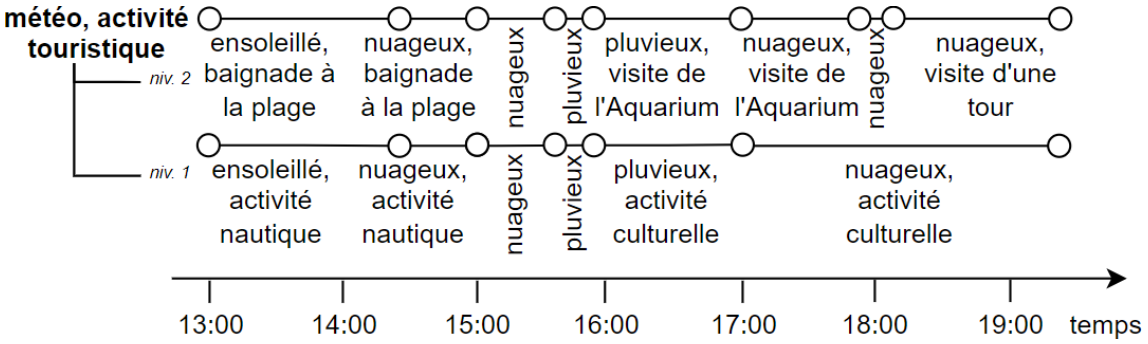


Figure 3.5 – Interprétation à deux critères basée sur les types d'aspects "météo" et "activité touristique"

La figure 3.5 montre une interprétation à deux critères et à deux niveaux, basée sur la météo et les activités touristiques.

Le fait de lier les données d'enrichissement aux trajectoires à l'aide d'interprétations et d'épisodes à plusieurs avantages. Le premier est de permettre la création d'interprétations personnalisées des trajectoires pour faciliter leur analyse. Le second concerne la construction de chaque interprétation qui se base sur un ou plusieurs types d'aspect pour plus de flexibilité. Enfin, le troisième est de permettre de préciser des épisodes par d'autres épisodes sur plusieurs niveaux de détail.

Gestion de versions

Nous avons remarqué que certaines données d'enrichissement peuvent changer au cours du temps. Ce changement se traduit, dans notre modèle, par une modification d'un ou de plusieurs attributs. Nous avons souhaité gérer les différentes versions de chaque attribut pour éviter d'instancier plusieurs fois l'intégralité d'un aspect, à la manière des graphes temporels [Andriamampianina et al., 2021].

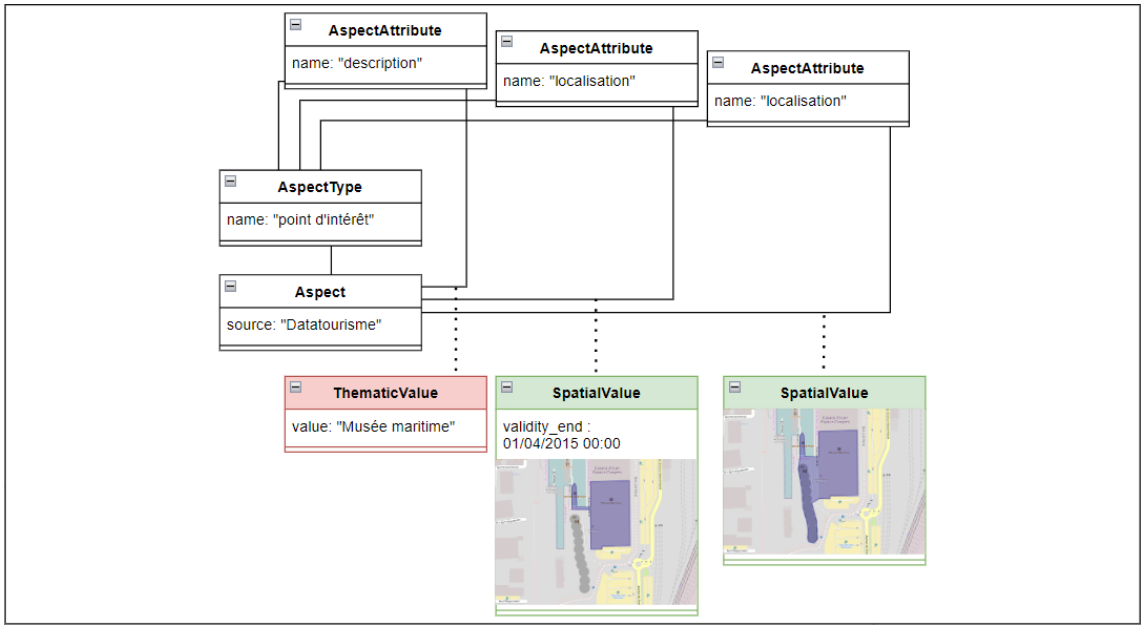


Figure 3.6 – Gestion de versions d'un attribut

La figure 3.6 montre une instanciation du modèle représentant un aspect de type "point d'intérêt". Il s'agit d'un aspect de type "point d'intérêt" dont l'attribut "description" a pour valeur thématique "Musée maritime". Ici, l'attribut "localisation" a été instancié à deux reprises. Les deux valeurs associées correspondent aux différentes versions de cet attribut. Comme nous le constatons sur la durée de validité de l'instance la plus ancienne, le musée s'est agrandi à partir du 01/04/2015.

Dans le projet DA3T, nous travaillons avec beaucoup de données et l'intégration de la gestion de versions à notre modèle permet d'éviter les surcharges de données. En effet, à la place d'avoir plusieurs instances d'un même aspect avec un unique attribut de différence, nous pouvons simplement avoir plusieurs instances de cet attribut liées à des durées de validité spécifiques.

Dans la partie suivante, nous présentons la structuration de notre modèle dans le langage JSON.

3.1.4 Structuration JSON

Le modèle DA3T est utilisé dans la plateforme DA3T sous la forme de fichier utilisant les normes JSON.

Listing 3.1 – Structure du fichier basé sur le modèle de trajectoire sémantique

```
{
    "raw_track" : {
        "type": "FeatureCollection",
        "features": [{
            "type": "Feature",
```

```
6             "geometry": {
7                 "type": "Point",
8                 "coordinates": [102.0, 0.5]
9             },
10            "properties": {
11                "mobile_object_id": 123456,
12                "timestamp": "2020-08-10T10:52:39", ...
13            }
14        }, ...]
15    },
16
17    "modified_trajectory" : {
18        "type": "FeatureCollection",
19        "features": [{
20            "type": "Feature",
21            "geometry": {
22                "type": "Point",
23                "coordinates": [102.0, 0.5]
24            },
25            "properties": {
26                "mobile_object_id": 123456,
27                "trajectory_id": 1,
28                "timestamp": "2020-08-10T10:52:39",
29                "aspects": [{
30                    "attributes": [{
31                        "name": ...,
32                        "type": ...,
33                        "value": ...
34                    }, ... ],
35                    "description": ...,
36                    "source": ...,
37                    "type_name": ...
38                }, ...]
39            }
40        }, ...]
41    }
42 }
```

Le code 3.1 montre la structure type d'un fichier basé sur le modèle. Elle est divisée en deux parties distinctes, à savoir les traces brutes *raw_track* (cf. code 3.1, lignes 1 à 15 du fichier) et les trajectoires brutes ou sémantiques *modified_trajectory* (cf. code 3.1, lignes 17 à 45 du fichier).

La partie concernant les traces brutes est structurée comme un simple fichier GeoJSON et elle est instanciée au début d'une chaîne de traitement, dès la phase de pré-traitement.

Chaque *feature* correspond à une position. Une position a une *geometry* qui correspond à sa localisation et des *properties* qui contiennent, à minima, *mobile_object_id* l'identifiant d'objet mobile à laquelle la position appartient et *timestamp* l'horodatage de la position. C'est également dans les propriétés que nous pouvons retrouver les données brutes (p. ex. la vitesse de déplacement, la précision de la capture, l'altitude, etc.).

La partie concernant les trajectoires brutes et sémantiques est également structurée comme un fichier GeoJSON et elle est instanciée dans un module de construction de trajectoire et/ou durant la phase d'enrichissement. Chaque *feature* correspond à une position avec une *geometry* qui correspond à sa localisation et des *properties* qui contiennent, en plus de toutes les données brutes, les *aspects* enrichissant les trajectoires sémantiques (cf. code 3.1, lignes 29 à 41 du fichier). Chaque aspect est composé de *description* correspondant à la description facultative de l'aspect, *source* correspondant à la source de données d'où provient l'aspect, *type_name* correspondant au type de l'aspect (p. ex. météo, quartier, point d'intérêt, etc.) et *attributes* correspondant à sa liste d'attributs. Un attribut est décrit par *name* le nom de l'attribut, *type* le type de l'attribut (p. ex. chaîne de caractères, instant, polygone, etc.) et *value* la valeur de l'attribut.

Toutes les données de mobilité circulant dans les chaînes de traitement exécutées sur la plateforme ont cette structure. Homogénéiser les formats et les structures au sein d'une chaîne permet de simplifier les traitements en minimisant les conversions. Cependant, pour chaque nouvelle source de données intégrée à la plateforme, un nouveau module d'extraction doit être développé. Ce module permet de créer un fichier ayant la structure du modèle à partir de ces nouvelles données.

3.1.5 Synthèse

Dans cette partie, nous avons présenté le modèle de trajectoire sémantique DA3T issu du verrou de recherche **(V1)** et de l'hypothèse **(H1)**.

Ce modèle est utilisé comme modèle de transition entre les modules de la plateforme ETL DA3T et permet, ainsi, de représenter les traces de mobilité à n'importe quels moments d'une chaîne de traitement (c.-à-d. traces de mobilité, trajectoires brutes et trajectoires sémantiques). Ce modèle combine des caractéristiques multi-interprétation (c.-à-d. l'enrichissement d'une trajectoire se fait grâce à une ou plusieurs interprétations de la trajectoire, une séquence d'épisodes enrichie ayant une thématique particulière), multi-niveau (c.-à-d. une interprétation issue d'une segmentation peut être détaillée sur plusieurs niveaux) et multi-aspect (c.-à-d. les données d'enrichissement sont présentées sous la forme d'aspect, des objets complexes permettant de représenter n'importe quels phénomènes du monde réel).

Pour plus de détails, nous avons publié deux articles au sujet du modèle DA3T [Cayèré et al., 2021a,b]. Développons maintenant à notre seconde contribution à savoir, deux mesures de similarité entre trajectoires sémantiques.

3.2 Mesures de calcul de similarité entre trajectoires sémantiques DA3T

Notre seconde contribution consiste en deux mesures permettant de calculer automatiquement la similarité entre deux trajectoires sémantiques de deux façons différentes en s'inspirant de l'évaluation manuelle d'experts.

3.2.1 Exemple d'évaluation de la similarité de deux trajectoires

Dans notre projet, nous avons eu l'occasion d'observer des experts en géographie du tourisme comparer des paires de trajectoires sémantiques de touristes. Lors de leur évaluation, ils s'intéressent aux trois dimensions des deux trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique). Nous avons noté que les dimensions étaient soit analysées de manière individuelle par certains géographes (p. ex. les empreintes spatiales des deux trajectoires sont comparée, puis les intervalles temporels des deux trajectoires sont comparées), soit analysées de manière combinée par d'autres (p. ex. l'enrichissement thématique et les intervalles temporels associés sont comparés simultanément). De plus les géographes comparent naturellement les trajectoires selon plusieurs niveaux de granularité (p. ex. pour la dimension spatiale, les géographes comparent d'abord les trajectoires à l'échelle de la région et de la ville puis à l'échelle des quartiers et enfin à l'échelle des rues et des arrêts des touristes).

Les paires de trajectoires sémantiques présentées en annexes E montrent la dimension spatiale des trajectoires sur une carte et les dimensions temporelle et thématique sur une frise chronologique. Par exemple, les trajectoires sémantiques n°71 et n°107 forment la paire n°13. La dimension spatiale de ces trajectoires sémantiques est présentée sur la figure E.13 et les dimensions thématique et temporelle sont présentées sur les figures E.42 pour la trajectoire n°71 et E.47 pour la trajectoire n°107. En s'appuyant sur ces supports visuels, les géographes peuvent comparer manuellement les deux trajectoires sémantiques. C'est un travail fastidieux car il y a trois dimensions et beaucoup de facteurs de comparaison à prendre en compte pour chaque dimension. Les experts donnent plus ou moins d'importance à chaque facteur en se basant sur leurs connaissances dans le domaine. Nous avons constater que les géographes s'intéressaient moins aux déplacements exacts qu'aux points et zones d'intérêt visités par les touristes. Malgré les différences notables entre les deux déplacements de la paire n°13, les experts accordent beaucoup d'importance aux zones d'intérêt communes comme celles de l'Aquarium de La Rochelle et de la promenade du Vieux-Port, visibles sur la figure E.13, lorsqu'ils comparent les deux trajectoires.

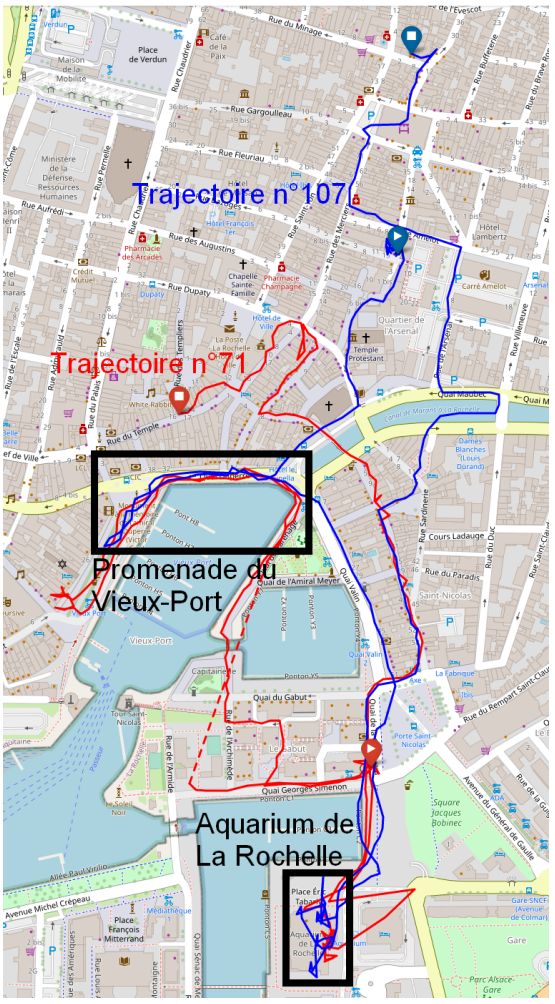


Figure 3.7 – Zones d'intérêt communes entre les trajectoires n°71 et n°107

En guise d'exemple des résultats produits par une mesure de similarité, nous avons sélectionné différentes mesures de similarité de l'état de l'art et nous les avons exécuté sur la paire de trajectoires sémantiques n°13 pour évaluer la similarité spatiale entre ces deux trajectoires. Les mesures de similarité entre séries temporelles sélectionnées intègrent la distance de Haversine pour calculer les distances entre deux positions. Notons que ces mesures ne produisent pas forcément des résultats sur un intervalle $[0, 1]$. Pour que les résultats soient comparables, nous les avons normalisé sur cet intervalle. Dans un premier temps, nous avons exécuté chaque mesure sur l'ensemble des paires de trajectoires de notre jeu de données et nous avons sélectionné le plus haut score. Dans un second temps, nous avons divisé l'intégralité des scores de notre jeux de données par ce score maximal. Notons également que certaine de ces mesures sont des mesures de similarité et d'autres sont des mesures de distance. Pour les comparer, nous avons transformé tous les scores en scores de similarité (c.-à-d. plus le score s'approche de 0 moins les trajectoires se ressemblent et plus il s'approche de 1 plus elles se ressemblent) à l'aide de l'équation 2.1.

Table 3.1 – Résultats des mesures de l'état de l'art exécutées sur la paire n°13

Mesures de similarité spatiale	**Échelle de comparaison**	**Résultat**
TRACLUS	Lignes	0,94
DTW	Points	0,93
ERP	Points	0,83
EDR	Points	0,52
LCSS	Points	0,47
RI spatiale	Polygones	0,03

Les résultats des mesures sélectionnées sont classés du plus élevé au plus faible dans le tableau 3.1. Nous constatons que les mesures sont assez partagées sur la similarité ou la non-similarité des deux trajectoires. TRACLUS, DTW et ERP donnent des scores très élevés, contrairement à la mesure de RI spatiale qui donne un score très faible. EDR et LCSS donne des scores assez similaires, proche de la moyenne. Nous en concluons que, selon l'échelle de comparaison, les résultats changent radicalement. À l'échelle des polygones avec la mesure de RI spatiale, ces deux trajectoires ne sont pas du tout similaires alors qu'à l'échelle des lignes avec TRACLUS, elles sont très similaires.

Nous souhaitons créer une mesure de similarité prenant en compte les subtilités de la comparaison de deux trajectoires sémantiques faite par des experts. Pour cela, nous nous appuyons sur le verrou et les hypothèses rappelés dans la partie suivante.

3.2.2 Rappel des hypothèses et verrous de recherche

Notre seconde contribution concerne un module de la plateforme permettant de calculer la similarité entre deux trajectoires sémantiques. À notre connaissance il n'existe pas de mesure de similarité entre trajectoires sémantiques intégrant les dimensions spatiale, temporelle et thématique tout en cherchant à imiter l'analyse que ferait un expert dans le domaine des trajectoires cibles (p. ex. tourisme pour les trajectoires de touristes, ornithologie pour des trajectoires d'oiseaux, etc.). Le verrou **(V2)** décrit le besoin d'un telle mesure de similarité entre trajectoires sémantiques intégrant les trois dimensions et dont les résultats s'approchent de l'avis d'experts.

Pour résoudre le verrou **(V2)**, nous émettons deux hypothèses : l'hypothèse **(H2.1)** de combiner des mesures de similarité spatiale, temporelle et thématique sur différents niveaux de granularité micro, méso et macro afin de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes et l'hypothèse **(H2.2)** de combiner deux mesures de similarité bidimensionnelles (p. ex. spatio-temporelle et tempo-thématique) afin de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes. Ces deux hypothèses ciblent différentes approches de comparaison faites par des experts que nous avons relevées lors de nos discussions avec eux. Certains experts préfèrent analyser les dimensions des trajectoires sémantiques de manière individuelle avant de se prononcer sur un score global et d'autres préfèrent considérer les dimensions simultanément pour évaluer la similarité de deux trajectoires. Nous considérons qu'une mesure de similarité a des performances supérieures à une autre si elle est plus proche de l'avis des

experts.

Les parties suivantes présentent les deux contributions découlant des hypothèses **(H2.1)** et **(H2.2)**.

3.2.3 Mesure unidimensionnelle sur plusieurs niveaux de granularité

Pour valider l'hypothèse **(H2.1)**, nous mettons en œuvre une mesure de similarité qui combine des sous-mesures spatiales, temporelles et thématiques pondérées par des coefficients. Chacune de ces sous-mesures est la combinaison de trois mesures de niveaux de granularité différents également pondérées par des coefficients (cf. tableau 1.1). Ainsi, par exemple, la dimension spatiale d'une trajectoire est considérée successivement au niveau des points capturés (grain micro, cf. tableau 1.1, figure 1), au niveau des segments ou des lignes (grain méso, cf. tableau 1.1, figure 2) et au niveau des boîtes ou polygones englobants (grain macro, cf. tableau 1.1, figure 3).

Notre première mesure $DA3T\_S1_{glb}$ est définie par l'équation 3.1.

$$DA3T\_S1_{glb} = \alpha_{spt} * S_{spt} + \beta_{tmp} * S_{tmp} + \gamma_{thm} * S_{thm} \tag{3.1}$$

Dans cette formule, chaque sous-mesure du calcul de similarité liée à une dimension spécifique peut être de nouveau détaillée en trois nouvelles sous-mesures liées à des granularités différentes. Le développement de l'équation 3.1 est détaillé par les équations 3.2, 3.3 et 3.4.

$$S_{spt} = \alpha_{spt-mic} * S_{spt-mic} + \beta_{spt-mes} * S_{spt-mes} + \gamma_{spt-mac} * S_{spt-mac} \tag{3.2}$$

$$S_{tmp} = \alpha_{tmp-mic} * S_{tmp-mic} + \beta_{tmp-mes} * S_{tmp-mes} + \gamma_{tmp-mac} * S_{tmp-mac} \tag{3.3}$$

$$S_{thm} = \alpha_{thm-mic} * S_{thm-mic} + \beta_{thm-mes} * S_{thm-mes} + \gamma_{thm-mac} * S_{thm-mac} \tag{3.4}$$

La somme des coefficients de pondération d'un même niveau (p. ex. α_*, β_* and γ_*) est toujours égale à 1 telle que : $\alpha_{spt} + \beta_{tmp} + \gamma_{thm} = 1$, $\alpha_{spt-mic} + \beta_{spt-mes} + \gamma_{spt-mac} = 1$, $\alpha_{tmp-mic} + \beta_{tmp-mes} + \gamma_{tmp-mac} = 1$ et $\alpha_{thm-mic} + \beta_{thm-mes} + \gamma_{thm-mac} = 1$. De plus toute mesure de similarité S est telle que : $0 \leq S \leq 1$.

Parmi neuf sous-mesures de l'équation 3.1, nous créons trois nouvelles mesures de similarité $S_{spt-mes}$ (c.f. Tableau 1.1, b), $S_{tmp-mac}$ (c. f. Tableau 1.1, h) et $S_{thm-mes}$ (c.f. Tableau 1.1, f) et mettons en œuvre six versions modifiées des mesures existantes. Premièrement, les mesures de la dimension spatiale (cf. équation 3.2) sont les suivantes :

— $S_{spt-mic}$: Pour comparer les trajectoires au niveau des points, nous adaptons la DTW [Vintsyuk, 1968] et nous y intégrons la distance Haversine [Andrew, 1805]. DTW se concentre sur les modèles de trajectoires et ne prend en compte ni le temps de déplacement, ni la vitesse de déplacement (ou le changement de vitesse), ni la fréquence de capture des positions. DTW a besoin de la distance entre chaque paire de points pour calculer la similarité globale des deux trajectoires et nous avons choisi d'utiliser la distance Haversine pour cela. La distance d'Haversine mesure la distance entre deux

points GPS.

— $S_{spt-mes}$: Pour comparer les trajectoires au niveau des lignes, nous créons un nouvel algorithme qui utilise TRACLUS [Lee et al., 2007]. TRACLUS permet de comparer deux segments en pondérant trois sous-mesures, chacune s'intéressant à une caractéristique particulière des segments (c.-à-d. leur parallélisme, leur distance et leur angle). Comme nous souhaitons comparer deux trajectoires, notre algorithme compare deux à deux les segments des trajectoires à l'aide de TRACLUS. Pour chacun des segments de la première trajectoire, il sélectionne le segment de la seconde qui est le plus ressemblant (c.-à-d. qui donne le meilleur score TRACLUS) sans contrainte d'ordre. La somme des scores toutes les meilleures paires correspond au score global des deux trajectoires. L'équation 3.5 formalise mathématiquement notre mesure :

$$spt\_mes(R,S) = \sum_{sgm_i \in R} min_{sgm_j \in S}(traclus(sgm_i, sgm_j)) \tag{3.5}$$

Avec R et S deux trajectoires, sgm_i et sgm_j les segments des deux trajectoires et $traclus(sgm_i, sgm_j)$ le score produit par TRACLUS pour les segments sgm_i et sgm_j. Le code 3.2 montre le code Python de notre sous-mesure $S_{spt-mes}$.

Listing 3.2 – Code Python de la sous-mesure $S_{spt-mes}$

```
def get_spt_mes(trj_1, trj_2, a, b, c):
    score = 0
    is_first = True
    segment_1 = ["", ""]
    segment_2 = ["", ""]

    for point_1 in trj_1:
        min_distance = INF
        segment_1[1] = point_1

        for point_2 in trj_2:
            segment_2[1] = point_2

            if (segment_1[0] != "")
            and (segment_2[0] != "")
            and (segment_1[0] != segment_1[1])
            and (segment_2[0] != segment_2[1]):
                distance = traclus(segment_1, segment_2, a, b, c)

                if(min_distance > distance):
                    min_distance = distance

            segment_2[0] = point_2

        if is_first == False:
```

```
                score = score + min_distance

            segment_1[0] = point_1
            is_first = False

        return score
```

Avec trj_1 et trj_2 les trajectoires comparées, a, b et c les trois coefficients de TRACLUS destinés à pondérer les sous-scores évaluant le parallélisme, la distance et l'angle des deux segments.

— $S_{spt-mac}$: Pour comparer les trajectoires au niveau des boîtes englobantes, nous implémentons la mesure de similarité appliquée à la RI spatiale Le Parc-Lacayrelle et al. [2007] qui utilise l'intersection de deux polygones pour calculer leur distance. Pour cette sous-mesure, nous avons choisi de travailler sur des polygones n'ayant aucune sémantique (contrairement p. ex. à des polygones représentant les quartiers d'une ville) afin de pouvoir l'exécuter sur n'importe quel jeu de trajectoires sémantiques (p. ex. trajectoires d'oiseaux). Nous pensons, dans la suite de nos travaux, faire évoluer cette sous-mesure pour se baser sur un découpage de l'espace en grille et comparer les emprises des trajectoires sur cette grille, pour rendre les résultats plus précis.

Ensuite, concernant la dimension temporelle, nous ne traitons que des trajectoires journalières et nos sous-mesures sont donc adaptées pour des trajectoires de cette durée ou d'une durée plus courte. Elles ne traitent donc pas les cas limites d'une trajectoire qui commence un jour pendant l'après-midi et termine le lendemain matin. Les mesures de similarité temporelle (cf. équation 3.3) sont les suivantes :

— $S_{tmp-mic}$: Pour comparer les trajectoires au niveau des horodatages, nous attribuons d'abord une période de la journée à chaque horodatage. Nous avons découpé une journée en six périodes distinctes, à savoir : tôt le matin de 00 :00 à 06 :00, petit-déjeuner de 06 :00 à 09 :00, matin de 09 :00 à 12 :00, déjeuner de 12 :00 à 14 :00, après-midi de 14 :00 à 19 :00 et dîner de 19 :00 à 21 :00 et soirée de 21 :00 à 00 :00. Nous lions ensuite chaque horodatage des trajectoires à la période qui lui correspond ce qui forme des séquences de périodes. Nous mettons en œuvre la mesure EDR [Chen et al., 2005] pour comparer deux séquences de périodes associées aux trajectoires. Par exemple, dans la figure 1.3, la trajectoire 1 a ses horodatages associés, dans l'ordre, aux périodes déjeuner et après-midi et la trajectoire 2 aux périodes matin, déjeuner et après-midi. Nous considérons qu'il y a correspondance entre deux périodes de temps si elles sont exactement égales.

— $S_{tmp-mes}$: Pour comparer les trajectoires au niveau des intervalles temporels, nous réduisons les intervalles temporels des trajectoires à une échelle quotidienne. Par exemple, dans la figure 1.3, la trajectoire 1 de 13h à 19h le 14/07/2020 et la trajectoire 2 de 9h à 15h le 23/08/2020 sont résumées comme deux trajectoires de 13h à 19h et de 9h à

15h respectivement, quelque soit le jour précis. Nous appliquons ensuite la mesure de similarité temporelle appliquée à la RI Le Parc-Lacayrelle et al. [2007] qui utilise l'intersection entre deux intervalles temporels pour calculer leur distance.

— $S_{tmp-mac}$: Pour comparer les trajectoires au niveau des contextes temporels, nous créons une nouvelle mesure de similarité pour comparer les vecteurs de données temporelles. Un vecteur de données temporelles est associé à la trajectoire entière (p. ex. "année : 2020, saison : automne, mois : 11, etc."). Ainsi, nous comparons deux trajectoires sur leurs vecteurs de données. Chaque élément des vecteurs qui diffère apporte une pénalité au score. Les valeurs des pénalités sont laissées au choix de l'utilisateur. L'équation 3.6 formalise mathématiquement notre mesure :

$$tmp\_mac(R,S,i) = \begin{cases} p_i + tmp\_mac(R,S,i+1) & \text{si } R.vect_i \neq S.vect_i \\ 0 & \text{sinon} \end{cases} \tag{3.6}$$

Avec p_i la pénalité associée à l'élément i du vecteur de données temporelles, $R.vect_i$ et $S.vect_i$ les éléments i des vecteurs associés respectivement aux trajectoires R et S. Le code 3.3 montre le code Python de notre sous-mesure $S_{tmp-mac}$.

Listing 3.3 – Code de la sous-mesure $S_{tmp-mac}$

```
def get_tmp_mac(tsp_1, tsp_2, penalties):
    score = 0

    vector_1 = [get_year(tsp_1),
                get_season(tsp_1),
                get_month(tsp_1),
                is_weekend(tsp_1),
                get_weekday(tsp_1),
                get_day_number(tsp_1)]
    vector_2 = [get_year(tsp_2),
                get_season(tsp_2),
                get_month(tsp_2),
                is_weekend(tsp_2),
                get_weekday(tsp_2),
                get_day_number(tsp_2)]

    if(vecteur_1[0] != vecteur_2[0]):
        score = score + penalties[0] * abs(vector_2[0] - vector_1[0])

    if(vecteur_1[1] != vecteur_2[1]):
        score = score + penalties[1]

    if(vecteur_1[2] != vecteur_2[2]):
        score = score + penalties[2]

    if(vecteur_1[3] != vecteur_2[3]):
```

```
        score = score + penalties[3]

    if(vecteur_1[4] != vecteur_2[4]):
        score = score + penalties[4]

    if(vecteur_1[5] != vecteur_2[5]):
        score = score + penalties[5]

    return score
```

Avec tsp_1 et tsp_2 les horodatages des premiers points des deux trajectoires, $penalties$ un tableau contenant les pénalités associées à chaque élément des vecteurs de données temporelles et définies par l'utilisateur. L'année présente un cas particulier car la pénalité qui lui est associée est multipliée par la distance entre les années des deux trajectoires. Nous avons mis en œuvre notre mesure uniquement sur des trajectoires journalières et le vecteur de données que nous avons jugé intéressantes pour comparer ces trajectoires est [année, saison, mois, weekend/pas weekend, jour de la semaine, numéro du jour de la semaine]. Dans un contexte avec des trajectoires plus longues (p. ex. de plusieurs jours), il faudrait changer les valeurs du vecteur pour qu'elles correspondent au contexte.

Enfin, les mesures de similarité thématique (cf. équation 3.4) sont les suivantes :

— $S_{thm-mic}$: Pour comparer les trajectoires au niveau des aspects, nous implémentons MUITAS May Petry et al. [2019] car il nous permet de comparer les aspects en prenant en compte tous les attributs de chaque aspect. Cette mesure permet de comparer deux trajectoires en s'appuyant sur points enrichis avec des aspects. Ici, nous utilisons l'ensemble des positions brutes des trajectoires associées aux aspects par le biais de leur dimension spatiale et/ou temporelle. Par exemple, dans la figure 1.3, les premières positions des trajectoires 1 et 2 ont une similarité avec l'aspect "plage des Minimes" et une différence avec les aspects "nuageux" et "ensoleillé" respectivement.

— $S_{thm-mes}$: Pour comparer deux trajectoires au niveau des séquences d'épisodes, nous créons une nouvelle mesure de similarité pour comparer deux épisodes et nous l'intégrons dans la mesure EDR [Chen et al., 2005]. Les épisodes peuvent être simples (p. ex. épisode lié à un unique aspect "nuageux") ou composés (p. ex. épisode lié à plusieurs aspects "nuageux" et "aquarium de La Rochelle"). Ainsi, si les deux épisodes sont simples, notre mesure renvoie 1 lorsque les aspects sont égaux et 0 sinon. Dans le cas d'épisodes composés, la mesure renvoie la moyenne des scores obtenus par chaque

paire d'aspects. L'équation 3.7 formalisent mathématiquement notre mesure.

$$episodes\_similarity(eps\_1, eps\_2) = \frac{\sum_{asp_i \in eps\_1 \text{ et } asp_j \in eps\_2}(aspects\_similarity(asp_i, asp_j))}{max(len(eps\_1), len(eps\_2))} \tag{3.7}$$

Avec eps_1 et eps_2 les épisodes comparés, asp_i les aspects composant le premier épisode, asp_j les aspects composant le second épisode et $aspects\_similarity(asp_i, asp_j)$ le score de similarité des aspect asp_i et asp_j dont la définition mathématique est spécifié par l'équation 3.8.

$$aspects\_similarity(asp\_1, asp\_2) = \begin{cases} 1 & \text{si } asp\_1 = asp\_2 \\ 0 & \text{sinon} \end{cases} \tag{3.8}$$

Avec asp_1 et asp_2 les aspects comparés. Le code 3.4 montre le code Python de notre mesure de similarité permettant de comparer deux épisodes.

Listing 3.4 – Code de la mesure de similarité permettant de comparer deux épisodes

```
def get_episodes_similarity(episode_1, episode_2):
    score = 0

    if(episode_1 == episode_2):
        score = 1

    else:
        for aspect_1 in episode_1:
            for aspect_2 in episode_2:
                if(aspect_1 == aspect_2):
                    score = score + 1

        score = score / max(len(episode_1), len(episode_2))

    return score
```

Une amélioration que nous souhaitons mettre en place pour faire évoluer cette mesure est la prise en compte de la sémantique des aspects. En effet, nous attribuons un score de similarité de 1 lorsque deux aspects sont identiques et de 0 autrement (p. ex. "nuageux" et "pluie" donne 0 et "ensoleillé" et "pluie" donne 0), ce qui rend la comparaison extrêmement binaire. En utilisant la sémantique des aspect pour les comparer, cela permet d'introduire plus de détails et de modération dans le score. La mise ne place d'une telle amélioration nécessite d'avoir un thésaurus ou une ontologie pour représenter chaque type d'aspect (p. ex. "nuageux" et "pluie" donne un score plus élevé que "ensoleillé" et "pluie" car ces concepts sont plus proches sémantiquement parlant).

— $S_{thm-mac}$: Pour comparer les trajectoires au niveau des thématiques principales, nous les résumons avec la valeur dominante de chaque aspect et utilisons la mesure LCSS,

qui est très adaptée à la comparaison de deux chaînes de caractères. Par exemple, dans la figure 1.3, en termes de météo, la trajectoire 1, qui est résumée par "nuageux", est similaire à la trajectoire 2, qui est également résumée par "nuageux". Ainsi les chaînes de caractères se ressemblant (p. ex. "nuageux" et "peu nuageux") donnent des scores plus élevés que les chaînes de caractères complètement différentes (p. ex. "nuageux" et "ensoleillé"). Comme pour la granularité méso, nous pouvons également améliorer cette sous-mesure pour prendre en compte la sémantique des concepts. Ici aussi, cela implique de prendre en compte un thésaurus ou une ontologie pour chaque type d'aspect.

3.2.4 Mesure bidimensionnelle

Pour valider l'hypothèse **(H2.2)**, nous mettons en œuvre une deuxième mesure de similarité qui combine des sous-mesures spatio-temporelles et tempo-thématiques pondérées par des coefficients. Notre mesure $DA3T\_S2_{glb}$ est définie par l'équation 3.9.

$$DA3T\_S2_{glb} = \alpha_{spt-tmp} * S_{spt-tmp} + \beta_{tmp-thm} * S_{tmp-thm} \tag{3.9}$$

Avec $\alpha_{spt-tmp} + \beta_{-tmp} = 1$ et $0 \leq S \leq 1$.

Dans l'équation 3.9, nous adaptons une mesure existante (c.f. $S_{tmp-thm}$) et implémentons une version modifiée d'une mesure existante (c.f. $S_{spt-tmp}$) :

— $S_{spt-tmp}$: Pour comparer les trajectoires sur la dimension spatio-temporelle, nous implémentons le STLCSS [Vlachos et al., 2002] qui utilise les distances spatiales et temporelles entre deux positions. Un seuil de distance temporelle est ajouté au seuil de distance spatiale existant. Ainsi, il permet d'identifier les trajectoires qui se rapprochent les unes des autres à la même heure de la journée. Par exemple, dans la figure 1.3, les trajectoires 1 et 2 ont un intervalle temporel commun entre 13h et 15h pendant lequel elles ne sont pas au même endroit ; leur score sera donc faible.

— $S_{tmp-thm}$: Pour comparer les trajectoires sur la dimension tempo-thématique, nous adaptons MUITAS [May Petry et al., 2019] afin de prendre en compte les horodatages des aspects, deux aspects correspondent lorsqu'ils sont similaires et proches temporellement. Par exemple, dans la figure 1.3, les trajectoires 1 et 2 ont un intervalle temporel commun entre 13h et 15h pendant lequel les activités et les points d'intérêt sont différents mais la météo et les marées sont similaires.

3.2.5 Synthèse

Dans cette partie, nous avons présenté notre seconde contribution, à savoir, deux mesures de similarité issues du verrou de recherche **(V2)** et des hypothèses **(H2.1)** et **(H2.2)**.

Ces mesures de similarité permettant de comparer deux trajectoires sémantiques sur les trois dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique). La mesure $DA3T\_S1_{glb}$ est une combinaison pondérée de trois sous-mesures unidimensionnelles

(c.-à-d. sous-mesures spatiale, temporelle et thématique) qui sont, à leur tour, des combinaisons pondérées de trois sous-mesures s'intéressant à des niveaux de granularité différents (c.-à-d. micro, méso et macro) pour la dimension en question. La mesure $DA3T\_S2_{glb}$ est une combinaison pondérée de deux sous-mesures bidimensionnelles (c.-à-d. spatio-temporel et tempo-thématique).

Notons que ces mesures sont très adaptées dans notre contexte de plateforme ETL mais ne peuvent pas être réutilisées telles quelles dans d'autres contextes. En effet, elles nécessitent certaines adaptations et/ou certains pré-traitements des données. Par exemple, la normalisation des sous-mesures nécessite de les exécuter sur chaque paire de trajectoires du jeu de données afin de diviser les scores par le score maximal. Un autre exemple concerne le jeu de données que nous avons choisi pour nos expérimentations qui se compose uniquement de trajectoires journalières. Ainsi, pour traiter des trajectoires plus longues dans le temps, certaines adaptations au niveau de sous-mesures temporelles doivent être faites (c.-à-d. changement d'échelle pour les mesures micro et méso et modification du vecteur de données temporelles pour la mesure macro).

Notre objectif est que les scores produits par nos deux mesures s'approchent de l'avis d'experts. Pour cela, nous avons mis en place une expérimentation chapitre 5, section 5.3 pour fixer les coefficients de pondération afin de produire des résultats satisfaisants pour un jeu de données donné.

Pour plus de détails, nous avons publié un article au sujet de ces mesures [Cayèré et al., 2022]. La partie suivante présente le développement de la plateforme DA3T qui sert d'environnement à nos deux contributions.

3.3 Synthèse générale

Ce chapitre a permis de détailler nos deux contributions qui s'intègrent dans la plateforme DA3T. Résumons ces deux contributions.

La première contribution est le modèle de trajectoire sémantique DA3T qui sert de modèle de transition dans les chaînes de traitement de la plateforme DA3T, c'est-à-dire que toutes les données circulant dans les chaînes respectent le format et la structure du modèle. C'est un modèle qui permet de représenter les traces de mobilité à n'importe quelle étape du traitement. Les données d'enrichissement sont sous la forme d'aspects, des objets complexes permettant de représenter n'importe quel phénomène du monde réel. Un aspect est décrit par son type (p. ex. météo, point d'intérêt, activité touristique, etc.) et par un ensemble d'attributs dimensionnels liés au type (p. ex. un monument peut être décrit par son nom, son prix d'entrée, sa localisation, ses heures d'ouverture). Pour enrichir les traces avec des aspects, le modèle utilise la notion d'interprétations de la trajectoire, c'est-à-dire des séquences d'épisodes. Une interprétation est construite sur la base d'un ou de plusieurs types d'aspect. À chaque fois qu'un nouvel aspect du type choisi enrichit la trajectoire, un nouvel épisode et créé et ajouté à l'interprétation. Un épisode peut être détaillé à l'aide d'autres épisodes, ce

qui produit une interprétation à plusieurs niveaux. Notre modèle est générique et extensible.

La seconde contribution regroupe deux mesures de similarité permettant de comparer deux trajectoires sémantiques. La première mesure $DA3T\_S1_{glb}$ combine, à l'aide de coefficients de pondération, des sous-mesures de similarité spatiale, temporelle et thématique ; elle mêmes combinant des sous-mesures de similarité s'intéressant à des niveaux de granularité micro, méso macro. Cela fait un total de neuf sous-mesures, trois pour chaque dimension. Parmi elles, nous en avons créé trois et les autres ont été adaptées de la littérature. La seconde mesure $DA3T\_S2_{glb}$ combine deux sous-mesures de similarité bidimensionnelle dont la dimension pivot et la dimension temporelle. Elles proviennent de la littérature et ont été adaptées à notre contexte. L'objectif de ces deux mesures est que leurs résultats s'approchent le plus de l'avis des géographes.

Nos deux contributions s'intègrent à la plateforme de conception de chaînes de traitement DA3T qui fait l'objet du chapitre suivant.

Chapitre 4

Plateforme de conception de chaînes de traitement DA3T

L'analyse des traces de mobilité est une tâche fastidieuse car elle implique d'exécuter un ensemble de traitements consécutifs sur ces données. Ces traitements peuvent se trouver sur différentes plateformes et ne prennent pas forcément des formats similaires en entrée, ce qui demande un effort de conversion. Nous souhaitons créer une plateforme de création et d'exécution de chaînes de traitement de type **ETL** (abrégé de l'anglais, ***E**xtract, **T**ransform, **L**oad*) dédié au traitement de trace de mobilité et plus spécifiquement des trajectoires sémantiques. Plutôt que d'être une contribution de cette thèse, cette plateforme est plus particulièrement l'objectif final qui exploite nos contributions (c.-à-d. le modèle et les mesures). Ce chapitre se concentre sur la plateforme DA3T en commençant par définir le concept d'ETL (cf. partie 4.1), puis, en énonçant les besoins des géographes et des aménageurs (cf. partie 4.3), ensuite, en présentant l'architecture de la plateforme (cf. partie 4.4) et enfin, en décrivant les parties serveur (cf. partie 4.4.1) et client (cf. partie 4.4.2) de la plateforme.

4.1 ETL dédiés aux données de mobilité

L'ETL décrit un processus comportant plusieurs étapes utilisé pour l'intégration de données, c'est-à-dire le transfert de données provenant de multiples sources hétérogènes (p. ex. fichiers, bases de données, etc.) vers un emplacement unique (p. ex. un entrepôt de données) selon des besoins spécifiques Sreemathy et al. [2020]. Une partie importante de ce processus est la transformation des données sources dans des formats différents en un seul format homogène. Le but est d'utiliser ce format homogène tout au long du processus de traitement.

L'étape d'extraction (en anglais, *Extract*) consiste à récupérer les données de diverses sources avec des formats hétérogènes. Ces données peuvent ensuite être inspectées et validées pour s'assurer que les plages de valeurs attendues sont présentes. En cas de corruption, elles peuvent être renvoyées à la source pour correction. Plusieurs types d'extractions sont possibles (p. ex. données sources complètes, extrait de données sources, etc.) [Sreemathy et al., 2020].
L'étape de transformation (en anglais, *Transform*) a pour but de transformer les différents

formats des données extraites en un format unique qui est conçu en fonction des besoins. Différents types de transformation peuvent être appliqués aux données (p. ex. le nettoyage, le filtrage, la normalisation, la suppression des doublons, la jointure, l'agrégation, etc.).
Enfin, l'étape de chargement (en anglais, *Load*) est une étape de propagation [Chakraborty et al., 2017], c'est-à-dire que les données sont physiquement chargées, écrites dans un autre endroit (généralement, un entrepôt de données). Le premier chargement est appelé chargement initial, et en cas de chargements supplémentaires, ils peuvent être soit incrémentaux (c.-à-d. ajout des données supplémentaires à celles qui existent déjà), soit complets (c.-à-d. rechargement de toutes les données).

Il existe deux principaux types de déploiement pour les outils ETL, à savoir :

— déploiement sur site : ces ETL sont installés sur l'ordinateur de l'utilisateur et se comporte comme des applications de bureau régulières.
— déploiement dans le *cloud* : ces ETL sont accessibles par le biais d'applications web et fonctionnent à distance sur le *cloud*.

De nombreux logiciels permettent de mettre en place des processus ETL. Une courte revue de ces logiciels ETL peut être trouvée dans l'annexe D issue de notre étude exposée dans Masson et al. [2022].

4.2 Scénario de motivation

Cette partie a pour objectif de présenter un scénario de motivation pour développer une telle plateforme. Dans notre projet, nous utilisons les traces de mobilité de touristes volontaires issues de l'application Géoluciole. Les géographes du projet souhaitent répondre à différentes questions sur les données, telles que :

1. **Quels sont les points d'intérêt qui ont intéressés les touristes quand la marée est haute et ceux qui ont intéressés les touristes quand la marée est basse ?**
2. Quelles sont les trajectoires qui sont passées par le quartier X et par le quartier Y dans la même journée ?
3. Quelles sont les activités touristiques des touristes quand il fait beau l'été dans les différents quartiers de La Rochelle ?

Notre scénario de motivation se base sur la question (1). Les deux questions suivantes sont utilisées dans notre expérimentation du modèle dans le chapitre 5.

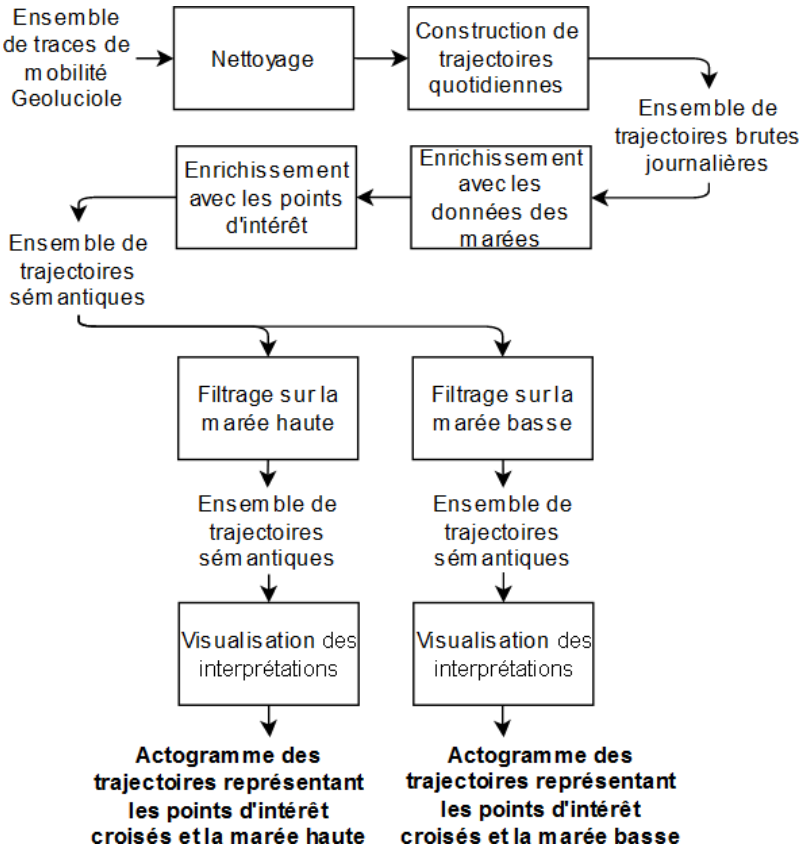


Figure 4.1 – Chaîne de traitement permettant de répondre à la question (1)

La figure 4.1 permet de répondre à cette question (1). Nous souhaitons pouvoir implémenter de telles chaînes de traitement sur notre plateforme ETL. Les rectangles représentent les modules de traitement et les textes non encadrés représentent les données circulant dans la chaîne.

Dans un premier temps, une étape importante du travail des géographes du projet est de nettoyer les traces (cf. module *Nettoyage*). Dû à certains types de matériels de capture (c.-à-d. les téléphones des touristes) ou à la mauvaise captation des signaux GPS, les traces Géoluciole ne sont pas toujours fidèles au déplacement réel et il est souvent nécessaire d'effectuer une suppression des positions aberrantes. Nous nous intéressons ici à des trajectoires quotidiennes ; or, les traces représentent les déplacements complets des touristes et durent souvent plusieurs jours. L'étape suivante est donc de construire les trajectoires brutes quotidiennes à partir des traces de mobilité en les découpant selon les jours (cf. module *Construction de trajectoires quotidiennes*). Ensuite, nous voulons enrichir les trajectoires avec les données relatives aux marées et avec celles relatives aux points d'intérêt de La Rochelle (cf. modules *Enrichissement avec les données des marées* et *Enrichissement avec les points d'intérêt*). Les trajectoires brutes deviennent, à cette étape, des trajectoires sémantiques car elles ont été enrichies avec des données externes. Ensuite, afin de faciliter la comparaison des deux cas mentionnés dans la question, deux filtrages parallèles sont faits : le premier filtre les trajectoires pour n'obtenir que celles qui comprennent un épisode de marée haute et le second filtre les trajectoires pour n'obtenir que celles qui comprennent un épisode de marée basse

(cf. modules *Filtrage sur la marée haute* et *Filtrage sur la marée basse*). Enfin, la dernière étape est la visualisation des résultats. Ici, nous affichons les interprétations des trajectoires où les points d'intérêt et les marées sont mis en parallèle. Ainsi les géographes peuvent voir efficacement les points d'intérêt croisés par les touristes durant les marées hautes et basses (cf. modules *Visualisation des interprétations*).

4.3 Description des besoins des géographes et aménageurs

Les géographes du projet DA3T veulent s'appuyer sur les trajectoires sémantiques afin de mieux comprendre le comportement des touristes. Les aménageurs ont pour objectif d'utiliser les résultats de leurs analyses pour aménager et valoriser le territoire touristique. Pour répondre à ces objectifs et faciliter l'analyse des traces de mobilité touristiques, nous avons fait le choix de concevoir une plateforme de type ETL permettant de traiter ces données. Nous avons identifié plusieurs besoins quant au développement d'une telle plateforme.

Tout d'abord, comme dans tout logiciel ETL, les outils de traitement des données doivent être proposés sous la forme de modules. Un module est une unité logicielle qui permet d'effectuer un traitement très spécifique sur des données qu'il reçoit en entrée et qui donne le résultat en sortie. Ces modules doivent pouvoir être paramétrés par l'utilisateur selon ses besoins. Une demande spécifique des géographes est que ces modules ne deviennent pas des boîtes noires dans lesquelles il est difficile de comprendre ce qui est réellement fait sur les données. Pour éviter cela, nous avons mis en place un cahier des charges consultable en annexe C détaillant chaque module à l'aide d'informations telles que les types de données en entrée et en sortie, les paramètres s'il y en a et la description du traitement effectué par le module. Ce document a été validé par les géographes et sert de documentation détaillée des modules.

Pour traiter des jeux de données de mobilité, la plateforme doit permettre aux utilisateurs de créer des chaînes de traitement en enchaînant des modules. Les modules sont organisés dans différentes catégories (p. ex. pré-traitement, enrichissement, visualisation, etc.). Au fil d'une chaîne de traitement, les traces de mobilité évoluent en trajectoires brutes (durant la phase de pré-traitement) puis en trajectoires sémantiques (durant la phase d'enrichissement). Parfois, pour répondre à un questionnement spécifique sur un jeu de données, un utilisateur veut pouvoir faire entrer la sortie d'un module dans plusieurs autres modules pour qu'elle soit traitée de différentes manières. L'utilisateur doit également pouvoir combiner les sorties de différents modules en utilisant un module spécifique.

Les données traitées dans la plateforme sont des données de mobilité et sont représentées à l'aide d'un modèle de trajectoire sémantique présenté dans la partie 3.1. Ce dernier doit servir de modèle de transition dans la plateforme, c'est-à-dire que seuls des fichiers utilisant la structuration JSON du modèle doivent circuler entre les modules d'une chaîne de traitement.

La plateforme, ainsi que les modules, doivent être simples à maintenir pour faire des mises

à jour suite aux retours utilisateurs. L'ajout, la suppression et la modification de modules, qui sont des tâches courantes dans la maintenance de la plateforme, doivent être rapides à exécuter.

Enfin, cette plateforme est dédiée à des utilisateurs non informaticiens. Elle doit être facile à prendre en main et à utiliser. La documentation doit être claire.

4.4 Architecture globale

La plateforme modulaire que nous avons développée a une architecture de type client/serveur. Ce choix de conception a été fait afin de centraliser les codes des modules de traitement pour faciliter leur maintenance.

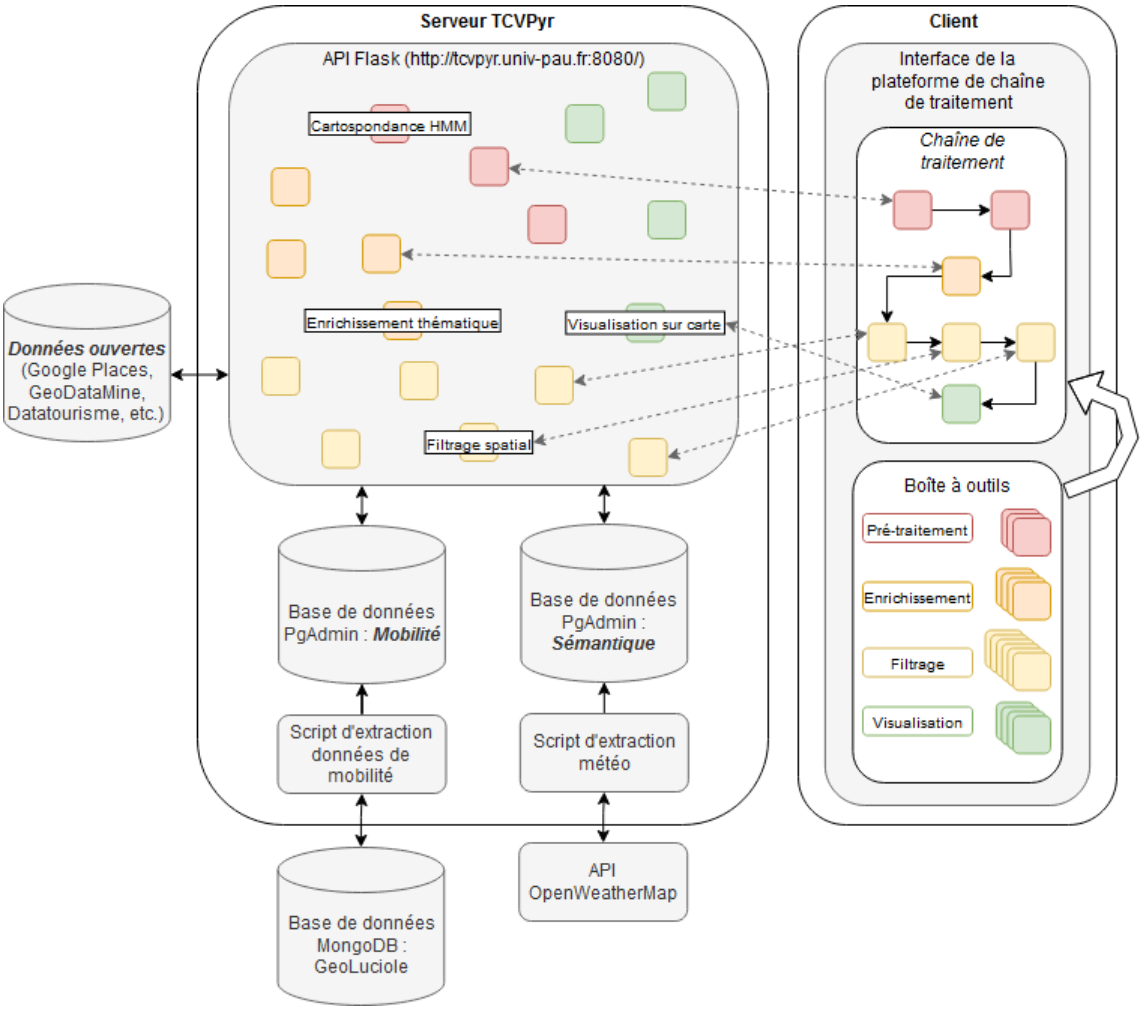


Figure 4.2 – Architecture de la plateforme de construction de chaînes de traitement DA3T

La figure 4.2 illustre l'architecture de la plateforme. Côté client se trouve l'interface sur laquelle l'utilisateur peut construire des chaînes de traitement à partir de modules de traitement. À l'exécution, une chaîne envoie des requêtes côté serveur sur lequel se trouve le code des modules. Les modules appartiennent à différentes catégories dépendant de leur fonction-

nalité. Certains modules accèdent à des données sur des bases de données. Il y en a deux : (1) une base de données *Mobilité* sur laquelle se trouve toutes les traces de mobilité brutes Géoluciole et les données sur les touristes et (2) une base de données *Sémantique* sur laquelle se trouve certaines données d'enrichissement. L'historique de la météo de La Rochelle se trouve sur cette base de données. La météo est collectée une fois par jour grâce à un script d'extraction qui fait des requêtes à l'API de OpenWeatherMap [OpenWeather]. Enfin, certains modules accèdent à des données ouvertes, notamment les modules d'extraction de données d'enrichissement sur les points d'intérêt qui font des requêtes à des API (p. ex. Google Places [Google], GéoDataMine [OpenDataFrance], DATAtourisme [DATAtourisme, a], etc.).

4.4.1 Serveur : banque de modules

Suite à de nombreuses discussions avec les experts en géographie, nous avons établi un cahier des charges recensant tous les modules à développer. Ce document est consultable en annexe C. Tous les modules sont développés en Python et localisés sur le serveur et s'exécute lorsqu'une requête leur est faite de la part d'un client. Concernant les modules en eux-mêmes, nous distinguons ceux qui ont été repris de la littérature et ceux qui ont été créés.

Identifiant du module	Nom du module	Intégré à la plateforme	Repris de la littérature	Référence (si repris)
111	Construction de trajectoires temporelles	✓	✗	/
112	Construction de trajectoires spatiales	✓	✗	/
121	Nettoyage basé sur la précision	✓	✗	/
122	Cartospondance naïve	✗	✗	/
123	Cartospondance avec les chaînes de Markov cachées	✓	✓	Newson and Krumm [2016]
131	Extraction Géoluciole	✓	✗	/
201	Filtrage spatial	✓	✗	/
202	Filtrage temporel	✓	✗	/
203	Filtrage thématique	✓	✗	/
204	Filtrage sur les données brutes	✓	✗	/
301	Annotation par aspects	✓	✗	/
302	Segmentation en épisodes	✓	✗	/
401	Modification d'une position	✗	✗	/
402	Suppression d'une position	✗	✗	/
403	Ajout d'une position	✗	✗	/
501	Visualisation cartographique	✓	✗	/
502	Visualisation sous la forme d'un cube spatio-temporel	✓	✓	Menin et al. [2019]
503	Visualisation des interprétations	✓	✗	/

Table 4.1 – Synthèse des modules spécifiés dans le cahier des charges

Le tableau 4.1 montre les différents modules spécifiés dans le cahier des charges. Les deux premières colonnes identifient le modules, la troisième colonne indique si le module à été intégré à la plateforme à ce jour et les quatrième et cinquième colonnes indique si le module à été repris de la littérature. Ainsi, deux modules ont été repris de la littérature, à savoir, le module de cartospondance avec les chaînes de Markov cachées et le module de visualisation sous la forme d'un cube spatio-temporel.

4.4.2 Client : interface utilisateur

L'interface utilisateur est un logiciel bureau permettant à l'utilisateur de créer des chaînes de traitement sur des jeux de traces de mobilité à partir de modules de traitement via un espace de travail.

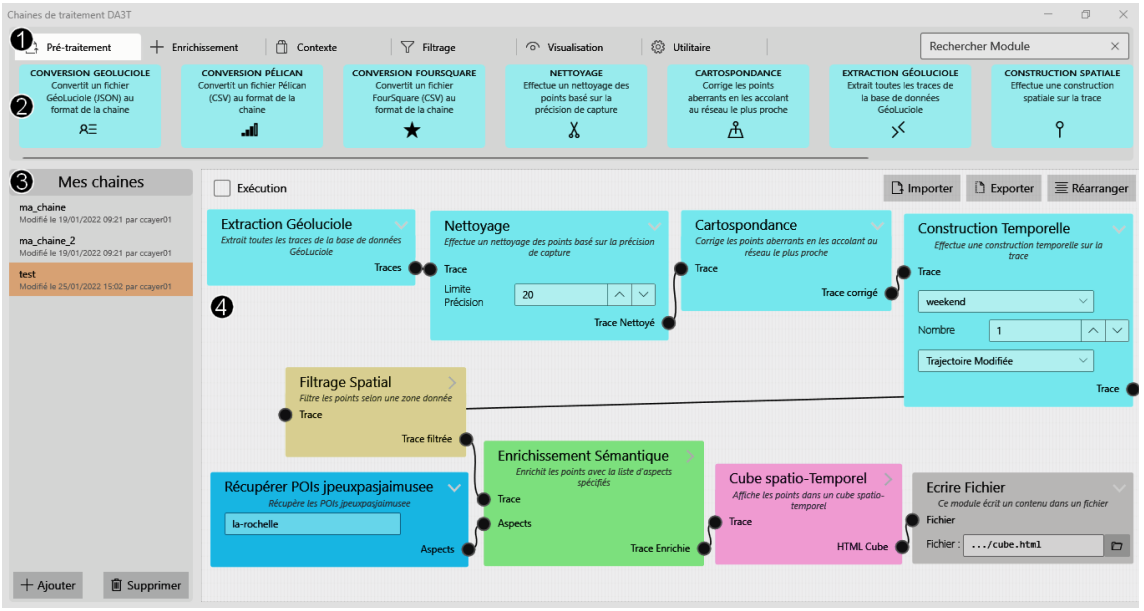


Figure 4.3 – Interface utilisateur de la plateforme

La figure 4.3 montre une capture d'écran de l'interface utilisateur de la plateforme DA3T. En (1), les modules de la plateforme sont rangés dans différentes catégories. Les catégories sont les suivantes : *Pré-traitement*, *Enrichissement*, *Contexte*, *Filtrage*, *Visualisation* et *Utilitaire*. En (2), les modules sont décrits par catégorie et sont définis par un titre et une description. Ils peuvent être glissés-déposés dans l'espace de travail en (4) où ils sont intégrés à une chaîne de traitement pour répondre à un questionnement particulier sur des traces de mobilité. Les chaînes de traitement peuvent être enregistrées, importées et exportées et sont visibles dans la liste en (3).

4.5 Application du scénario de motivations

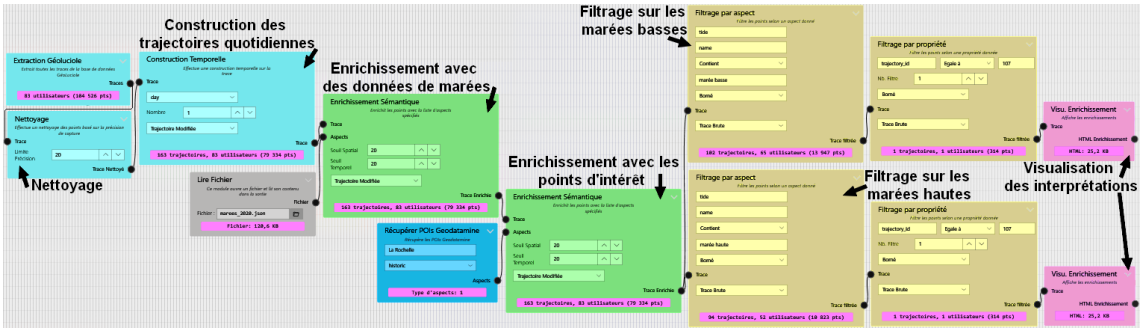


Figure 4.4 – Chaîne de traitement permettant de répondre à la question (1) implémentée dans la plateforme DA3T

La figure 4.4 montre la chaîne de traitement 4.1 permettant de répondre à la question (1) implémentée dans la plateforme DA3T. Nous l'avons exécutée sur le jeu de données Géoluciole pour découvrir quels points d'intérêt intéressent les touristes quand la marée est haute

et quand la marée est basse. Les modules de filtrage permettent d'obtenir deux groupes de trajectoires, les premières se déroulent à marée basse et les deuxièmes se déroulent à marée haut. Après les modules de filtrage sur les marées, nous avons ajouté des modules pour sélectionner une trajectoire sémantique par groupes obtenus, une pour illustrer les segments de trajectoires se passant pendant la marée basse et l'autre pour illustrer les segments de trajectoires se passant pendant la marée haute.

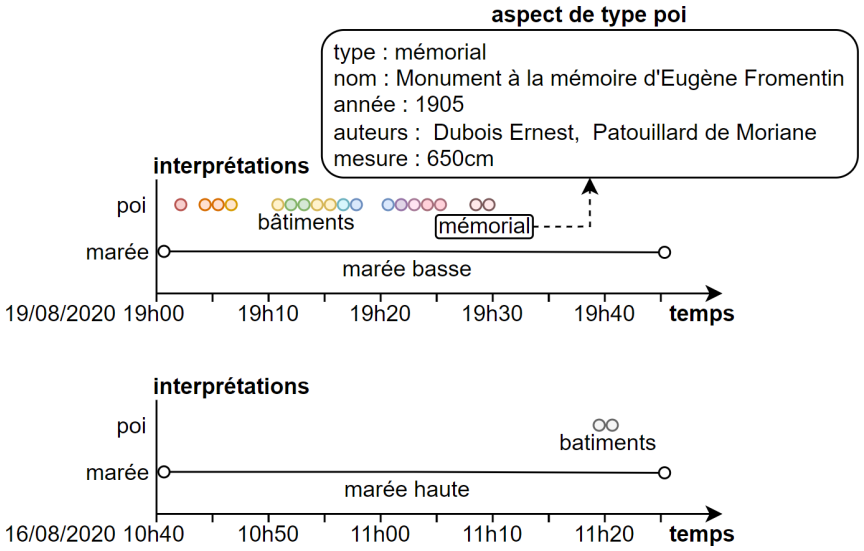


Figure 4.5 – Dimensions thématique de deux trajectoires après le filtrage sur les marées (marée basse en haut, marée haute en bas)

La figure 4.5 montre les résultats visuels obtenus à la fin de la chaîne de traitement pour chacune des deux trajectoires sélectionnées (reconstruits à partir des copies d'écran des résultats générés, consultables sur les figures E.48 E.52 en annexe). Nous pouvons voir les interprétations des deux trajectoires et nous pouvons répondre à la question (1) en nous appuyant sur l'interprétation POI. Ainsi, ces deux touristes sont majoritairement passé à côté de bâtiments et de monuments que ce soit à marée basse ou à marée haute. Rappelons que ces deux touristes ont été sélectionnés au hasard dans les deux groupes et leurs comportements sont pas représentatifs des comportements des autres membres des groupes.

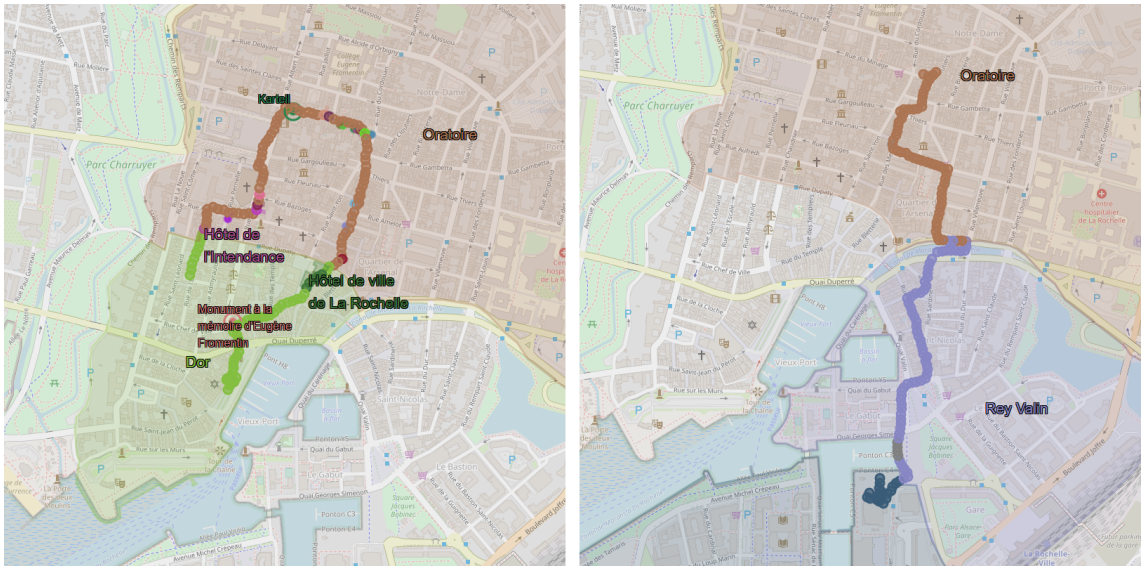


Figure 4.6 – Carte représentant les deux trajectoires pendant les marées (marée basse à gauche, marée haute à droite)

Enfin, la figure 4.6 montre une visualisation cartographique pour chacune des deux trajectoires sélectionnées.

4.6 Synthèse générale

L'objectif principal de cette thèse est de développer un outil de traitement des traces de mobilité afin d'aider les géographes du projet dans leurs analyses de ces données. Nous avons décidé de concevoir une plateforme ETL dédiée aux traces de mobilité et destinée à être utilisée par des non-informaticiens. Cette plateforme a une architecture de type client/serveur. Le client est l'interface que chaque utilisateur de la plateforme doit avoir sur son bureau et le serveur est une banque de modules situé sur une machine distante. À l'exécution d'une chaîne de traitement, le client envoie des requêtes au serveur pour chaque module composant la chaîne et le serveur répond. Beaucoup de modules sont disponibles sur la plateforme et le processus d'ajout de nouveaux modules a été penser pour être le plus simple possible. La plateforme est ergonomique et facile d'utilisation pour des géographes. Certains d'entre eux l'ont déjà testé et leurs retours sont positifs. Nous souhaitons publier et laisser libre d'accès le code de la plateforme au terme de cette thèse.

Le prochain chapitre détaille les expérimentations faites sur nos deux contributions

Chapitre 5

Expérimentations

Ce chapitre a pour but d'expérimenter nos contributions et propositions. Nous commençons par présenter les jeux de données que nous utilisons dans les expérimentations (cf. partie 5.1). Puis, nous testons le fonctionnement de la plateforme DA3T et la puissance de représentation de notre modèle de trajectoire sémantique (cf. partie 5.2). Enfin nous évaluons nos deux mesures de similarité (cf. partie 5.3).

5.1 Jeux de données

Cette partie a pour objectif de présenter les différents jeux de données que nous utilisons dans ce mémoire, notamment pour expérimenter nos contributions et propositions. Nous allons, tout d'abord, présenter le jeu de données Géoluciole contenant des traces de mobilité de touristes (cf. partie 5.1.1). Puis, nous présentons le jeu de données Foursquare contenant des traces de mobilité d'utilisateurs de réseau sociaux (cf. partie 5.1.2). Enfin, nous présentons le jeu de données Pélicans contenant des traces de mobilité d'oiseaux migrateurs (cf. partie 5.1.3).

5.1.1 Jeu de données Géoluciole

Le jeu de données principal dans nos tests et expérimentations est le jeu de données Géoluciole. Nous avons introduit succinctement le processus de collecte lors de l'introduction, partie 1.1.1, nous allons ci-après donner plus de détails sur le processus et sur le jeu de données.

L'application Géoluciole[1] est une application mobile, disponible sur iOS et Android, que nous avons développée dans le cadre du projet DA3T. Elle est destinée aux touristes visitant La Rochelle, volontaires pour nous partager leur trace de mobilité. Une fois inscrit, l'utilisateur répond à un bref questionnaire sur son contexte de visite (p. ex. Voyage-t-il accompagné ? Est-ce la première fois qu'il visite la ville ? etc.). Il a ensuite le choix d'activer ou pas la capture de son déplacement. Activée, l'application capture à intervalles de temps réguliers la longitude, la latitude, l'altitude et la vitesse de déplacement du téléphone mobile ainsi que

1. Lien GooglePlay Géoluciole : https://play.google.com/store/apps/details?id=fr.univ_lr.geoluciole&hl=fr&gl=US

l'horodatage et la précision de la capture. Il s'agit d'une valeur, en mètre, correspondant au rayon autour de la position capturée dans lequel se trouve la position réelle ; plus cette valeur est élevée moins la précision de la capture est bonne et inversement. L'envoi des données capturées se fait, après autorisation de l'utilisateur, à une base de données MongoDB hébergée sur un serveur mis en place par nos soins. Nous avons veillé à ce que Géoluciole soit conforme aux réglementations françaises et européennes concernant la protection de la vie privée.

À l'été 2020, nous avons lancé une campagne de collecte à La Rochelle durant laquelle nous avons collecté 92 traces de mobilité de touristes volontaires avec un total de 118 951 captures et une moyenne d'environ 1 293 captures par trace avec un écart-type d'environ 1 585 et une médiane de 774. Tous les touristes n'ont pas été sollicités au même moment durant leur séjour, certains venaient tout juste d'arriver à La Rochelle, d'autres repartaient le lendemain. Ainsi, dans notre jeu de données, la moyenne des durées des séjours capturés est de 2 jours par touriste.

Parmi l'ensemble des touristes volontaires ayant lancé l'application, 15 ont passé un entretien semi-directif, une fois leurs séjour terminé, afin d'enrichir et de combler d'éventuels blancs dans les traces (dus aux coupures GPS, aux zones blanches, etc.). Ainsi, les entretiens permettent de récupérer des données sur les pratiques touristiques d'un visiteur au cours de sa visite (p. ex. sortie restaurant, baignade, etc.). Ces entretiens ont été retranscrits numériquement.

5.1.2 Jeu de données Foursquare

Nous utilisons également un jeu de données contenant des trajectoires sémantiques d'utilisateurs de Foursquare, utilisé dans Varlamis et al. [2021a]. Foursquare est un réseau social qui permet à ses utilisateurs d'écrire des publications sur des points d'intérêt sur lesquels ils se situent. Les jeux de données issus de Foursquare sont souvent utilisés dans les travaux liés aux traces de mobilité touristiques May Petry et al. [2019].

Le jeu de données comprend 3 079 trajectoires hebdomadaires appartenant à 193 utilisateurs de Foursquare pour un total de 66 962 publications enrichies. La collecte a été effectuée d'avril 2012 à février 2013 à New York. Chaque publication, composant la trajectoire d'un utilisateur, comprend la longitude et la latitude de l'utilisateur au moment de la publication, l'horodatage de la publication, le nom et le type du point d'intérêt où se situe l'utilisateur, le prix (s'il y en a un) pour accéder au point d'intérêt, la note moyenne du point d'intérêt (attribuée par les utilisateurs) et la météo au moment de la publication.

5.1.3 Traces d'oiseaux migrateurs : jeu de données Pélican

Bien que nous travaillons principalement avec des trajectoires appartenant à des touristes dans le but de mieux comprendre leur comportement, nous avons utilisé un jeu de traces de mobilité retraçant les déplacements de pélicans pour montrer la généricité de nos contributions. Elles ont été recueillies par Juliet Lamb dans le cadre de son observation des activités des pélicans bruns du Golfe du Mexique [Lamb et al., 2017].

Le jeu de données comporte les traces brutes de 81 pélicans pour un total de 168 041 captures. La collecte a été effectuée entre 2013 et 2016 avec une capture par pélican toutes les 90 minutes. Chaque capture comprend la longitude et la latitude, l'horodatage, l'identifiant et la colonie de l'individu.

5.2 Évaluation du modèle de trajectoire sémantique et de la plateforme ETL

L'objectif de cette partie est de présenter une expérimentation démontrant la puissance de représentation de notre modèle de trajectoire sémantique et de vérifiant que le verrou **(V1)** a été levé. Ce verrou réside dans la construction d'un modèle de représentation de traces de mobilité *indoor* et *outdoor* prenant en compte plusieurs niveaux d'enrichissement avec des données complexes. Pour lever ce verrou, rappelons que nous émettons l'hypothèse **(H1)** que combiner certaines caractéristiques de modèles existants dans un même modèle de trajectoire sémantique permet d'enrichir les trajectoires avec plusieurs interprétations composées de données complexes, sur plusieurs niveaux de détail. Cette expérimentation permet, par ailleurs, de tester notre plateforme ETL de construction de chaînes de traitement.

Dans un premier temps, nous testons le fonctionnement de la plateforme ETL et l'instanciation du modèle à travers deux cas d'usage relatifs aux données Géoluciole collectées dans le cadre du projet DA3T. Dans un second temps, nous testons le fonctionnement de la plateforme ETL et la généricité de notre modèle avec des cas d'usage relatifs à des traces de mobilité issues d'autres domaines d'application (c.-à-d. traces de mobilité d'oiseaux et traces de mobilité d'utilisateurs d'un réseau social). Ces cas d'usage correspondent à des questionnements de haut niveau qu'ont des experts du domaine par rapport à ces différents jeux de données. Les chaînes de traitement présentées dans cette partie ont été crées et exécutées sur la plateforme ETL mais pour des raisons de lisibilité nous les avons représentées sous la forme de diagrammes.

5.2.1 Étude de deux quartiers de La Rochelle

Nous souhaitons répondre à la question **(Q1)** :

Quelles sont les trajectoires qui sont passées par le quartier *Plage Minimes* et par le quartier du *Vieux-Port* dans la même journée ?

Dans cette partie, nous présentons d'abord rapidement les spécificités du jeu de données utilisé. Puis, nous mettons en place une chaîne de traitement pour répondre à la question **(Q1)**. Nous testons, ensuite, l'évolution de l'instanciation de notre modèle au cours de cette chaîne.

Présentation des spécificités du jeu de données

Pour les deux cas d'usage suivants, nous utilisons le jeu de données Géoluciole issues de la période de collecte réalisées durant l'été 2020. Ce jeu de données regroupe les traces GPS de

mobilité de 92 touristes volontaires durant leur séjour à La Rochelle dont les retranscriptions de 15 entretiens. Nous avons décrit ce jeu de traces de mobilité plus en détail dans la partie 5.1.1.

Nous utilisons également des données d'enrichissement pour donner plus de sens à ces traces brutes. Certains de ces jeux de données sont collectés et stockés dans une base de données du projet comme la météo, d'autres sont directement interrogés l'*Open Data* à travers des API comme les points d'intérêt (cf. figure 4.2). Pour les deux cas d'usage présentés ci-après, nous utilisons le découpage administratif de la ville en quartiers et la météo. Le découpage en quartiers de La Rochelle est issue de l'*Open Data* de la ville et dénombre 23 quartiers, chacun représenté par un nom et un polygone (cf. annexe B.1). Quant aux données météorologiques, elles ont été collectées sur l'*Open Data* via une API. Ces données décrivent la météo de La Rochelle heure par heure sur 357 jours (c.-à-d. 8 463 relevés météorologiques).

Mise en place de la chaîne de traitement

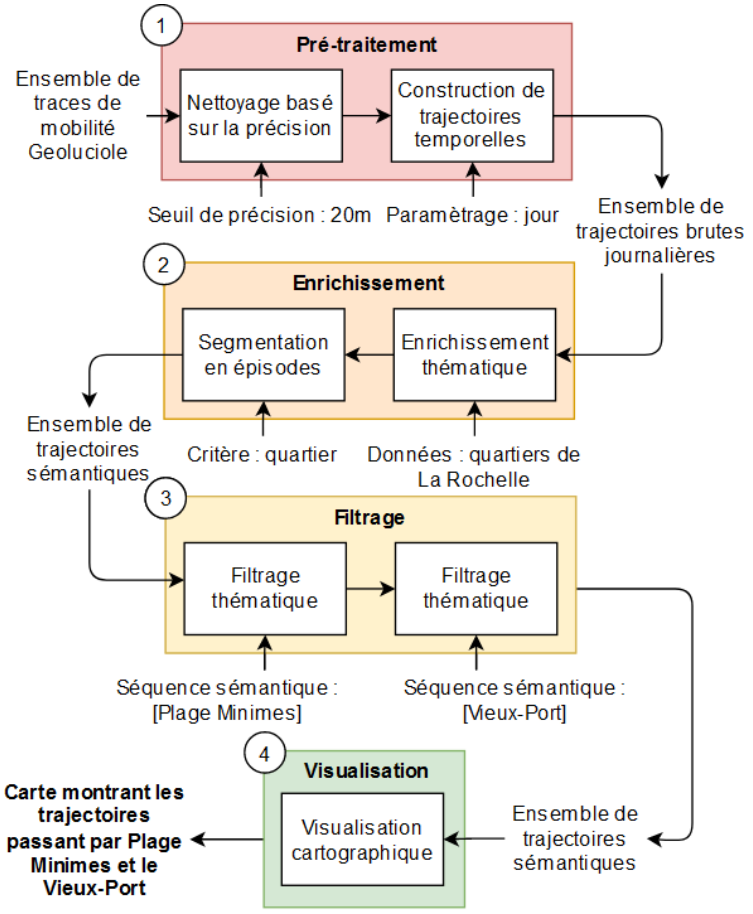


Figure 5.1 – Chaîne de traitement personnalisée permettant de répondre à la question **(Q1)**

La figure 5.1 montre la chaîne de traitement permettant de répondre à la question **(Q1)**.

Cette chaîne de traitement (cf. figure 5.1) prend en entrée des traces de mobilité brutes

et affiche en sortie une carte avec toutes les trajectoires répondant à la question **(Q1)**. Pour cela, elle enchaîne des modules issus de quatre catégories différentes.

Pour commencer, dans la catégorie *Pré-traitement* (cf. figure 5.1, catégorie 1), les traces de mobilité des touristes sont nettoyées avec le module *Nettoyage basé sur la précision*. Ce module s'appuie sur la valeur de précision enregistrée au moment de la capture de la position et indique, en mètres, le rayon d'incertitude autour de la position. Le module suivant *Construction de trajectoires temporelles* construit des trajectoires brutes à partir des traces de mobilité nettoyées. La question pousse à nous intéresser aux trajectoires journalières des touristes à l'échelle temporelle de la journée. Ainsi, le paramétrage *jour* indique que le module construit une trajectoire par jour et par personne.

La catégorie *Enrichissement* (cf. figure 5.1, catégorie 2) regroupe les modules qui transforment une trajectoire brute en une trajectoire sémantique. Le module *Enrichissement thématique* lie des aspects sémantiques aux positions de la trajectoire. Ici, chaque position des trajectoires est liée à l'aspect représentant le quartier de La Rochelle dans lequel elle se trouve en utilisant les polygones spatiaux des quartiers. Le module suivant *Segmentation en épisodes* construit, pour chaque trajectoire, une interprétation basée sur l'enrichissement réalisé précédemment, c.-à-d. une séquence d'épisodes composée des différents quartiers de La Rochelle traversés. Le résultat de ces traitements est un ensemble de trajectoires sémantiques.

La catégorie *Filtrage* (cf. figure 5.1, catégorie 3) regroupe les modules qui filtrent les trajectoires selon des critères spécifiques. Ici, le module *Filtrage thématique* filtre les trajectoires selon un quartier particulier et donne comme résultat celles dont la séquence d'épisodes comprend ce quartier. Remarquez que le module est appelé deux fois : la première fois, il sélectionne les trajectoires qui sont passées par le quartier *Plage Minimes* et la seconde fois, celles qui sont passées par le quartier du *Vieux-Port*.

Enfin, la catégorie *Visualisation* (cf. figure 5.1, catégorie 4) regroupe les modules qui permettent de visualiser des résultats. Le module *Visualisation cartographique* permet de représenter les données en entrée sur une carte. En somme, seules les trajectoires correspondant au déplacement journalier de personnes passées à la fois par les quartiers *Plage Minimes* et *Vieux-Port* sont visibles sur la carte finale (cf. question **(Q1)** initiale).

Instanciation du modèle

Nous allons maintenant nous appuyer sur la chaîne de traitement présentée figure 5.1 pour illustrer un exemple d'instanciation du modèle. Rappelons que les entrées et sorties de chaque module sont des instances du modèle : des traces, des trajectoires brutes ou sémantiques. La figure 5.2 montre l'instanciation du modèle juste avant le filtrage (cf. figure 5.1, catégorie 3).

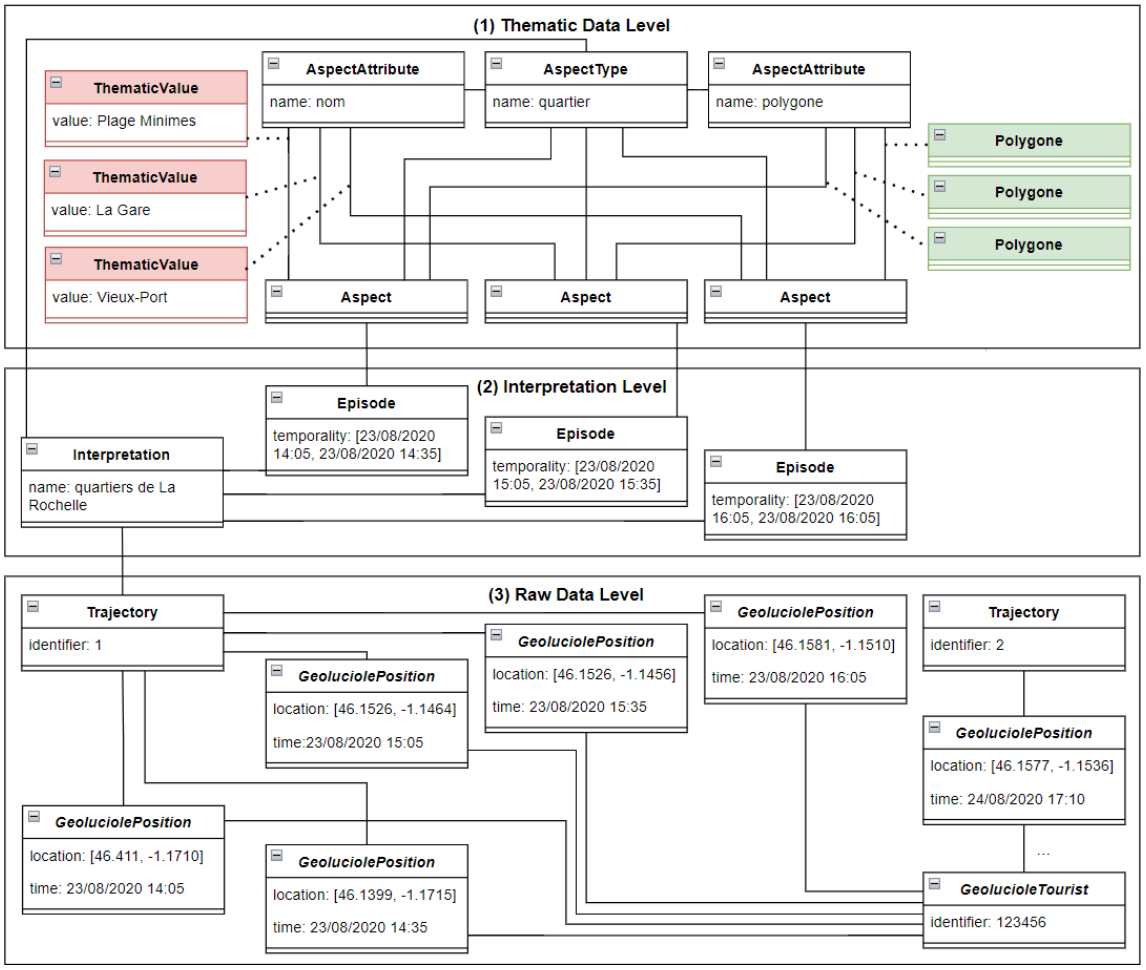


Figure 5.2 – Instanciation du modèle durant le processus de traitement pour répondre à la question **(Q1)**

La partie *Raw data level* (cf. figure 5.2, bloc 3) montre les données brutes collectées pour le touriste Géoluciole GeolucioleTourist. Dans cet exemple d'instanciation, la trace est composée de 6 positions GeoluciolePosition. Pour une question de lisibilité, nous n'avons pas fait apparaître toutes les données brutes accompagnant chaque capture Géoluciole (c.-à-d. la vitesse de déplacement, la précision de la capture, l'orientation et l'altitude) sur la figure. Deux trajectoires Trajectory résultent du module de construction temporelle pour une trace s'étendant sur deux jours. Ici, nous nous intéressons à la trajectoire d'identifiant 1 se passant le 23/08/2020.

La partie *Semantic data level* (cf. figure 5.2, bloc 1) montre trois aspects Aspect qui représentent chacun un quartier différent (c.-à-d. quartiers de la plage de Minimes, de la gare et du vieux-port). Ces aspects sont de type AspectType "quartier" et ont comme attributs AspectAttribute "nom" et "polygone". Chaque relation entre un aspect et un attribut est instanciée avec la valeur correspondante qu'elle soit spatiale, temporelle ou thématique à l'aide des classes d'association SpatialValue, TemporalValue et ThematicValue et leur sous-classes. Ces instanciations se matérialisent grâce aux liens en pointillés sur la figure. L'attribut "nom" est instancié avec une valeur thématique textuelle et l'attribut "polygone" avec une valeur spa-

tiale et plus spécifiquement un polygone.

La partie *Interpretation level* (cf. figure 5.2, bloc 2) montre ce qui résulte du module de segmentation qui a créé l'interprétation et les épisodes basés sur l'enrichissement des positions. Cette étape de la chaîne a créé les liens entre les épisodes et les quartiers ainsi que l'interprétation correspondante. Nous pouvons voir trois épisodes Episode rattachés à l'instance Interpretation ayant pour nom "quartiers de La Rochelle" car la trace a traversé trois quartiers à différents moments du déplacement. Les relations entre les positions et les quartiers résultant de l'enrichissement ne sont pas représentées pour ne pas surcharger la figure.

Après le filtrage, la trajectoire 1 apparaîtra en sortie car elle passe par les deux quartiers spécifiés dans la question **(Q1)**. Par souci de lisibilité, la figure représente la trace de mobilité d'un seul touriste Géoluciole.

5.2.2 Étude des activités touristiques en centre ville

Nous souhaitons maintenant répondre à la question **(Q2)** :

Quelles sont les activités touristiques des touristes quand il fait beau l'été dans les différents quartiers de La Rochelle ?

Cette question se base sur le même jeu de données que celui du premier cas d'usage présenté dans la partie 5.2.1 ; nous ne le représenterons pas dans cette partie. Dans un premier temps, nous mettons en place une chaîne de traitement pour répondre à la question **(Q1)**. Dans un second temps, nous testons l'évolution de l'instanciation de notre modèle durant l'exécution de cette chaîne.

Mise en place de la chaîne de traitement

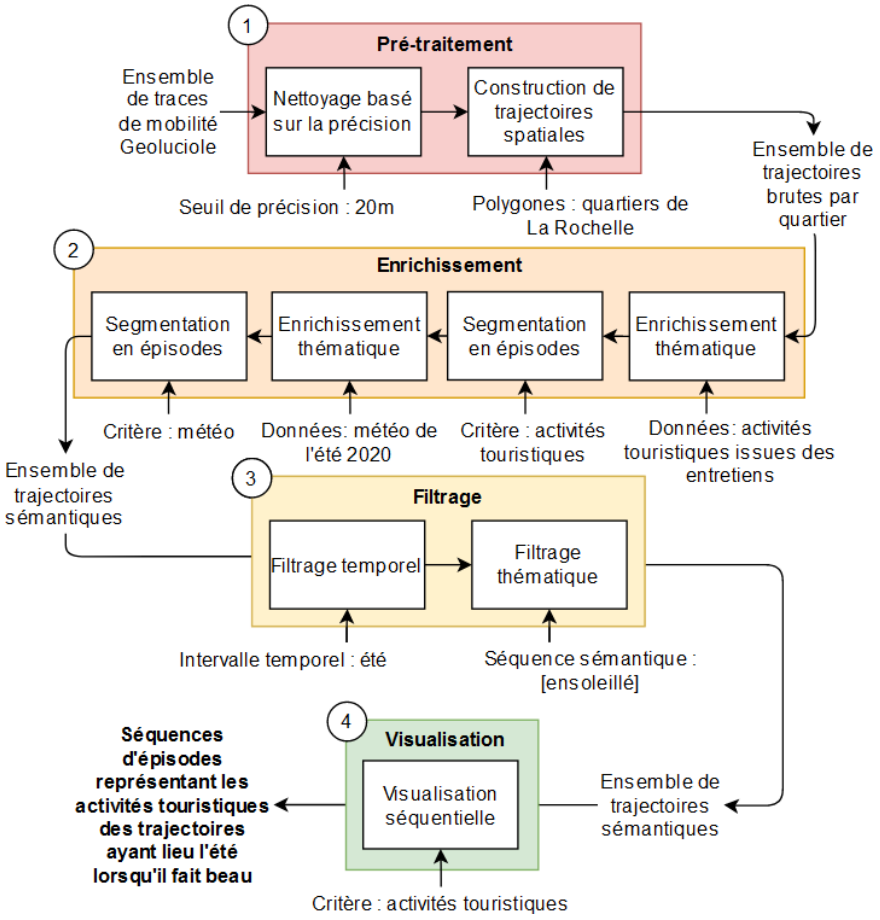


Figure 5.3 – Chaîne de traitement personnalisée permettant de répondre à la question **(Q2)**

La figure 5.3 montre la chaîne de traitement permettant de répondre à la question **(Q2)**.

Dans un premier temps, nous partons des traces de mobilité des touristes qui entrent d'abord dans une phase de pré-traitement (cf. figure 5.3, catégorie 1). Les traces de mobilité des touristes sont d'abord nettoyées avec le module *Nettoyage basé sur la précision*. Nous nous intéressons aux activités touristiques dans les différents quartiers, par conséquent, la construction des trajectoires se fonde sur des critères spatiaux et utilise les zones décrivant les quartiers de La Rochelle. Ainsi, à la suite du module *Construction de trajectoires spatiales*, nous obtenons une trajectoire pour chaque quartier traversé.

L'ensemble de trajectoires va ensuite entrer dans la phase d'enrichissement (cf. figure 5.3, catégorie 2) qui va lier à chaque position des aspects décrivant les activités touristiques et la météo qui lui sont potentiellement associés, grâce au module *Enrichissement thématique*, puis construire pour chaque trajectoire une interprétation basée sur l'enrichissement avec les activités touristiques et une autre basée sur l'enrichissement météorologique, à l'aide du module *Segmentation en épisodes*.

La phase suivante (cf. figure 5.3, catégorie 3) enchaîne un module de *Filtrage temporel* qui va filtrer uniquement les trajectoires qui se passent en été en se basant sur la date des positions puis un module de *Filtrage thématique* qui va filtrer les trajectoires qui présentent un épisode d'ensoleillement dans l'interprétation créée précédemment.

Enfin, la phase de visualisation (cf. figure 5.3, catégorie 4) permet de visualiser le résultat des traitements précédents sous la forme de séquences d'activités touristiques à l'aide du module *Visualisation séquentielle*. Nous obtenons, pour chaque trajectoire, une séquence des activités touristiques réalisées dans un certain quartier de La Rochelle pendant l'été lorsqu'il fait beau (cf. question **(Q2)** initiale).

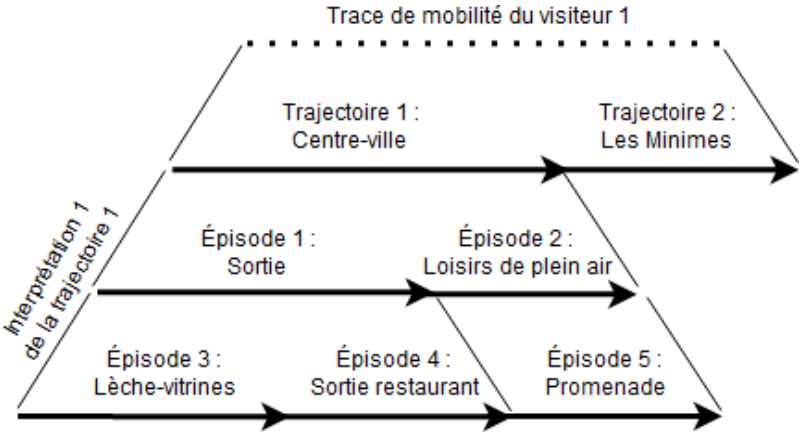


Figure 5.4 – Exemple d'interprétation d'une trajectoire relative aux activités touristiques

À la suite de la segmentation basée sur l'enrichissement avec les activités touristiques (cf. figure 5.3, catégorie 2), nous obtenons pour chaque trajectoire une interprétation basée sur ce critère. À ce stade du projet, cet enrichissement a été réalisé manuellement à partir de données collectées dans le cadre d'entretiens menés auprès des touristes. La figure 5.4 schématise les différents niveaux de données après cette première segmentation. À la suite de la phase de pré-traitement, nous avons obtenu, pour le touriste 1, deux trajectoires : l'une se déroule dans le quartier *Centre-ville*, l'autre, dans le quartier *Les Minimes*. Nous montrons seulement l'interprétation pour la trajectoire du centre-ville dans la figure 5.4. Ainsi, l'interprétation 1 est composée de cinq épisodes décrivant les activités touristiques réalisées par le touriste à ces moments-là. Il est à noter que les épisodes 3 et 4 (c.-à-d. *Lèche-vitrines* et *Sortie restaurant*) sont imbriqués dans l'épisode 1 (c.-à-d. *Sortie*) et que l'épisode 5 (c.-à-d. *Promenade*) est imbriqué dans l'épisode 2 (c.-à-d. *Loisirs de plein air*).

Instanciation du modèle

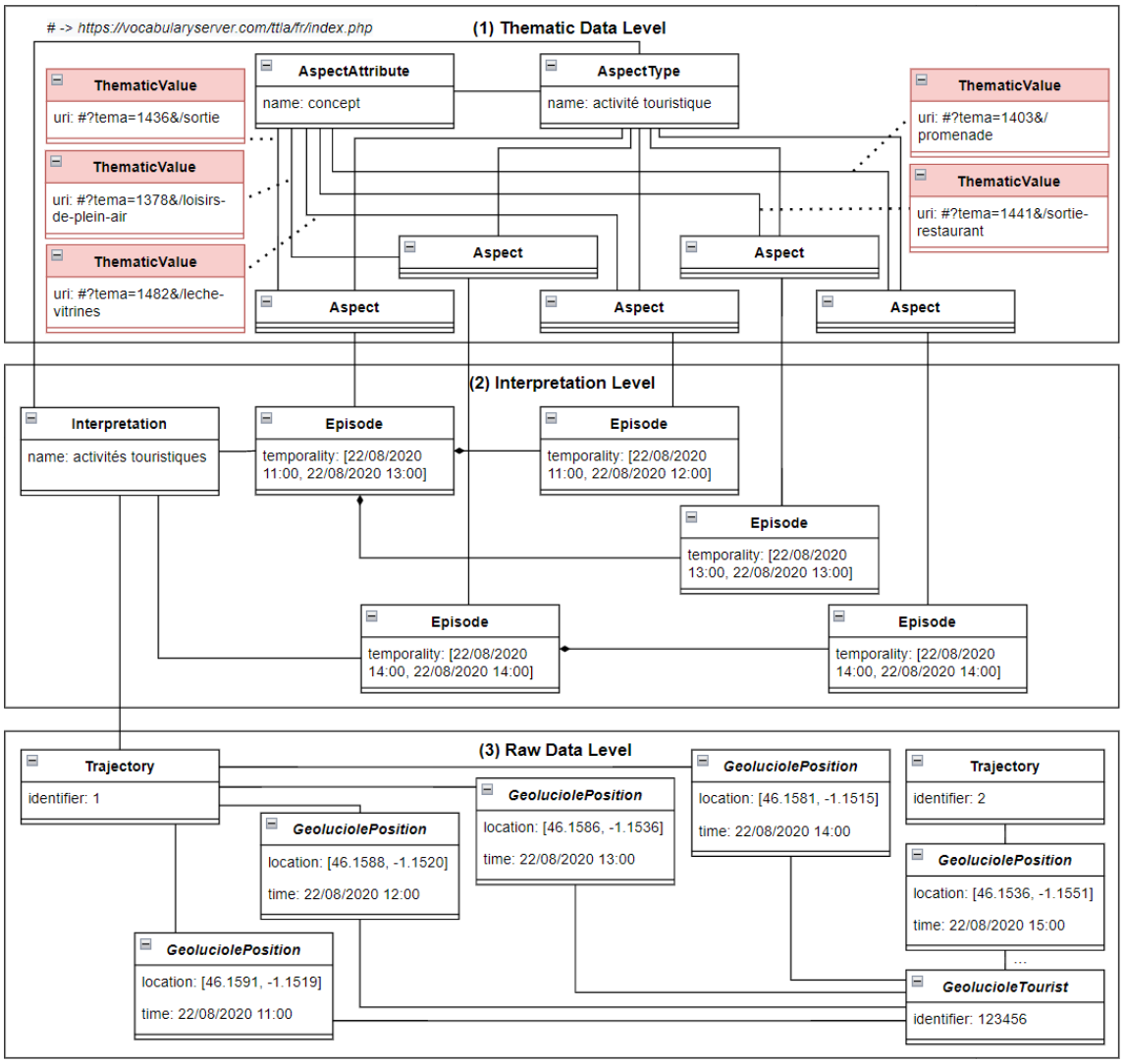


Figure 5.5 – Instanciation du modèle durant le processus de traitement pour répondre à la question **(Q2)**

La figure 5.5 montre l'instanciation du modèle juste après l'enrichissement thématique avec des données relatives aux activités touristiques.

La partie *Raw data level* (cf. figure 5.5, bloc 3) montre les données brutes collectées appartenant à un touriste Géoluciole spécifique GeolucioleTourist. Dans cet exemple d'instanciation, la trace de mobilité est composée de 6 positions GeolucioleTourist. Deux trajectoires Trajectory résultent du module de construction spatiale car la trace a parcouru deux quartiers. Dans cet exemple d'instanciation, nous nous intéressons à la trajectoire d'identifiant 1 se passant au centre-ville.

La partie *Semantic data level* (cf. figure 5.5, bloc 1) montre cinq Aspect qui représentent chacune une activité touristique différente (c.-à-d. activités de sortie, de loisirs de plein air, de

lèches-vitrines, de promenades et de sortie restaurant) issue d'un thésaurus de l'Organisation Mondiale du Tourisme décrivant les activités touristiques que nous utilisons dans le projet. Les aspects sont de type AspectType "activité touristique" et ont comme attribut AspectAttribute "concept" qui correspond au concept d'ontologie représentant une activité touristique donnée. Chaque relation entre un aspect et un attribut est instanciée avec la valeur thématique correspondante, c.-à-d. un concept du thésaurus (c.-à-d. concepts *Sortie*, *Loisirs de plein air*, *Lèche-vitrines*, *Sortie restaurant* et *Promenade*). La partie du thésaurus qui concerne les activités touristiques est représentée en figure 5.6 et les concepts que nous utilisons dans notre cas d'usage sont encadrés en rouge.

La partie *Interpretation level* (cf. figure 5.5, bloc 2) montre ce qui résulte du premier module de segmentation qui a créé l'interprétation et les épisodes basés sur l'enrichissement des positions. Cette étape de la chaîne a créé les liens entre les positions et les activités touristiques ainsi que l'interprétation correspondante. Nous pouvons voir cinq épisodes Episode liés à l'interprétation Interpretation, soient cinq activités touristiques mentionnées dans l'entretien avec le touriste. Nous nous sommes basés sur la structure du thésaurus du tourisme et des loisirs [Organisation Mondiale du Tourisme] pour hiérarchiser les épisodes. L'épisode 1 correspond à l'aspect *Sortie* et l'épisode 2 correspond à *Loisirs de plein air*. Les épisodes 3 et 4 composent l'épisode 1 et correspondent respectivement à *Lèche-vitrines* et *Sortie restaurant* et l'épisode 5 compose l'épisode 2 et correspond à *Promenade*. Comme pour le premier cas d'usage, les relations entre les positions et les aspects ne sont pas affichées mais les positions 1, 2 et 3 sont associées à l'activité de *Sortie*, les positions 1 et 2 sont aussi associées à l'activité *Lèche-vitrines*, la position 3 est aussi associée à l'activité de *Sortie restaurant*, enfin, la position 4 est associée aux activités *Loisirs de plein air* et *Promenade*. La segmentation s'appuie ensuite sur cet enrichissement pour créer une interprétation relative aux activités touristiques. Nous pouvons voir que cinq épisodes ont été créés.

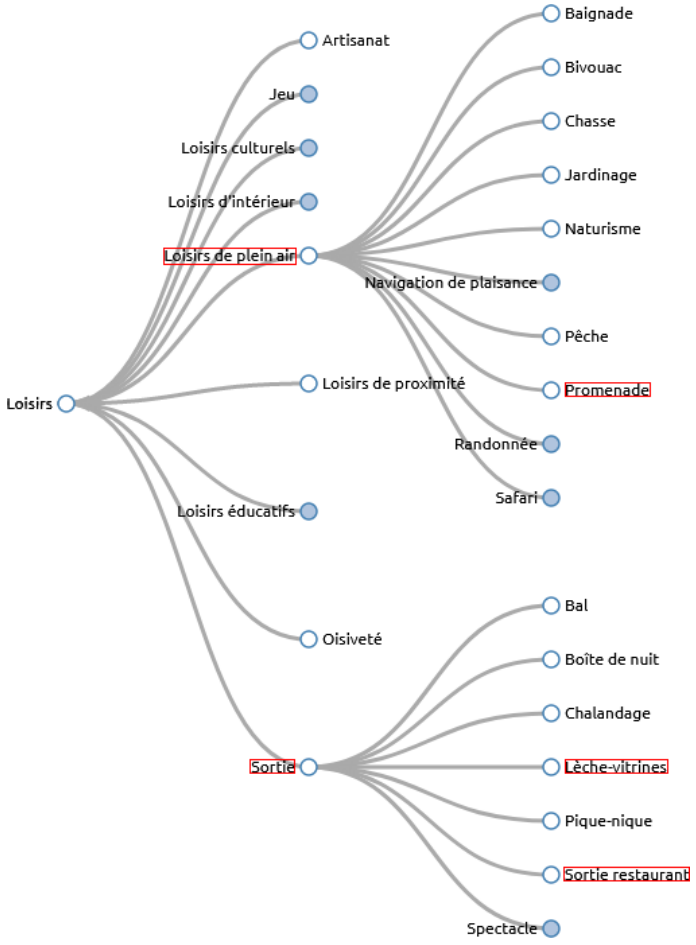


Figure 5.6 – Concept *Loisirs* du thésaurus du tourisme et des loisirs

Cette partie a permis de démontrer l'intérêt du modèle à travers deux cas d'usage réels. Le modèle supporte tous les types de données transitant entre les modules d'une chaîne de traitement. Nous avons pu voir, notamment, que le modèle supporte la description des traces de mobilité brutes, des données d'enrichissement sous la forme d'aspects et l'enrichissement des traces de mobilité par des aspects grâce à des séquences multi-niveau (c-à-d. séquences d'épisodes imbriquées).

5.2.3 Étude du jeu de données Foursquare

Afin de s'assurer de la généricité de notre modèle et des modules dans d'autres contextes d'analyse de traces, nous l'avons testé sur le jeu de données Foursquare.

Nous souhaitons répondre à la question **(Q3)** :

Quelles sont les habitudes de restauration d'un utilisateur pendant le week-end et quelles sont toutes les autres habitudes de ce même utilisateur pendant le week-end ?

Cette question se base sur le jeu de données Foursquare tel qu'il est présenté dans la partie 5.1.2. Nous mettons en place une chaîne de traitement pour répondre à la question **(Q3)** mais

nous ne montrons pas d'instanciation du modèle car le jeu de données utilisé reste à l'état brut durant son traitement et il n'y a pas d'enrichissement ou de segmentation supplémentaire.

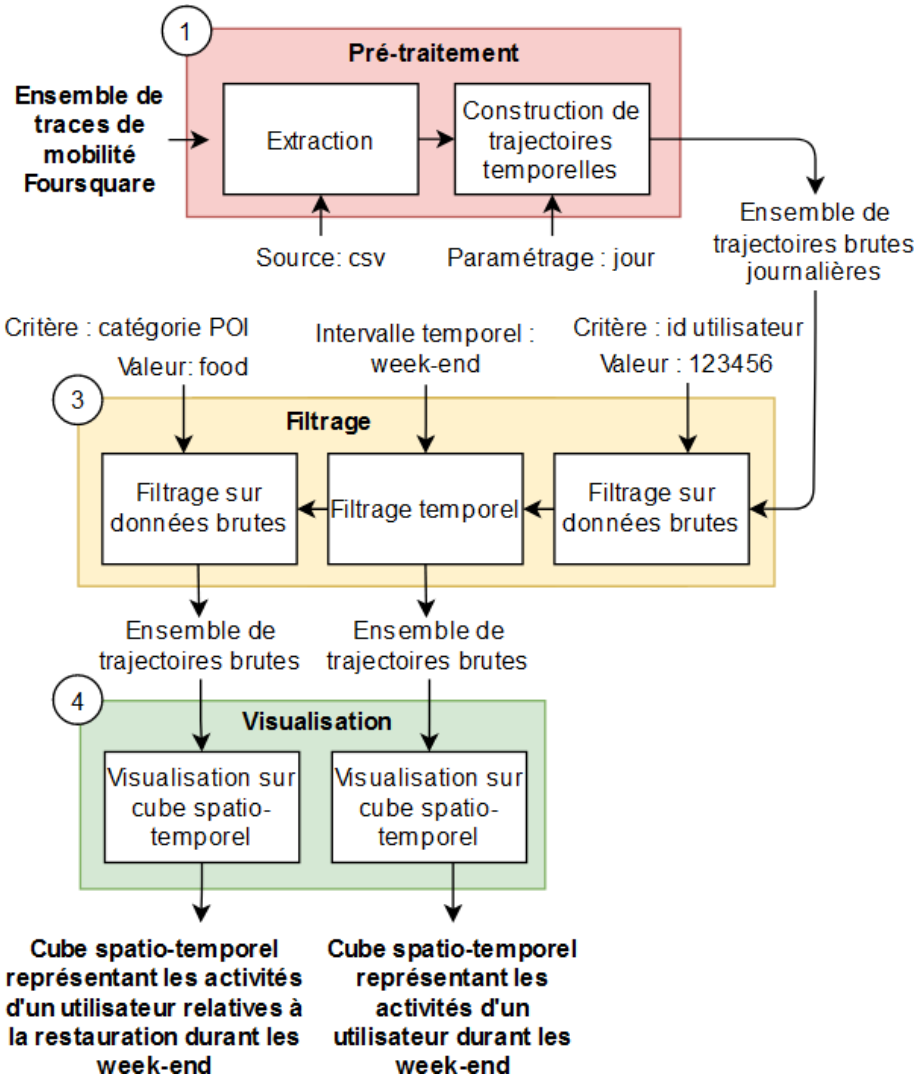


Figure 5.7 – Chaîne de traitement personnalisée permettant de répondre à la question **(Q3)**

Pour cela, nous avons mis en place la chaîne de traitement présentée sur la figure 5.7.

La phase de pré-traitement (cf. figure 5.7, catégorie 1) commence par un module *Extraction* qui permet d'importer des données sources au format CSV et de les transformer au format de notre modèle DA3T. Le module suivant *Construction de trajectoires temporelles* construit des trajectoires journalières brutes à partir des traces de mobilité grâce au paramétrage *jour*.

Il n'y a pas de phase d'enrichissement car nous nous servons uniquement des données brutes pour répondre à la question **(Q3)**.

La phase de filtrage (cf. figure 5.7, catégorie 3) enchaîne trois modules de filtrage : un module de *Filtrage sur données brutes* qui permet de ne retenir que les publications de l'utilisateur 123456, un module de *Filtrage temporel* qui permet de retenir les trajectoires qui se

déroulent le week-end et un module de *Filtrage sur données brutes* qui permet de retenir les publications relatives à des points d'intérêt de type *food*.

Enfin, la chaîne se termine par une phase de visualisation phase de filtrage (cf. figure 5.7, catégorie 4). Deux modules de *Visualisation sur cube spatio-temporel* permettent respectivement d'analyser toutes les publications de l'utilisateur 123456 sur l'ensemble des week-ends et d'analyser ces publications relatives à la restauration ces mêmes week-ends. Ces cubes spatio-temporel sont montrés sur la figure 5.8.

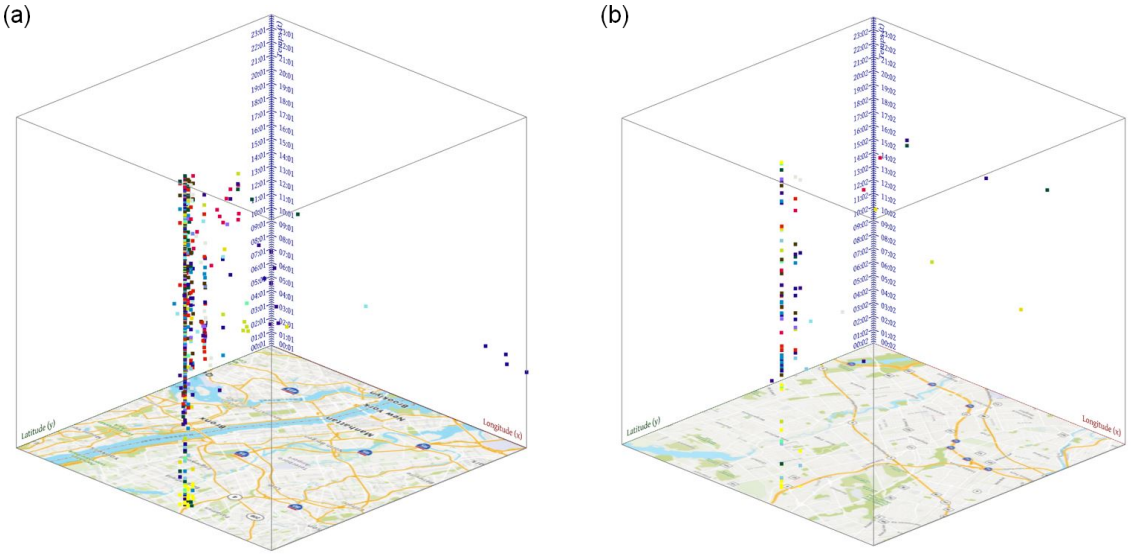


Figure 5.8 – Visualisation des publications Foursquare pour l'utilisateur 123456 le week-end : (a) toutes les activités ; (b) les activités restauration

La figure 5.8 (a) montre une vue globale sur 3 dimensions des publications du week-end d'un utilisateur de Foursquare. Nous voyons que cet utilisateur reste dans une même zone géographique. En revanche, dans la figure 5.8 (b), qui montre uniquement les publications du week-end relatives à des points d'intérêt de type *food*, nous voyons distinctement deux couloirs spatiaux d'activités récurrentes. Cet utilisateur privilégie deux quartiers de New-York pour ses activités de restauration. Pour chacun des couloirs spatiaux, nous remarquons que l'utilisateur fréquente de manière systématique le même ensemble d'établissements.

5.2.4 Étude du jeu de données Pélican

Enfin, nous avons finalement testé notre modèle sur le jeu de données Pélican qui s'éloigne encore plus de notre domaine de départ. Ici, nous ne représentons plus des déplacements humains mais des déplacements d'animaux.

Nous voulons répondre à la question **(Q4)** :

Quel est le comportement d'un ensemble de colonies de pélicans en période d'hivernage (c.-à-d. de février à mars) et en fin de saison de reproduction (c.-à-d. août) ?

Cette question se base sur le jeu de données Pélican tel qu'il est présenté dans la partie 5.1.3. Nous mettons en place une chaîne de traitement pour répondre à la question **(Q4)** mais nous ne montrons pas d'instanciation du modèle car le jeu de données utilisé reste à l'état brut durant son traitement et il n'y a pas d'enrichissement ou de segmentation supplémentaire.

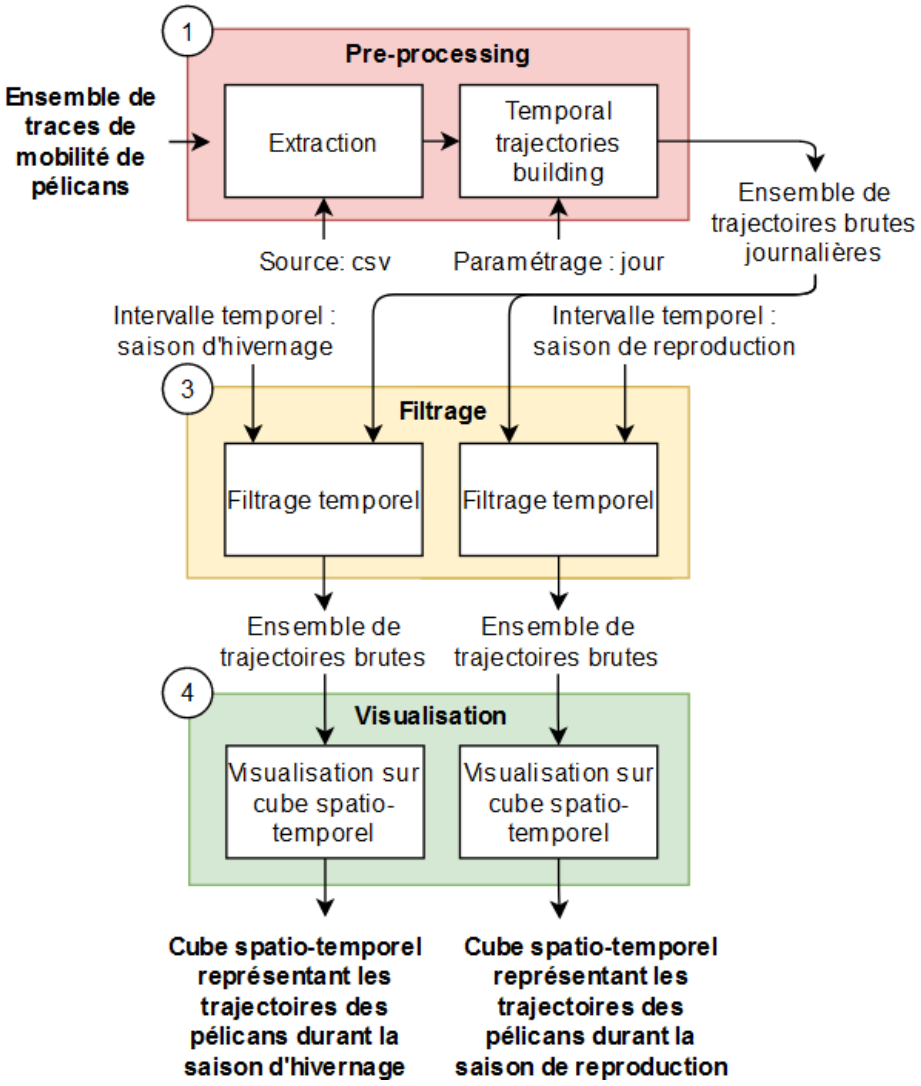


Figure 5.9 – Chaîne de traitement personnalisée permettant de répondre à la question **(Q4)**

À cette fin, nous avons mis en place la chaîne de traitement décrite figure 5.9.

La phase de pré-traitement (cf. figure 5.9, catégorie 1) commence par un module *Extraction* qui permet d'importer les traces de mobilité des pélicans depuis un fichier CSV et de les transformer au format du modèle DA3T. Le module suivant *Construction de trajectoires temporelles* construit des trajectoires journalières brutes à partir des traces de mobilité à l'aide du paramètre *jour*.

Ensuite, l'ensemble des trajectoires brutes est divisé en deux dans la phase de filtrage (cf. figure 5.9, catégorie 3) : un premier module *Filtrage temporel* permet de retenir les trajectoires correspondant à la saison d'hivernage et un second module *Filtrage temporel* permet

de retenir les trajectoires correspondant à la saison de reproduction.

Enfin, pour chacune des deux périodes, la phase de visualisation (cf. figure 5.9, catégorie 4) utilise deux modules de *Visualisation sur cube spatio-temporel*, pour les deux branches de la chaîne, afin de présenter les trajectoires des pélicans en saison d'hivernage et en saison de reproduction, respectivement. Ces cubes spatio-temporel sont montrés sur la figure 5.10.

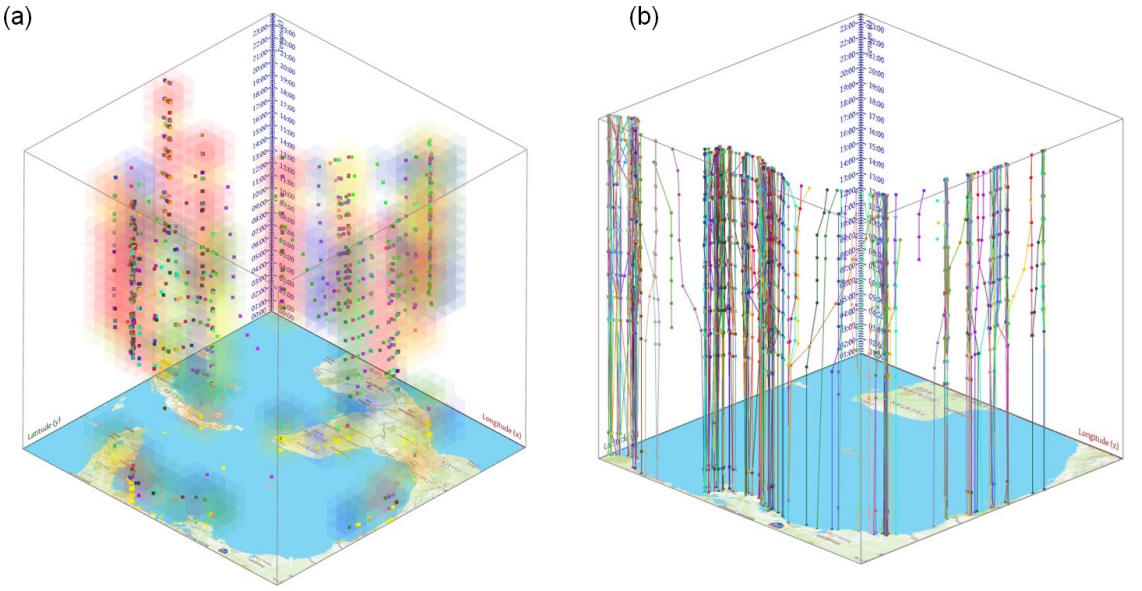


Figure 5.10 – Visualisation des données de mobilité des pélicans : (a) hivernage ; (b) reproduction

Nous avons étudié les trajectoires des 81 pélicans sur deux périodes remarquables : la saison d'hivernage (c.-à-d. entre mi-février et mi-mars 2014) et la fin de la saison de reproduction (c.-à-d. août 2014). Nous pouvons comparer les comportements des pélicans durant la même année (c.-à-d. 2014) sur ces deux périodes (cf. figure 5.10). Nous constatons qu'en période d'hivernage (cf. figure 5.10 (a)), pour chaque colonie, la zone de reproduction reste occupée à longueur de journée (cf. tubes verticaux). En survolant les points GPS d'une latitude en particulier, nous constatons que les pélicans sont en mouvement mais restent globalement dans le même périmètre. Par contre, en fin de saison de reproduction (cf. figure 5.10 (b)), nous remarquons deux types de comportements : des pélicans migrant (cf. traces non rectilignes) vers d'autres zones géographiques et des pélicans résidents (cf. traces rectilignes). Chaque trait dans le cube correspond au déplacement journalier d'un pélican.

5.2.5 Discussion

Les expérimentations menées sur des jeux de données issues de domaines d'application très différents ont permis de montrer la généricité et la robustesse du modèle et de la plateforme DA3T. Nous avons résumé ces dernières dans le tableau 5.1. Ainsi, la question **(Q1)** (cf. 5.1, **(Q1)**) illustre la dimension multi-aspect du modèle proposé à travers un exemple classique de traitement de traces de mobilité. La question **(Q2)** (cf. 5.1, **(Q2)**) vise davantage la dimension multi-niveau du modèle, en l'occurrence, les activités touristiques illustrées à tra-

vers des descriptions externes (c.-à-d. URI des ressources). Ces activités sont observées sous la forme d'épisodes avec différents niveaux d'imbrication. La question **(Q3)** (cf. 5.1, **(Q3)**), quant à elle, démontre que le modèle et la plateforme fonctionnent avec des données issues de réseaux sociaux moins précises que des données GPS. Enfin, la question **(Q4)** (cf. 5.1, **(Q4)**) montre que le modèle et la plateforme fonctionnent avec un jeu de données volumineux.

Question	**Jeu de données**	**Période(s) et lieu(x) de collecte**	**Caractéristique ciblée**
(Q1) Quelles sont les trajectoires qui sont passées par le quartier *Plage Minimes* et par le quartier du *Vieux-Port* dans la même journée ?	traces de mobilité GPS de 92 touristes avec un total de 118 951 positions spatio-temporelles	La Rochelle, été 2020	Dimension multi-aspect du modèle
(Q2) Quelles sont les activités touristiques des touristes quand il fait beau l'été dans les différents quartiers de La Rochelle ?	Idem **(Q1)**	Idem **(Q1)**	Dimension multi-niveau du modèle
(Q3) Quelles sont les habitudes de restauration d'un utilisateur pendant le week-end et quelles sont toutes les autres habitudes de ce même utilisateur pendant le week-end ?	66 962 publications géolocalisées et horodatées de 193 utilisateurs de Foursquare	New-York, entre 2012 et 2013	Généricité du modèle et de la plateforme
(Q4) Quel est le comportement d'un ensemble de colonies de pélicans en période d'hivernage (c.-à-d. de février à mars) et en fin de saison de reproduction (c.-à-d. août) ?	168 041 positions correspondants aux déplacements de 81 pélicans	Golfe du Mexique, entre 2013 et 2016	Généricité du modèle, robustesse de la plateforme

Table 5.1 – Tableau récapitulatif des expérimentations

Ainsi, nous pouvons conclure que le verrou **(V1)** est levé. Nous avons conçu un modèle permettant de représenter (1) des traces de mobilité de tout type et contexte à n'importe quelle étape de son traitement (c.-à-d. trace de mobilité, trajectoire brute et trajectoire sémantique), (2) des données d'enrichissement complexes décrites par des attributs spatiaux, temporels et thématiques et (3) un enrichissement des trajectoires brutes par les données d'enrichissement grâce à des interprétations de trajectoires (c.-à-d. séquence d'épisode enrichie) pouvant être détaillées sur plusieurs niveaux hiérarchiques.

Dans la prochaine partie, nous présentons l'évaluation de nos deux mesures de similarité entre trajectoires sémantiques.

5.3 Évaluation des mesures de similarité entre trajectoires sémantiques

L'objectif de cette partie est de présenter l'expérimentation proposant un cadre de validation de nos deux mesures de similarité entre trajectoires sémantiques et de vérifier si le verrou

(V2) est levé. Ce verrou réside dans la définition d'une mesure de similarité entre trajectoires sémantiques intégrant les dimensions spatiale, temporelle et thématique et dont les résultats s'approchent de l'avis d'experts. Rappelons que pour lever ce verrou, deux hypothèses ont été émises : l'hypothèse **(H2.1)** que la combinaison de mesures de similarité relatives à chaque dimension selon différents niveaux de granularité (c.-à-d. micro, méso, et macro) permet de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes et l'hypothèse **(H2.2)** que la combinaison de deux mesures bidimensionnelles permet de bâtir une mesure de similarité globale aux performances supérieures aux mesures de similarité existantes.

Dans un premier temps, nous présentons l'expérimentation sur la mesure s'intéressant aux trois dimensions selon plusieurs niveaux de granularité et dans un second temps, celle sur la mesure combinant des sous-mesures bidimensionnelles.

5.3.1 Description du jeu de données et présentation des experts

Le corpus utilisé pour cette expérience est composé de 30 paires de trajectoires touristiques journalières issues de la campagne de collecte Géoluciole. Ces paires représentent une variété de cas différents où les trajectoires peuvent être similaires sur toutes, plusieurs, une ou aucune des dimensions spatiale, temporelle et thématique. Nous avons enrichi ces trajectoires avec des données météorologiques, de lever et de coucher du soleil provenant d'OpenWeatherMap [OpenWeather], des données sur les points d'intérêt de La Rochelle provenant de DATAtourisme [DATAtourisme, a], des données sur les quartiers, les espaces verts, les plages et la marée provenant de La Rochelle Open Data [La Rochelle].

5.3.2 Protocole expérimental

Le protocole de l'expérience est défini par les étapes suivantes :

1. Recueillir l'avis de quatre géographes du projet sur la similarité ou la non similarité de chaque paire de trajectoires, globalement et selon chaque dimension. Ces experts en géographie du tourisme expriment leur avis en notant chaque paire de trajectoires sémantiques selon chaque dimension et globalement (c.-à-d. quatre notes par paire, au total) entre 0 et 5, avec un score de 0 pour les trajectoires qui ne se ressemblent en rien selon eux et une note de 5 pour celles qui se ressemblent parfaitement selon eux.
2. Collecter les résultats issus des deux mesures DA3T $DA3T\_S1_{glb}$ et $DA3T\_S2_{glb}$ pour chaque paire de trajectoires globalement, selon chaque dimension ainsi que selon chaque niveau de granularité par dimension en ayant fixé les seuils (c.-à-d. les valeurs associées à chaque mesure au-delà desquelles deux trajectoires sont considérées comme similaires par cette mesure) et coefficients.
3. Collecter les résultats issus des mesures DTW [Keogh and Ratanamahatana, 2005], de similarité de RI temporelle [Le Parc-Lacayrelle et al., 2007] et MUITAS [May Petry et al., 2019] correspondant respectivement aux mesures spatiale, temporelle et thématique de l'état de l'art que nous avons choisi comme référence.
4. Utiliser les métriques de précision, de rappel et de F1-score pour calculer la pertinence des mesures DA3T par rapport à l'avis des experts puis pour la comparer avec

les mesures de l'état de l'art. Ces métriques sont essentielles pour comparer les résultats obtenus des mesures par rapport aux résultats attendus par les experts et nous détaillons leurs calculs dans la suite de cette partie.

5. Réitérer les étapes (2) et (4) avec des seuils et coefficients différents afin d'optimiser ces valeurs.

Pour calculer les métriques de précision et de rappel, nous devons d'abord vérifier si les résultats de la mesure correspondent à l'avis des experts (qui nous sert de vérité). Chaque paire de trajectoires est comptée comme un vrai positif (VP), comme un vrai négatif (VN), comme un faux positif (FP) ou comme un faux négatif (FN).

Table 5.2 – Catégorisation d'une paire de trajectoires après l'analyse des experts et l'exécution de la mesure

Avis des experts → **↓ Résultats de la mesure**	**Similaires**	**Non similaires**
Similaires	VP	FP
Non similaires	FN	VN

Le tableau 5.2 montre à quoi correspondent ces catégories : une paire de trajectoires est comptée comme un VP si les experts et la mesure évaluent les trajectoires comme similaires, comme VN si les experts et la mesure évaluent les trajectoires comme non similaires, comme FN si les experts évaluent les trajectoires comme similaires et la mesure les évalue comme non similaires et enfin comme FP si les experts évaluent les trajectoires comme non similaires et la mesure les évalue comme similaires.

Ainsi, la précision qui correspond au nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par la mesure mais pas forcément par les experts, se calcule avec l'équation 5.1.

$$precision = \frac{nb\_VP}{nb\_VP + nb\_FP} \tag{5.1}$$

Le rappel, qui correspond au nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par les experts mais pas forcément par la mesure, se calcule avec l'équation 5.2.

$$rappel = \frac{nb\_VP}{nb\_VP + nb\_FN} \tag{5.2}$$

Enfin, le F1-score, qui combine la précision et le rappel, se calcule avec l'équation 5.3.

$$f1 - score = 2 * \frac{precision * rappel}{precision + rappel} \tag{5.3}$$

Un des objectifs de cette expérimentation est de fixer les valeurs des coefficients de pondération afin d'obtenir des résultats les plus proches possible de l'avis des experts. Notons que parmi les mesures réutilisées, certaines demandent des paramètres (p. ex. coefficients de pondération de TRACLUS). Nous avons fixé les valeurs de ces paramètres sur des valeurs par défaut pour ne pas ajouter plus de complexité à notre mesure qui comporte déjà neuf coefficients de pondération et d'alourdir notre expérimentation. Nous envisageons de prendre

en compte la paramétrisation des sous-mesures dans une expérimentation de validation plus complète, impliquant plus d'experts.

Présentons maintenant les résultats de l'expérimentation.

5.3.3 Résultats

Après avoir collecté l'avis des experts concernant la similarité des trajectoires sémantiques pour chaque paire selon chaque dimension et globalement, nous avons calculé leur taux d'accord. Pour cela, nous nous sommes appuyé sur le kappa de Fleiss créé par Joseph L. Fleiss [Fleiss, 1971], une mesure statistique permettant de mesurer l'accord entre plusieurs évaluateurs.

Table 5.3 – Évaluation de l'accord entre les experts avec le kappa de Fleiss

	Évaluation spatiale	**Évaluation temporelle**	**Évaluation thématique**	**Évaluation globale**
K-Fleiss	0,53	0,37	0,57	0,66

Le tableau 5.3 montre les kappa de Fleiss calculés pour les évaluations spatiale, temporelle, thématique et globale. Pour les évaluations spatiale et thématique, il atteint respectivement 0,53 et 0,57, ce qui montre une concordance moyenne entre les avis des experts. Nous constatons que leurs avis sont beaucoup plus mitigés concernant la dimension temporelle pour laquelle le kappa de Fleiss n'atteint que 0,37, ce qui montre une légère concordance entre leurs avis. Pour finir, le kappa de Fleiss atteint 0,66 concernant l'évaluation globale, ce qui implique une concordance importante.

Nous allons maintenant passer à la présentation et à la discussion des résultats de l'évaluation concernant les deux mesures.

Résultats de l'expérimentation sur la mesure $DA3T\_S1_{glb}$

Table 5.4 – Résultats de l'expérimentation sur la mesure $DA3T\_S1_{glb}$

Score	α_*	β_*	γ_*	**Seuil**	**Précision**	**Rappel**	**F1-score**
$S_{spt-mic}$				0,9	1	0,7	0,824
$S_{spt-mes}$				0,892	1	0,609	0,757
$S_{spt-mac}$				0,010	1	0,583	0,737
S_{spt}	0,33	0,33	0,34	0,616	1	0,737	0,848
$S_{tmp-mic}$				0,15	1	0,680	0,81
$S_{tmp-mes}$				0,2	0,882	0,833	0,857
$S_{tmp-mac}$				0,34	0,941	0,64	0,762
S_{tmp}	0,4	0,4	0,2	0,393	0,941	0,842	0,889
$S_{thm-mic}$				0,56	0,7	0,583	0,636
$S_{thm-mes}$				0,4	0,8	0,444	0,571
$S_{thm-mac}$				0,25	0,9	0,563	0,692
S_{thm}	0,4	0,2	0,4	0,5	0,727	0,727	0,727
$DA3T\_S1_{glb}$	0,7	0,2	0,1	0,55	0,938	0,882	0,909

Le tableau 5.4 présente les résultats de l'optimisation des coefficients et du seuil de $DA3T\_S1_{glb}$ effectuée en s'appuyant sur les avis des experts. Il présente les scores micro, méso et macro en matière de rappel, précision et F1-score pour chaque dimension. Rappelons que pour chaque dimension (c.-à-d. les lignes S_{spt}, S_{tmp} et S_{thm}), la colonne α_* correspond au coefficient appliqué au grain micro, la colonne β_* correspond au coefficient appliqué au grain méso et la colonne γ_* correspond au coefficient appliqué au grain macro. Globalement (c.-à-d. la ligne $DA3T\_S1_{glb}$), la colonne α_* correspond au coefficient appliqué à la dimension spatiale, le la colonne β_* correspond au coefficient appliqué à la dimension temporelle et la colonne γ_* correspond au coefficient appliqué à la dimension thématique. Les coefficients du tableau 5.4 sont les meilleures valeurs trouvées lors de l'application du protocole expérimental.

Concernant la dimension spatiale, nous observons de légères disparités dans les résultats pour les grains micro, méso et macro. Le grain micro (c.-à-d. la ligne $S_{spt-mic}$) donne un F1-score de 0,824, supérieur aux deux autres grains (c.-à-d. les lignes $S_{spt-mes}$ et $S_{spt-mac}$). Le score spatial global (c.-à-d. la ligne S_{spt}) prend en compte les trois niveaux de granularité de manière équivalente et nous constatons une légère amélioration des résultats par rapport à ceux des niveaux de granularité pris individuellement : F1-score de 0,848. Pour ce jeu de données, nous pouvons conclure que les experts utilisent tous les niveaux de granularité dans leur observation d'une trajectoire touristique dans la ville.

Concernant la dimension temporelle, nous observons que, parmi les trois niveaux de granularité, les grains micro et méso (c.-à-d. les lignes $S_{tmp-mic}$ et $S_{tmp-mes}$) donnent des F1-scores légèrement supérieurs à celui du grain macro (c.-à-d. la ligne $S_{tmp-mac}$). Ce grain a moins de poids dans le calcul du score temporel global (c.-à-d. la ligne S_{tmp}) et nous constatons une amélioration : F1-score de 0,889. Ainsi, là encore, il est intéressant de considérer les trois niveaux de granularité pour calculer la similarité temporelle de deux trajectoires sémantiques de ce jeu de données.

En ce qui concerne la dimension thématique, nous observons des scores plutôt faibles pour tous les niveaux de granularité. Les grains micro, méso et macro (c.-à-d. les lignes $S_{thm-mic}$, $S_{thm-mes}$ et $S_{thm-mac}$) donnent respectivement les F1-scores 0,636, 0,571 et 0,692. Nous supposons que ces faibles scores sont dû à un manque de données pour enrichir les trajectoires des touristes. Notons, en revanche, que le score thématique global (c.-à-d. la ligne S_{thm}) montre une amélioration par rapport aux niveaux de granularité séparés : F1-score de 0,727.

Enfin, les résultats obtenus avec la mesure $DA3T\_S1_{glb}$ sont supérieurs à ceux des trois mesures dimensionnelles considérées séparément, ce qui valide **(H2.1)**. Le coefficient attribué à la dimension thématique montre que les experts sont légèrement moins intéressés par cette dimension lorsqu'ils comparent deux trajectoires touristiques ; ceci est cohérent avec les résultats plus faibles obtenus pour cette dimension.

Évaluons $DA3T\_S1_{glb}$ par rapport aux mesures de référence existantes dans les différentes dimensions. Les mesures de référence choisies sont : DTW présentée dans Keogh and Ratanamahatana [2005] pour la dimension spatiale, la mesure de similarité IR temporelle présentée dans Le Parc-Lacayrelle et al. [2007] pour la dimension temporelle et enfin MUITAS présentée

dans May Petry et al. [2019] pour la dimension thématique.

Table 5.5 – Comparaison de $DA3T\_S1_{glb}$ avec les mesures de référence grâce au F1-score

Dimension	**Mesures de référence**			$DA3T\_S1_{glb}$
	DTW	**RI temp.**	**MUITAS**	
Spatiale	0,824			0,848
Temporelle		0,857		0,889
Thématique			0,636	0,727

Le tableau 5.5 montre que $DA3T\_S1_{glb}$ donne des résultats plus proches de l'opinion des experts que les mesures de référence sélectionnées grâce au F1-score, dans toutes les dimensions.

Passons maintenant à la présentation et à la discussion des résultats de l'évaluation concernant la seconde mesure.

Résultats de l'expérimentation sur la mesure $DA3T\_S2_{glb}$

Table 5.6 – Résultats de l'expérimentation sur la mesure $DA3T\_S2_{glb}$

Score	$\alpha_{spt-tmp}$	$\beta_{tmp-thm}$	**Seuil**	**Précision**	**Rappel**	**F1-score**
$S_{spt-tmp}$			0,001	0,8	0,857	0,828
$S_{tmp-thm}$			0,28	0,6	0,818	0,692
$DA3T\_S2_{glb}$	0,8	0,2	0,001	0,875	0,824	0,848

Le tableau 5.6 présente les résultats de l'optimisation des coefficients et du seuil de $DA3T\_S2_{glb}$ effectuée en s'appuyant sur les avis des experts. Il présente les scores spatio-temporel et tempo-thématique en matière de rappel, précision et F1-score. Les coefficients du tableau 5.6 sont les meilleures valeurs trouvées au cours du protocole expérimental.

Les résultats obtenus avec $DA3T\_S2_{glb}$ sont supérieurs à ceux des deux mesures bidimensionnelles considérées séparément, ce qui valide **(H2.2)**. Le coefficient attribué à la mesure spatio-temporelle (c.-à-d. la ligne $S_{spt-tmp}$) est beaucoup plus faible que celui attribué à la tempo-thématique (c.-à-d. la ligne $S_{tmp-thm}$). Cela renforce encore l'idée que les géographes utilisent moins la dimension thématique pour comparer deux trajectoires sémantiques de ce jeu de données.

5.3.4 Discussion

Les expérimentations menées sur les deux mesures donnent des résultats assez similaires. Cependant, $DA3T\_S1_{glb}$ est un peu plus proche de l'opinion des experts que $DA3T\_S2_{glb}$. Pour les deux mesures, les sous-mesures incorporant la dimension thématique donnent des résultats faibles. Cela est dû à un manque de données d'enrichissement utiles pour la comparaison. Dans $DA3T\_S2_{glb}$, nous nous sommes concentrés sur la dimension temporelle, les deux sous-mesures intègrent cette dimension. Cependant, il est rare que deux trajectoires se chevauchent temporellement et nous avons dû aligner temporellement leurs premiers horodatages pour avoir des résultats pertinents. Ainsi, pour éviter ce type de pré-traitement,

nous recommandons d'utiliser $DA3T\_S1_{glb}$ pour comparer deux trajectoires sémantiques de touristes et nous recommandons d'utiliser la $DA3T\_S2_{glb}$ pour comparer une trajectoire sémantique d'un touriste avec un itinéraire enrichi de l'office du tourisme où les temps de déplacement et de visite sont estimés.

Ainsi, nous pouvons conclure que le verrou **(V2)** est levé. Nous avons conçu deux mesures de similarité intégrant les dimensions spatiale, temporelle et thématique des trajectoires sémantiques qui donnent des résultats s'approchant plus de l'avis des experts que ceux des mesures existantes.

L'expérimentation décrite dans cette section est une première approche vers la validation de nos deux mesures. Nous l'avons réalisée avec seulement quatre experts ce qui est insuffisant. Nous espérons pouvoir réaliser une validation avec plus d'experts dans la suite de ces travaux.

Chapitre 6

Conclusion

Dans ce dernier chapitre, nous concluons ce mémoire. La première partie fait le bilan de l'intégralité des travaux menés pendant la thèse (cf. partie 6.1). La seconde partie établit des perspectives pour de futurs travaux (cf. partie 6.2).

6.1 Bilan

Dans ce mémoire, nous avons détaillé les travaux réalisés dans cette thèse, menée dans le cadre du projet régional et pluridisciplinaire DA3T. Ce projet s'appuie sur l'hypothèse que l'utilisation des traces de mobilité peut servir à analyser les comportements des touristes en ville et ainsi aider les aménageurs à gérer et valoriser le territoire touristique. Dans cette partie, nous faisons le récapitulatif de ces travaux.

Rappelons que l'objectif de cette thèse est de fournir aux géographes du projet des outils logiciels et des méthodes pour les aider à analyser les traces de mobilité touristiques. Dans ce but, une plateforme modulaire, de type ETL, permettant de traiter des traces de mobilité est conçue et développée. Elle est destinée à des utilisateurs non-informaticiens et leur permet de construire des chaînes de traitement personnalisables et paramétrables à partir de modules de bas niveau pour répondre à des questionnements de plus haut niveau sur un jeu de données. Chaque module appartient à une catégorie de traitement (p. ex. pré-traitement, enrichissement, filtrage, visualisation, etc.). Cette plateforme a une architecture client-serveur où l'interface utilisateur de construction de chaînes de traitement est localisée coté client et où l'exécution des calculs des modules de traitement s'effectue côté serveur. Elle est extensible car de nouveaux modules peuvent être ajoutés sans trop de difficulté. L'extensibilité de la plateforme est utile, entre autres, lorsqu'il s'agit d'intégrer de nouvelles sources de données et qu'il est nécessaire d'ajouter des modules d'extraction. Elle est également générique car elle accepte n'importe quels types de traces de mobilité (p. ex. traces d'humains, d'oiseaux migrateurs, etc.) et permet de les enrichir avec n'importe quels phénomènes du monde réel (p. ex. météo, points d'intérêt, activités, évènements, etc.). Les modules des chaînes de traitement peuvent être paramétrés selon les besoins de l'utilisateur.

Par exemple, nous pouvons utiliser la plateforme pour répondre à la question "*Quels sont*

les touristes qui sont passés par le quartier du Vieux-Port lors d'une journée ensoleillée ?". Nous créons une chaîne de traitement composée des modules suivants : (1) un module de construction temporelle de trajectoires pour obtenir des trajectoires journalières, puis (2) un module d'enrichissement thématique pour enrichir les trajectoires avec la météo de La Rochelle au moment du déplacement, (3) un module de filtrage spatial avec en entrée le polygone représentant le quartier du Vieux-Port pour ne garder que les trajectoires qui sont passées par ce quartier, (4) un module de filtrage thématique sur la météo pour ne garder que les trajectoires qui se sont déroulées un jour de beau temps et enfin (5) un module de visualisation cartographique des trajectoires.

Le développement de cette plateforme ETL est détaillé dans le chapitre 4 de ce mémoire. Trois publications sont au sujet de la plateforme : le premier, présenté à l'occasion de la conférence MDM 2020 [Cayèré et al., 2020] et le second, présenté durant le forum JCJC de la conférence INFORSID 2020 [Cayèré, 2020], ont été écrits en début de thèse et introduisent l'idée de créer une plateforme ETL pour répondre à l'objectif du projet et le troisième, présenté a l'occasion de la conférence SAC 2022 [Masson et al., 2022], détaille le fonctionnement de la plateforme une fois développée.

La conception d'une telle plateforme et le développement de tels modules ont soulevé le besoin d'un modèle de représentation des données. Ce modèle doit servir de modèle de transition entre les modules et permettre de représenter n'importe quels types de traces de mobilité ou de données d'enrichissement à n'importe quels moments de la chaîne de traitement (c.-à-d. des traces de mobilité aux trajectoires sémantiques en passant par les trajectoires brutes). Le verrou **(V1)** et l'hypothèse **(H1)** découlent de ce besoin.

Ainsi, nous avons développé un modèle de trajectoire sémantique **(C1)**. Ce modèle peut être séparé en trois grandes parties : (1) *Thematic Data Level*, (2) *Interpretation Level* et (3) *Raw Data Level*. La partie (1) permet de représenter les données d'enrichissement. Nous réutilisons le concept d'aspect qui permet de décrire n'importe quels phénomènes du monde réel et nous l'élargissons en précisant la dimension de chaque attribut (c.-à-d. spatiale, temporelle ou thématique). Nous qualifions notre modèle de modèle multi-aspect. La partie (3) permet de représenter les traces de mobilité brutes. Enfin, la partie (2) permet de faire le lien entre les données d'enrichissement décrites en (1) et les données brutes décrites en (3). Nous réutilisons les concepts d'interprétations et d'épisodes pour faire ce lien. Une interprétation est une séquence d'épisodes représentant un axe thématique que l'utilisateur souhaite analyser et sa construction est basée sur un ou plusieurs types d'aspect (p. ex. météo, points d'intérêt, etc.). Les épisodes peuvent être hiérarchisés en plusieurs niveaux de détail. Nous qualifions notre modèle de modèle multi-niveau. Ce modèle permet de représenter n'importe quels types de traces et n'importe quelles données d'enrichissement, ce qui en fait un modèle générique. De plus, certaines classes (p. ex. *MobileObject*, *Position*, etc.) peuvent être étendues selon les besoins de l'utilisateur pour représenter les spécificités de ses données brutes, ce qui rend le modèle extensible.

Notre modèle a été mis à l'épreuve en instanciant des trajectoires sémantiques issues de différents jeux de données de domaines variés. Nous avons instancié des traces de mobilité

de touristes Géoluciole, des traces de mobilité d'utilisateurs du réseau social Foursquare et des traces de mobilité d'oiseaux migrateurs.

Cette contribution est détaillée dans la partie 3.1 du chapitre 3 et l'expérimentation du modèle en partie 5.2 du chapitre 5. Deux publications sont au sujet du modèle : l'une a été acceptée à la conférence nationale IC 2021 [Cayèré et al., 2021a] et l'autre est parue dans la revue IJGI en 2021 également [Cayèré et al., 2021b].

Un autre besoin, cette fois-ci identifié par les géographes du projet, est celui d'une mesure de similarité permettant de comparer deux trajectoires sémantiques sur les trois dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) et dont les scores de similarité calculés s'approchent de l'avis des experts. Le verrou **(V2)** et les hypothèses **(H2.1)** et **(H2.2)** découlent de ce besoin.

Nous avons développé deux mesures de similarité pour répondre à ce besoin. L'intérêt de ces deux mesures repose sur leur décomposition en sous-mesures et sur la mise en place de coefficients de pondération pour contrôler l'influence des sous-scores sur le score final. La première mesure **(C2.1)**, nommée $DA3T\_S1_{glb}$, est décomposée en trois sous-mesures, chacune s'intéressant à une dimension (c.-à-d. spatiale, temporelle et thématique). Ces sous-mesures dimensionnelles sont, à leur tour, décomposées en trois autres sous-mesures, chacune s'intéressant à un niveau de granularité spécifique (c.-à-d. micro, méso et macro). La seconde **(C2.2)**, nommée $DA3T\_S2_{glb}$, est décomposée en deux sous-mesures, chacune s'intéressant à deux dimensions simultanément (c.-à-d. spatio-temporel et tempo-thématique). Nous avons centré cette mesure autour de la dimension temporelle car cette dernière est centrale dans notre modèle de trajectoire sémantique et dans la notion de trajectoire de manière générale.

L'objectif de nos mesures est que leurs résultats s'approchent de l'avis d'un expert sur la similarité de deux trajectoires sémantiques. L'expérimentation mise en place a deux objectifs : le premier est de fixer les coefficients de pondération pour obtenir les résultats les plus proches de l'avis des experts et le second est d'évaluer la mesure vis-à-vis d'autres mesures de la littérature et vis-à-vis de l'avis des experts.

Ces mesures sont détaillées dans la partie 3.2 du chapitre 3 et l'expérimentation du modèle en partie 5.3 du chapitre 5. Ces mesures ont été présentées dans un article de la conférence nationale INFORSID 2022 [Cayèré et al., 2022].

Pour conclure, nous avons mis en place une plateforme générique et extensible qui permet de créer et d'exécuter des chaînes de traitement allant du nettoyage jusqu'à la visualisation en passant par l'enrichissement et le filtrage d'un jeu de traces de mobilité. Deux contributions ont découlé de cet objectif et ont donné des résultats d'expérimentations satisfaisants. Notre plateforme est actuellement utilisée par les géographes du projet DA3T. Dans la partie suivante, nous abordons les perspectives et les travaux futurs de cette thèse.

6.2 Perspectives

Cette partie a pour but de présenter des perspectives aux travaux réalisés dans cette thèse.

6.2.1 Expérimentation de la plateforme par les utilisateurs finaux

Les géographes du projet DA3T sont les utilisateurs finaux de la plateforme. Ils ont été présents tout au long des phases de spécification, de conception et de développement de la plateforme DA3T. Nous avons beaucoup échangé avec eux pour identifier leurs besoins en terme de modules et d'interface utilisateur. Cependant, aucune expérimentation encadrée n'a été menée pour valider la plateforme.

Ainsi, nous souhaitons mettre en place une expérimentation de la plateforme sous la forme de tests utilisateur afin d'évaluer, d'une part, son efficacité pour aider les géographes à analyser un jeu de traces de mobilité et, d'autre part, l'ergonomie de l'interface utilisateur.

6.2.2 Optimisation automatique des coefficients

Nous avons identifié une perspective d'amélioration concernant les mesures de similarité. Lorsqu'un expert compare deux trajectoires sémantiques, il va s'appuyer sur ses connaissances pour donner plus ou moins d'importance aux différents points de comparaison. Par exemple, un expert en géographie va donner beaucoup d'importance aux dimensions spatiale et temporelle et un peu moins à la dimension thématique sur laquelle il va principalement s'intéresser aux aspects qui ressortent majoritairement dans l'enrichissement (p. ex. il a fait beau la majorité du temps). Actuellement, nos mesures permettent d'imiter cette manière de comparer deux trajectoires sémantiques grâce aux coefficients de pondération. En effet, ils permettent de donner plus ou moins de poids à chaque sous-score dans le calcul du score final. Pour optimiser ces coefficients de pondération à des trajectoires sémantiques provenant de Géoluciole, nous avons réalisé une expérimentation avec des experts en géographie du tourisme dans laquelle ils doivent évaluer manuellement la similarité spatiale, temporelle, thématique et globale de plusieurs paires de trajectoires sémantiques. Le problème réside dans le fait que si nous voulons travailler avec un autre jeu de données appartenant à un domaine différent (p. ex. l'ornithologie), il faut refaire l'expérimentation avec les experts du domaine en question et cela est coûteux en temps.

Pour optimiser les valeurs des coefficients de pondération de la mesure afin que les résultats s'approchent de l'avis d'experts, nous avons besoin de collecter ces avis. Pour cela, notre idée est de mettre en place une nouvelle fonctionnalité dans la plateforme qui s'active à chaque fois qu'un utilisateur importe un nouveau jeu de traces de mobilité et qu'il souhaite utiliser le module de similarité dans sa chaîne de traitement. Avant l'exécution de la chaîne de traitement, à travers une boite de dialogue, il lui est proposé de donner son avis sur la similarité spatiale, temporelle, thématique et globale de quelques couples de trajectoires sémantiques tirées au hasard des données entrant dans le module de similarité. Toutes ces évaluations sont stockées et participent à l'optimisation automatique des coefficients de pondération pour ce jeu de traces de mobilité. Pour trouver les valeurs optimales, une série

d'essais est réalisée sur le jeu de paires de trajectoires sémantiques avec des valeurs de coefficients différentes. À terme, les valeurs ayant donné les meilleurs résultats vis-à-vis de l'avis des experts sont conservées et le module s'exécute avec ces valeurs de coefficients.

Une fois un assez grand nombre d'avis d'experts collectés pour des jeux de données différents, nous pensons pouvoir utiliser des solutions d'apprentissage machine pour trouver les coefficients optimaux d'un nouveau jeu de données sans passer par les étapes de collecte de l'avis de l'expert et de calcul des valeurs optimales. Cela permettrait d'éviter de solliciter l'utilisateur quand il souhaite juste utiliser la plateforme et de gagner du temps dans l'exécution de la chaîne de traitement.

6.2.3 Intégration d'une banque de patrons de chaînes de traitement

Bien que nous ayons conçu l'interface utilisateur de la plateforme pour qu'elle soit facile d'utilisation et ergonomique, nous pensons qu'il existe encore des points d'amélioration. Pour répondre à un questionnement de haut niveau sur un jeu de traces de mobilité, il n'est pas toujours aisé de choisir les bons modules de bas niveau ou de les enchaîner correctement. Parfois, il peut même exister plusieurs chaînes de traitement répondant à une même question.

Nous souhaitons aider l'utilisateur dans la tâche de construction de chaînes de traitement. Nous proposons d'intégrer une banque de patrons de chaînes de traitement basée sur des questionnements récurrents. Par exemple, prenons les chaînes de traitement suivantes :

(1) *Nettoyage basé sur la précision*
-> *Construction de trajectoires temporelles* (avec paramétrage "jour")
-> *Filtrage spatial* (avec paramétrage laissé au choix de l'utilisateur)
-> *Visualisation cartographique*

(2) *Nettoyage basé sur la précision*
-> *Construction de trajectoires spatiales* (avec paramétrage laissé au choix de l'utilisateur)
-> *Enrichissement* (avec des points d'intérêt)
-> *Visualisation cartographique*

La première chaîne de traitement (1) permet de visualiser les trajectoires journalières passant par un certain endroit (p. ex. une ville, un quartier, un lieu-dit, etc.) défini par le paramétrage du module de *Filtrage spatial*. La seconde chaîne de traitement (2) permet de visualiser les points d'intérêt qui enrichissent les trajectoires construites par rapport à un ensemble de polygones spatiaux définis par le paramétrage du module *Construction de trajectoires spatiale*. Ces deux chaînes de traitement sont des exemples de patrons pouvant s'avérer utiles sur différents jeux de données dans différents domaines.

Nous souhaitons utiliser la fouille de motifs pour trouver des patrons de chaînes de modules récurrents dans l'historique des chaînes de traitement construites avec la plateforme. Ainsi, nous pourrons construire une banque de patrons de chaîne de traitement en nous basant sur ces patrons récurrents.

6.2.4 Mise en place d'un DSL formel

Enfin pour terminer cette partie, nous proposons une dernière perspective, toujours au sujet de la plateforme. Actuellement, lorsqu'un utilisateur crée une chaîne de traitement, il sélectionne les modules qu'il souhaite dans la boîte à outils de modules et il les intègre à la chaîne de traitement. Il y a peu de vérifications quand il s'agit de faire une chaîne de traitement cohérente. S'il y a une erreur, elle apparaît au moment de l'exécution du module. Cela peut laisser un sentiment d'incompréhension problématique chez l'utilisateur lorsque notre objectif est de lui faciliter l'utilisation.

Pour résoudre ce problème, nous proposons d'utiliser la notion de **DSL** (***D****omain* ***S****pecific* ***L****anguage* ou langage dédié à un domaine en français), une grammaire dédiée à un certain domaine d'application pouvant amener à un langage de programmation (ici, le traitement des trajectoires sémantiques). Actuellement, dans notre plateforme, il y a un DSL informel qui est sous-jacent à chaque chaîne de traitement qui permet leur exportation et importation. Cependant la plateforme, qui sert de interpréteur, effectue peu de vérification sur la cohérence des chaînes. Nous souhaitons mettre en place un DSL formel avec des règles (contraintes) rigoureuses d'enchaînement des modules (p. ex. un module de segmentation ne peut être placé qu'après un module d'enrichissement ou un module d'enrichissement ne peut pas être placé avant un module de construction de trajectoires). Pour établir ces règles formelles, nous nous appuyons sur les entrées, les sorties des modules et les traitements effectués par les modules. Par exemple, un module de *Construction de trajectoires* qui construit des trajectoires brutes à partir de traces de mobilité ne peut pas aller après un module d'*Enrichissement* qui prend en entrée des trajectoires brutes, les enrichit, et renvoie en sortie des trajectoires sémantiques. Une piste à explorer est de contraindre l'exécution d'un module à l'aide de précondition et postcondition sur les données respectivement en entrée et en sortie.

Annexe A

Classification des mesures de similarité

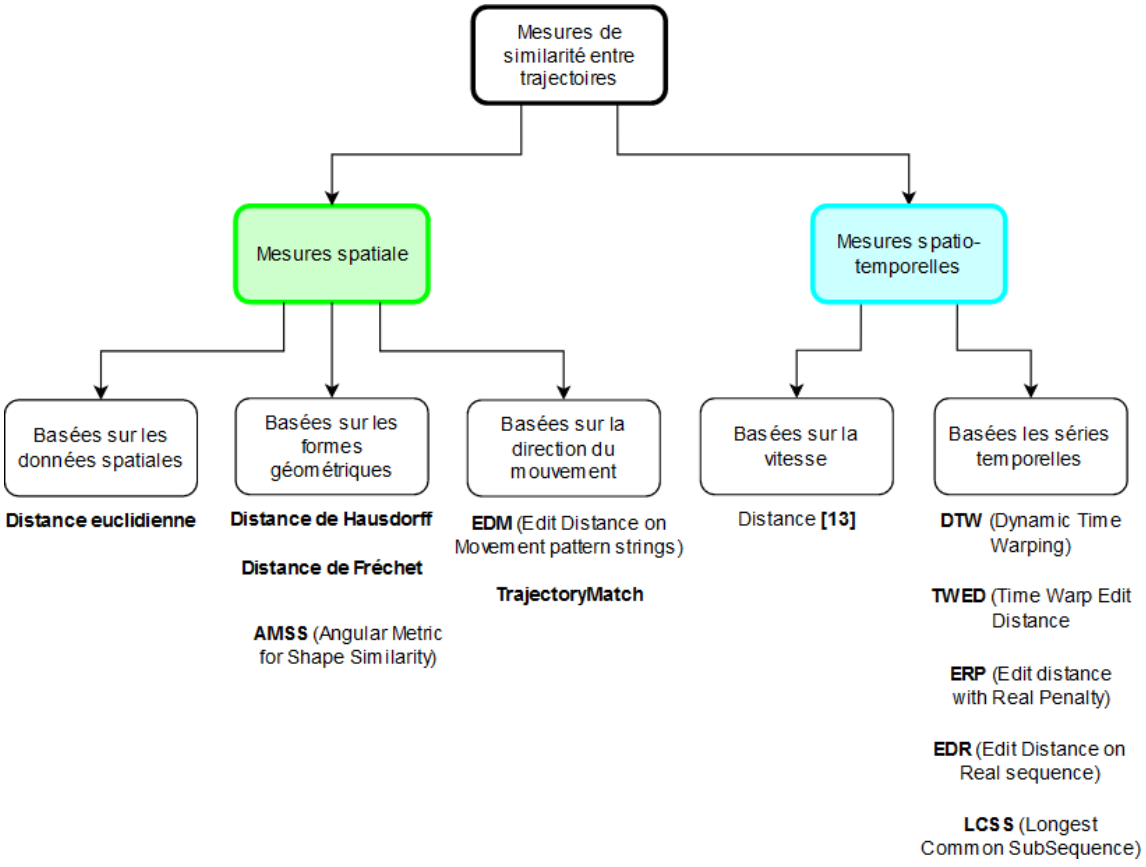


Figure A.1 – Classification des mesures de similarité de trajectoires selon Magdy et al. [2015]

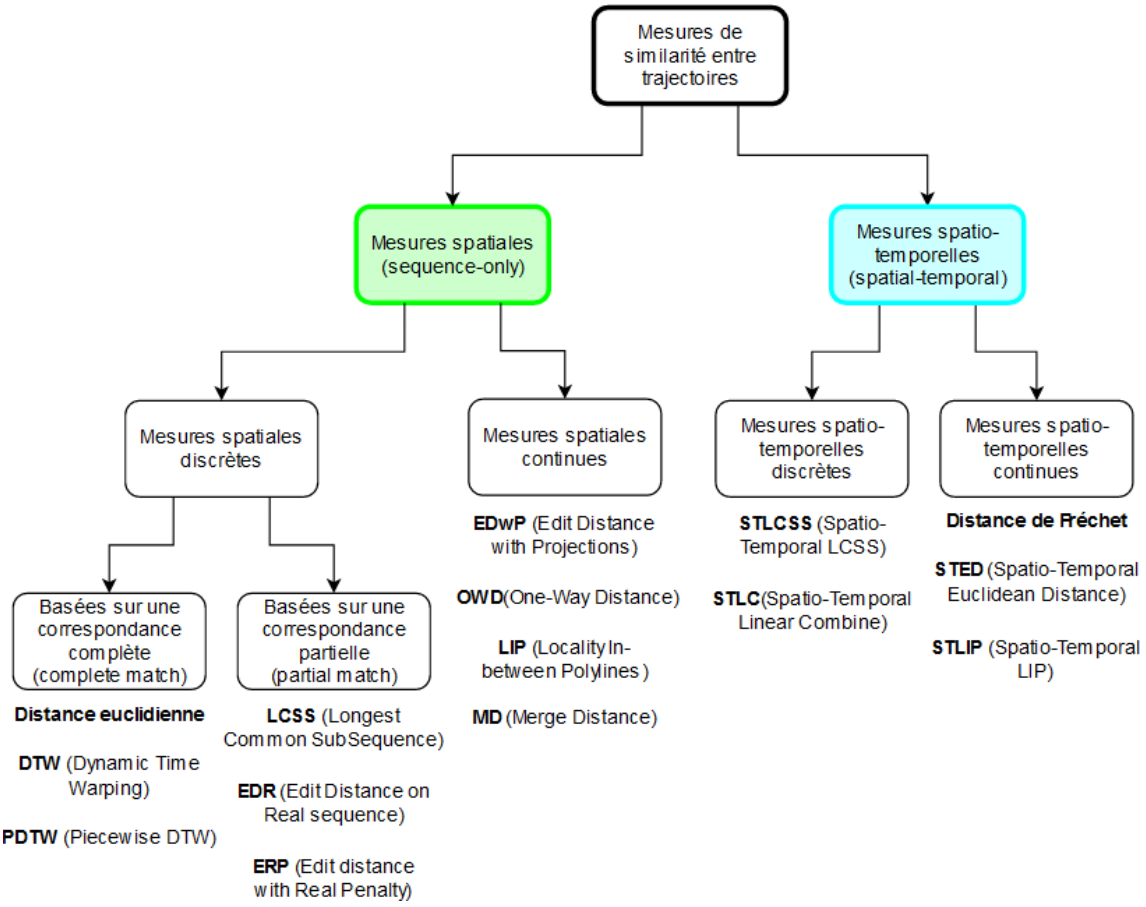


Figure A.2 – Classification des mesures de similarité de trajectoires selon Su et al. [2020]

Annexe B

Données d’enrichissement

B.1 Découpage de La Rochelle

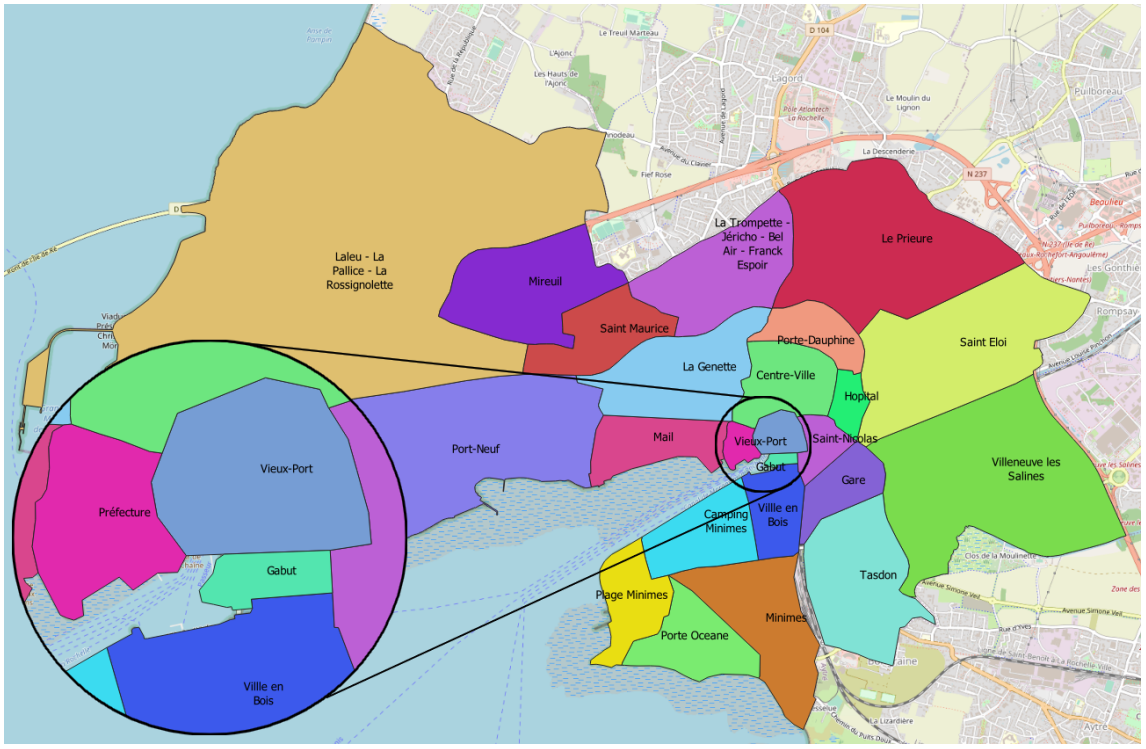


Figure B.1 – Découpage de La Rochelle en quartiers réalisé avec les géographes du projet

B.2 Données météorologiques

	datetime timestamp without time zone	description character varying (100)	temperature smallint	pressure smallint	humidity smallint	wind_speed smallint	wind_degree smallint
1	2020-07-09 22:00:00	couvert	18	1018	82	7	300
2	2020-09-03 12:00:00	partiellement nuageux	23	1023	53	4	300
3	2020-11-07 09:00:00	couvert	16	1015	72	6	120
4	2021-01-09 19:00:00	nuageux	1	1018	86	8	50
5	2021-04-10 21:00:00	ciel dégagé	12	1009	87	3	230
6	2021-05-06 14:00:00	couvert	14	1010	88	10	260
7	2021-05-16 11:00:00	couvert	15	1007	82	9	270
8	2021-05-17 08:00:00	nuageux	12	1013	76	9	270
9	2021-06-10 21:00:00	ciel dégagé	22	1020	78	4	290
10	2021-08-20 07:00:00	couvert	11	1017	100	2	90
11	2021-08-24 07:00:00	couvert	17	1021	82	4	40
12	2022-01-02 16:00:00	couvert	14	1024	88	7	230
13	2022-01-25 02:00:00	brume	-1	1032	100	4	100
14	2022-02-11 13:00:00	ciel dégagé	9	1033	57	7	80
15	2022-03-18 03:00:00	ciel dégagé	6	1033	81	5	30
16	2022-03-20 00:00:00	ciel dégagé	11	1024	76	5	100
17	2022-04-01 15:00:00	nuageux	8	1014	42	11	350
18	2022-04-01 23:00:00	ciel dégagé	4	1018	75	9	340
19	2022-05-04 02:00:00	ciel dégagé	11	1019	82	3	10
20	2022-06-10 20:00:00	ciel dégagé	20	1022	73	6	300

Figure B.2 – Extrait de la table Weather de la base de données Sémantique

Annexe C

Cahier des charges

C.1 Modules de pré-traitement (100)

C.1.1 Modules de construction de trajectoires (110)

Construction de trajectoires temporelles (111)

- description : Ce module permet de construire des trajectoires brutes à partir d'une ou de plusieurs traces de mobilité brutes. Il s'agit de découper les traces selon une périodicité donnée.
- entrées :
  - ensemble non vide de traces brutes ;
  - paramétrage :
    - *year* : les trajectoires sortantes représentent chacune le déplacement d'une année ;
    - *month* : les trajectoires sortantes représentent chacune le déplacement d'un mois ;
    - *week* : les trajectoires sortantes représentent chacune le déplacement d'une semaine.
    - *day* : les trajectoires sortantes représentent chacune le déplacement d'un jour.
    - *custom* + nombre de jours : les trajectoires sortantes représentent chacune le déplacement d'un nombre de jours donné à partir du début de la trace ;
- sorties :
  - ensemble de trajectoires brutes.

Construction de trajectoires spatiales (112)

- description : Ce module permet de construire des trajectoires brutes à partir d'une ou de plusieurs traces de mobilité brutes. Il s'agit de découper la trace selon un ensemble de polygones donné.
- entrées :
  - ensemble non vide de traces spatio-temporelles brutes ;
  - ensemble non vide de polygones spatiaux ;
- sorties :
  - ensemble de trajectoires brutes.

C.1.2 Modules de nettoyage (120)

Nettoyage basé sur la précision (121)

- description : Ce module permet de nettoyer une ou plusieurs trajectoires en supprimant les positions spatio-temporelles qui ont un diamètre de précision trop grand. Il s'agit d'un filtrage sur les positions aberrantes de la trajectoire.
- entrées :
  - ensemble non vide de trajectoires brutes ;
  - seuil limite de longueur du diamètre au delà de laquelle la position ne passe pas dans le résultat ;
- sorties :
  - ensemble de trajectoires brutes.

Cartospondance naïve (122)

- description : Ce module permet de recaler les positions spatio-temporelles d'une ou de plusieurs trajectoires brutes sur le segment de route le plus proche du réseau routier et piétonnier.
- entrées :
  - ensemble non vide de trajectoires brutes ;
- sorties :
  - ensemble de trajectoires brutes.

Cartospondance avec les chaînes de Markov cachées (123)

- description : Ce module permet de recaler les positions spatio-temporelles d'une ou de plusieurs trajectoires brutes sur le segment de route le plus probable du réseau routier et piétonnier en utilisant les chaînes de Markov cachées.
- entrées :
  - ensemble non vide de trajectoires brutes ;
- sorties :
  - ensemble de trajectoires brutes.

C.1.3 Modules d'extraction (130)

Extraction (131)
- description : Ce module permet d'extraire des traces de mobilité d'un fichier particulier.
- entrées :
  - nom d'un fichier ;
- sorties :
  - ensemble de traces spatio-temporelles brutes.

C.2 Modules de filtrage (200)

Filtrage spatial (201)
- description : Ce module permet de filtrer un ensemble de trajectoires sur des critères spatiaux. Les trajectoires sortantes sont celles qui sont en totalité ou en partie comprises dans un polygone selon un paramétrage donné.
- entrées :
  - ensemble non vide de trajectoires brutes ou sémantiques ;
  - polygone ;
  - paramétrage :
    - *complete* : les trajectoires sortantes sont celles qui sont entièrement comprises dans le polygone ;
    - *partial* : les trajectoires sortantes sont celles qui sont partiellement comprises dans le polygone ;
- sorties :
  - ensemble de trajectoires brutes ou sémantiques.

Filtrage temporel (202)
- description : Ce module permet de filtrer un ensemble de trajectoires sur des critères temporels. Les trajectoires sortantes sont celles qui sont en totalité ou en partie comprises dans un intervalle temporel selon un paramétrage donné.
- entrées :
  - ensemble non vide de trajectoires brutes ou sémantiques ;
  - intervalle temporel ;
  - paramétrage :
    - *complete* : les trajectoires sortantes sont celles qui sont entièrement comprises dans l'intervalle temporel ;
    - *partial* : les trajectoires sortantes sont celles qui sont partiellement comprises dans l'intervalle temporel ;
- sorties :
  - ensemble de trajectoires brutes ou sémantiques.

Filtrage thématique (203)

- description : Ce module permet de filtrer un ensemble de trajectoires sur des critères thématiques. Les trajectoires sortantes sont celles qui correspondent en totalité ou en partie à une séquence thématique selon un paramétrage donné.
- entrées :
  - ensemble non vide de trajectoires sémantiques ;
  - séquence sémantique ;
  - axe thématique ;
  - paramétrage de correspondance complète/partielle :
    - *complete* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend tous les éléments de la séquence sémantique en entrée ;
    - *partial* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend une partie des éléments de la séquence sémantique en entrée ;
  - paramétrage de correspondance ordonnée/désordonnée :
    - *ordered* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend des éléments de la séquence sémantique en entrée de manière ordonnée.
    - *unordered* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend des éléments de la séquence sémantique en entrée de manière désordonnée.
  - paramétrage de correspondance continue/discontinue :
    - *continuous* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend des éléments de la séquence sémantique en entrée de manière continue, c.-à-d. qu'il n'y a pas d'élément étranger à la séquence en entrée entre deux éléments de la séquence d'enrichissement ;
    - *discontinuous* : les trajectoires sortantes sont celles dont la séquence d'enrichissement pour l'axe thématique donné comprend des éléments de la séquence sémantique en entrée de manière discontinue, c.-à-d. qu'il peut y avoir des éléments étrangers à la séquence en entrée entre deux éléments de la séquence d'enrichissement ;
- sorties :
  - ensemble de trajectoires sémantiques.

Filtrage sur données brutes (204)

- description : Ce module permet de filtrer un ensemble de trajectoires sur des critères s'intéressant aux données brutes (p. ex. vitesse, orientation, altitude, etc.). Les trajectoires sortantes sont celles dont la valeur d'une donnée brute spécifiée correspond à une valeur donnée en paramètre.
- entrées :
  - ensemble non vide de trajectoires brutes ou sémantiques ;
  - nom d'une donnée brute ;
  - valeur ciblée de la donnée brute ;
- sorties :
  - ensemble de trajectoires brutes ou sémantiques.

C.3 Modules d'enrichissement (300)

Annotation par aspects (301)

- description : Ce module permet d'enrichir les trajectoires avec des données géolocalisées et/ou horodatées. Il s'agit d'annoter les positions spatio-temporelles correspondantes avec ces données additionnelles. Ces données peuvent être des données de contexte telles que la météo, les quartiers d'une ville, les points d'intérêt, etc. Cela peut également être des données liées à l'objet mobile lui-même telles que ses activités, ses moyens de transport, ses interactions, etc.
- entrées :
  - ensemble non vide de trajectoires brutes ou sémantiques ;
  - ensemble de données thématiques horodatées ou géolocalisées ;
- sorties :
  - ensemble de trajectoires sémantiques.

Segmentation en épisodes (302)

- description : Ce module permet de segmenter les trajectoires sémantiques en séquence sémantique d'épisodes selon un axe thématique donné (c.-à-d. un type de données d'enrichissement, comme p. ex. la météo).
- entrées :
  - ensemble non vide de trajectoires sémantiques ;
  - axe thématique ;
- sorties :
  - ensemble de trajectoires sémantiques sous la forme de séquences d'épisodes sémantiques.

C.3.1 Modification (400)

Modification d'une position (401)

- description : Ce module permet de modifier une position en lui attribuant une nouvelle valeur.
- entrées :
  - trajectoire brute ou sémantique ;
  - position à modifier ;
  - nouvelle position.
- sorties :
  - trajectoire brute ou sémantique modifiée.

Suppression d'une position (402)

- description : Ce module permet de supprimer une position d'une trajectoire.
- entrées :
  - trajectoire brute ou sémantique ;
  - position à supprimer.
- sorties :
  - trajectoire brute ou sémantique modifiée.

Ajout d'une position (403)

- description : Ce module permet d'ajouter une position à une trajectoire.
- entrées :
  - trajectoire brute ou sémantique ;
  - position à ajouter.
- sorties :
  - trajectoire brute ou sémantique modifiée.

C.3.2 Calcul de similarité (500)

Calcul de similarité basé sur trois dimensions et trois niveaux de granularité (501)

- description : Ce module permet d'attribuer un score de similarité à une paire de trajectoires sémantiques en se basant sur la combinaison de trois sous-mesures de similarité dimensionnelle (c.-à-d. spatiale, temporelle et thématique), elles-mêmes étant la combinaison de trois sous-mesures de similarité s'intéressant à des niveaux de granularité différents (c.-à-d. micro, méso et macro).
- entrées :
  - première trajectoire sémantique ;
  - seconde trajectoire sémantique ;
  - paramétrage :
    - coefficient α_{spt} : coefficient de pondération donnant plus ou moins d'importance à la sous-mesure spatiale ;
    - coefficient β_{tmp} : coefficient de pondération donnant plus ou moins d'importance à la sous-mesure temporelle ;
    - coefficient γ_{thm} : coefficient de pondération donnant plus ou moins d'importance à la sous-mesure thématique ;
    - coefficient $\alpha_{spt-mic}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure spatiale micro ;
    - coefficient $\beta_{spt-mes}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure spatiale méso ;
    - coefficient $\gamma_{spt-mac}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure spatiale macro ;
    - coefficient $\alpha_{tmp-mic}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure temporelle micro ;
    - coefficient $\beta_{tmp-mes}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure temporelle méso ;
    - coefficient $\gamma_{tmp-mac}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure temporelle macro ;
    - coefficient $\alpha_{thm-mic}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure thématique micro ;
    - coefficient $\beta_{thm-mes}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure thématique méso ;
    - coefficient $\gamma_{thm-mac}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure thématique macro.
- sorties :
  - score de similarité entre 0 et 1.

Calcul de similarité basé sur deux sous-mesures de similarité bidimensionnelle (502)

- description : Ce module permet d'attribuer un score de similarité à une paire de trajectoires sémantiques en se basant sur la combinaison de deux sous-mesures de similarité bidimensionnelle (c.-à-d. spatio-temporelle et tempo-thématique).
- entrées :
  - première trajectoire sémantique ;
  - seconde trajectoire sémantique ;
  - paramétrage :
    - coefficient $\alpha_{spt-tmp}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure spatio-temporelle ;
    - coefficient $\beta_{tmp-thm}$: coefficient de pondération donnant plus ou moins d'importance à la sous-mesure tempo-thématique ;
- sorties :
  - score de similarité entre 0 et 1.

C.3.3 Visualisation (600)

Visualisation cartographique (601)

- description : Ce module permet de visualiser un ensemble de trajectoires sur un fond cartographique.
- entrées :
  - ensemble non vide de trajectoires ;
  - paramétrage :
    - *default-color* + couleur : les positions des trajectoires apparaissent avec la couleur spécifiée ;
    - *default-icon* + nom de l'icône : les positions des trajectoires apparaissent avec l'icône spécifiée ;
    - *default-size* + taille : les positions des trajectoires apparaissent avec la taille spécifiée ;
    - *default-opacity* + opacité : les positions des trajectoires apparaissent avec l'opacité spécifiée ;
    - *color* : pour chaque trajectoire, les positions apparaissent d'une couleur différente ;
    - *icon* : pour chaque trajectoire, les positions apparaissent avec une icône différente ;
- sorties :
  - carte.

Visualisation sous la forme d'un cube spatio-temporel (602)

- description : Ce module permet de visualiser un ensemble de trajectoires sur un cube spatio-temporel.
- entrées :
  - ensemble non vide de trajectoires ;
- sorties :
  - cube spatio-temporel.

Visualisation des interprétations (603)

- description : Ce module permet de visualiser un ensemble de trajectoires sémantiques sous la forme de séquences d'épisodes.
- entrées :
  - ensemble non vide de trajectoires sémantiques ;
- sorties :
  - séquences d'épisodes.

Annexe D

ETL dédiés aux données de mobilité

De nombreux logiciels permettent de mettre en place des processus ETL. Dans cette annexe issue de notre étude exposée dans **??**, nous allons présenter certains d'entre eux sous deux angles : (1) sous l'angle des modules de traitement (cf. annexe D.1) puis (2) en nous concentrant sur les interactions homme-ordinateur (cf. annexe D.2). Pour chacun de ces deux angles d'approche, nous nous intéressons d'abord aux ETL orientés informatique décisionnelle (en anglais, *BI* pour ***B**usiness **I**ntelligence*) puis aux ETL spatiaux.

D.1 Modules de traitement

Nous nous intéressons d'abord aux différents types d'outils ETL existants du point de vue des modules de traitement.

D.1.1 ETL orientés informatique décisionnelle

Un ETL orienté informatique décisionnelle a pour but d'extraire et de préparer les données d'une entreprise destinées à la communication des données et l'informatique décisionnelle. L'informatique décisionnelle a de nombreuses définitions, certaines plus axées sur le côté technologique, d'autres sur le côté managérial [Foley and Guillemette, 2010] mais nous pouvons la définir comme l'ensemble des techniques et des structures permettant la communication et la valorisation des données de l'entreprise pour la prise de décision. La plupart des ETL entrent dans cette catégorie (p. ex. Talend Open Studio, Pentaho, etc.).
Un grand nombre de modules de traitement sont génériques et peuvent être trouvés dans tous les outils ETL orientés informatique décisionnelle. Les modules d'extraction permettent de lire la plupart des formats de fichiers et de bases de données, d'extraire des données via des API ou de lire des applications. Au niveau de la transformation, nous pouvons trouver des modules tels que la jointure, le *mapping* ou encore le filtrage permettant de passer de données aux formats hétérogènes à des données à un format commun. Enfin, pour cette catégorie d'ETL, le chargement consiste à charger les données homogénéisées dans un entrepôt de données ou un lac de données pour pouvoir faire de la communication des données.

Nous présentons, maintenant, des outils ETL orientés informatique décisionnelle.

Talend Open Studio[1] est un outil ETL gratuit et open source. Le logiciel fournit plus de 600 modules répartis en 22 catégories. Il est basé sur le langage de programmation Java et permet à l'utilisateur de créer ses propres modules personnalisés et de les partager avec d'autres.

RapidMiner[2] est une plateforme de science des données et d'apprentissage automatique intégrant un outil ETL. Il fournit plus de 400 modules d'analyse de données. Il permet aux utilisateurs d'en créer de nouveaux à l'aide du langage de programmation Python et de les proposer sur un espace en ligne [Ristoski et al., 2015]. Les modules sont répartis sur 8 catégories principales (p. ex. *Blending*, *Cleansing*, *Modelling*, etc.) et plus de 35 sous-catégories. Il utilise une architecture client-serveur dont le serveur peut être déployé sur place ou dans le nuage. Il est principalement destiné à l'apprentissage automatique, à l'apprentissage profond et peut même être utilisé pour l'exploration de texte [Hofmann and Klinkenberg, 2016].

Pentaho (également appelé Kettle)[3] est une suite orientée informatique décisionnelle lancée en 2006 qui fournit de l'intégration de données, des services OLAP et des services ETL [Bouman and Van Dongen, 2009]. Elle possède près de 250 modules de traitement (appelés étapes de transformation).

D.1.2 ETL spatiaux

Les outils ETL orientés informatique décisionnelle sont inadaptés au traitement des données spatiales. Les outils ETL spatiaux sont un type spécifique d'ETL qui les supporte et sont dédiés à l'extraction, la transformation et le chargement de données spatiales hétérogènes. Les algorithmes courants de géotraitement (p. ex. la validation de la géométrie ou la vérification de la topologie) sont également inclus [Drešček et al., 2020]. Certains ETL orientés informatique décisionnelle ont également des extensions leur permettant de traiter des données spatiales.

GeoKettle[4] est un outil ETL spatial qui prend en charge les données géométriques vectorielles (p. ex. points, lignes, polygones, etc.) [Astriani and Trisminingsih, 2016] et tous les processus associés (p. ex. centroïde, distance, tampon, etc.).

FME Desktop[5] (pour ***F**eature **M**anipulation **E**ngine*) est un outil ETL réalisé par SAFE Software spécialisé dans les données géographiques et les images. Il offre plus de 400 modules répartis sur 16 catégories allant du traitement matriciel à l'analyse spatiale.

Spatial Extension for Talend[6] est un plugin qui peut être installé sur Talend et permet de transformer et d'intégrer des données entre des systèmes d'information géographique. Il

1. https ://www.talend.com/products/talend-open-studio/
2. https ://rapidminer.com/
3. https ://www.hitachivantara.com/en-us/products/data-management- analytics/pentaho.html
4. https ://live.osgeo.org/archive/10.5/fr/overview/geokettle$_overview.html$
5. https ://www.safe.com/fme/fme-desktop/
6. https ://talend-spatial.github.io/

ajoute le support de la plupart des formats SIG pour l'extraction (p. ex. PostGIS, Shapefile, KML, etc.) et la transformation SIG (p. ex. tampon, centroïde, surface et longueur, distance, etc.).

D.1.3 Synthèse

Nom	**Besoins DA3T**	**Talend**	**RapidMiner**	**Pentaho**	**GeoKettle**	**FME Desktop**
Type		ETL	ETL	ETL	ETL spatial	ETL spatial
Extraction						
Sources courantes						
Bases de données	✗	✓	✓	✓	✓	✓
Formats de fichiers communs (p. ex. CSV, JSON, etc.)	✓	✓	✓	✓	✓	✓
Cloud / Services web / API	✓	✓	✓	✓	✓	✓
Applications web (p. ex. Twitter, etc.)	✗	✓	✓	✓	✓	✓
(Spatial) Format de fichier SIG (p.ex. KML, SHP, GeoJSON, etc.)	✓	+	✗	+	✓	✓
Transform						
Fonctions courantes						
Mapping	✗	✓	✓	✓	✓	✓
Nettoyage	✓	✓	✓	✓	✓	✓
Filtrage	✓	✓	✓	✓	✓	✓
Normalisation	✗	✓	✓	✓	✓	✓
Réduction de carte	✗	✓	✓	✓	✓	✓
Jointure	✗	✓	✓	✓	✓	✓
Stats	✗	✓	✓	✓	✓	✓
Suppression des doubles	✗	+	✗	+	✓	✓

Nom	**Besoins DA3T**	**Talend**	**RapidMiner**	**Pentaho**	**GeoKettle**	**FME Desktop**
(Spatial) Reprojection	✗	+	✗	+	✓	✓
(Spatial) Simplification	✗	+	✗	+	✓	✓
(Spatial) Mesure (taille, distance, etc.)	✗	+	✗	+	✓	✓
(Spatial) Géométries	✗	+	✗	+	✓	✓
Fonctions requises						
Nettoyage basé sur la précision des points	✓	✓	✓	✓	✓	✓
Cartospondance	✓	✗	✗	✗	✗	✗
Construction spatio-temporelle de trajectoires	✓	✗	✗	✗	✗	✗
Fusion de traces	✓	✗	✗	✗	✗	✗
Enrichissement sémantique	✓	✗	✗	✗	✗	✗
Détection des arrêts	✓	✗	✗	✗	✗	✗
Filtrage avancé des trajectoires	✓	✗	✗	✗	✗	✗
Calcul de similarité entre trajectoires sémantiques	✓	✗	✗	✗	✗	✗
Segmentation	✓	✓	✓	✓	✗	✓
Load						
Cibles communes						
Entrepôts de données	✗	✓	✓	✓	✓	✓
Bases de données	✗	✓	✓	✓	✓	✓
Fichiers communs	✓	✓	✓	✓	✓	✓

Nom	**Besoins DA3T**	**Talend**	**RapidMiner**	**Pentaho**	**GeoKettle**	**FME Desktop**
(Spatial) Fichiers de géographie	✓	+	✗	+	✓	✓
Cibles requises						
Visualisation sur cube spatio-temporel	✓	✗	✗	✗	✗	✗
Visualisation de l'enrichissement	✓	✗	✗	✗	✗	✗
Visualisation sur carte	✓	+	✓	+	✓	✓
Visualisation du temps	✓	✗	✗	✗	✗	✓

Table D.1 – Comparaison de plusieurs ETL par rapport à leurs modules et aux besoins relatifs à la plateforme DA3T

Le tableau D.1 représente la liste les modules nécessaires pour notre plateforme de chaîne de traitement. Un ✓ signifie que le module est présent dans le logiciel, une ✗ signifie que le module n'existe pas et un + signifie que le module n'existe pas, par défaut, mais peut être ajouté grâce à des *plugins* externes. Nous pouvons constater que les cinq ETL présentés ne répondent pas à tous les besoins de traitement du projet, d'où la nécessité de développer une nouvelle plateforme. D'une part, les outils ETL orientés informatique décisionnelle (p. ex. Talend) proposent des modules trop génériques et peu adaptés au seul traitement des traces de mobilité et, d'autre part, les ETL spatiaux ont tendance à offrir des fonctionnalités trop étendues qui ne sont pas nécessaires dans nos cas d'utilisation et compliquent les choses pour les utilisateurs non spécialisés. Certains des modules nécessaires au traitement, à l'enrichissement et à la visualisation des traces de mobilité ne sont proposés par aucun outil ETL à l'heure actuelle (p. ex. la visualisation sur un cube spatio-temporel, la détection des arrêts et des déplacements, la similarité des trajectoires, etc.). Enfin, ces ETL proposent un nombre assez important de modules de traitement, dont la plupart n'ont aucune utilité pour le traitement des données de mobilité, ce qui complique grandement l'utilisation du logiciel pour les utilisateurs novices.

D.2 Interactions homme-machine

Nous allons maintenant examiner les différents types d'outils ETL existants du point de vue des interactions homme-machine.

D.2.1 ETL orientés informatique décisionnelle

La plupart des outils ETL ont une interface graphique structurée de la même manière et ont des interactions similaires, à savoir :

— Une palette de modules permet de parcourir tous les modules de traitement disponibles. Ces modules sont répartis en catégories (p. ex. accès aux données, transformation, utilitaire, etc.) et parfois en sous-catégories.

— Un écran permet de visualiser la structure du projet et d'ajouter des fichiers, ou d'autres ressources externes.

— Une zone principale permet à l'utilisateur de construire un pipeline de traitement. Il est généralement basé sur une approche de réseau de nœuds. L'utilisateur n'a qu'à glisser et déposer les modules qu'il souhaite ajouter depuis l'écran de la palette de modules vers celui-ci, et peut les organiser comme il le souhaite. Les modules peuvent être reliés par des connecteurs. Des contrôles de validité sont effectués pour s'assurer que le pipeline de traitement reste cohérent (les connecteurs doivent relier deux paramètres de modules de traitement compatibles). compatibles avec les paramètres du module de traitement).

— Un écran contextuel permet de modifier les paramètres du module actuellement module actuellement sélectionné et éventuellement de visualiser ses sorties.

ETL spatiaux

Les ETL spatiaux se distinguent en offrant des écrans dédiés aux données spatiales. La plupart d'entre eux disposent d'une vue montrant l'arborescence des entités géographiques en cours de traitement. De même, les outils de prévisualisation sont adaptés à ce type de données, à la place de montrer les résultats uniquement sous forme de tableau, ils proposent souvent une vue une vue cartographique avec des options de personnalisation (p. ex. couleur des objets, basculement de la visibilité de l'élément, etc.).

D.2.2 Synthèse

Nom	**Besoins DA3T**	**Talend**	**RapidMiner**	**Pentaho**	**GeoKettle**	**FME Desktop**
Type		ETL	ETL	ETL	ETL spatial	ETL spatial
Caractéristiques courantes						
Palette de modules	✓	✓	✓	✓	✓	✓
Structure de projet en arbre	✗	✓	✓	✓	✓	✓
Zone de construction des chaînes de traitement	✓	✓	✓	✓	✓	✓

Nom	**Besoins DA3T**	**Talend**	**RapidMiner**	**Pentaho**	**GeoKettle**	**FME Desktop**
Tableau de résultats	✗	✓	✓	✓	✓	✓
Menu de paramétrage	✓	✓	✓	✓	✓	✓
Caractéristiques spatiales						
Écran cartographique de prévisualisation	✓	✗	✗	✗	✓	✓
Visualisation des objets géographiques	✗	✗	✗	✗	✓	✓
Caractéristiques requises						
Spécification de haut niveau	✓	✗	✗	✗	✗	✗
Facilité d'utilisation pour des utilisateurs novices (p. ex. peu de modules, interface simple, etc.)	✓	✗	✗	✗	✗	✗
Évolutivité (p. ex. facile à déployer, possibilité d'ajout de nouveaux modules, mise à jour, etc.)	✓	✗	✗	✗	✗	✗
Flexibilité et modularité	✓	✓	✓	✓	✓	✓

Table D.2 – Comparaison de plusieurs ETL par rapport à leurs interfaces homme-machine et aux besoins relatifs à la plateforme DA3T

Pour conclure, les outils ETL ont souvent un champ d'utilisation très large, à la fois pour les données d'entreprise avec l'informatique décisionnelle mais également pour les données géographiques au sens large. Il existe même des ETL dédiés au traitement du texte et du langage naturel (p. ex. GATE [7] ou LinguaStream [8]). Dans le cadre du projet DA3T, nous nous concentrons sur le traitement des données de mobilité. Comme le montre le tableau D.2, il

7. https ://gate.ac.uk/
8. http ://www.linguastream.org/

n’existe pas d’outil ETL strictement axé sur ces données. De plus, notre logiciel est dédié à des utilisateurs non informaticiens qui doit donc être aussi accessible et clair que possible. Ce n’est pas le cas des ETL existants. Ils doivent être le plus génériques possible et s’adapter à tout type de données (p. ex. données d’entreprise ou données spatiales) et ont donc une complexité de compréhension assez élevée et beaucoup de modules. Il est également assez difficile de les faire évoluer car cela doit passer par la mise en place de *plugins* qui doivent être redéployés auprès de tous les utilisateurs à chaque changement.

Annexe E

Paires de trajectoires sémantiques

E.0.1 Dimensions spatiale

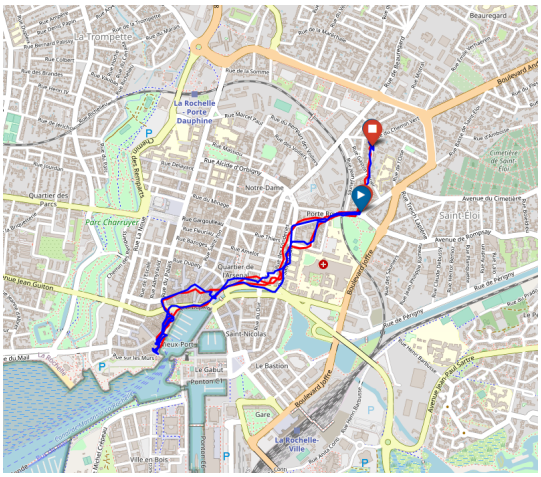


Figure E.1 – Paire n°1 : trajectoire n°5 (en rouge) et trajectoire n°6 (en bleu)

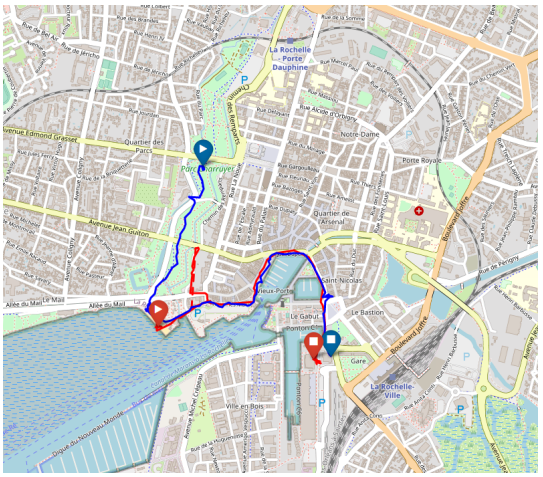


Figure E.2 – Paire n°2 : trajectoire n°93 (en rouge) et trajectoire n°103 (en bleu)

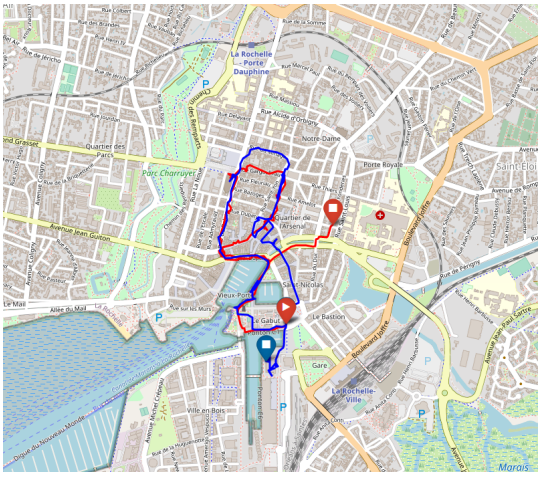


Figure E.3 – Paire n°3 : trajectoire n°21 (en rouge) et trajectoire n°31 (en bleu)

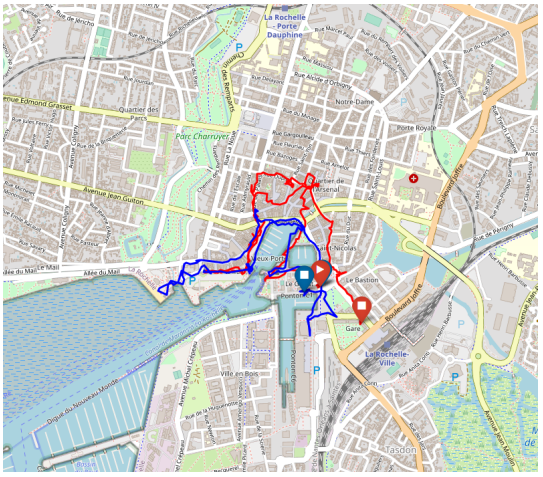


Figure E.4 – Paire n°4 : trajectoire n°68 (en rouge) et trajectoire n°90 (en bleu)

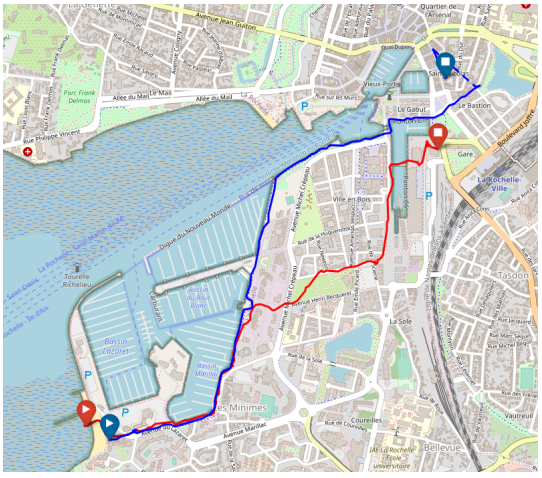


Figure E.5 – Paire n°5 : trajectoire n°92 (en rouge) et trajectoire n°137 (en bleu)

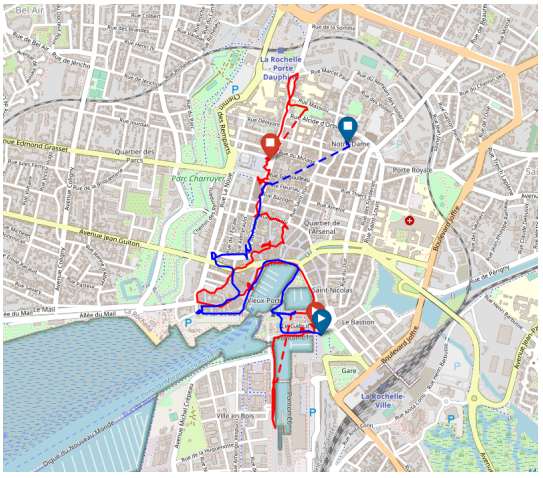


Figure E.6 – Paire n°6 : trajectoire n°65 (en rouge) et trajectoire n°136 (en bleu)

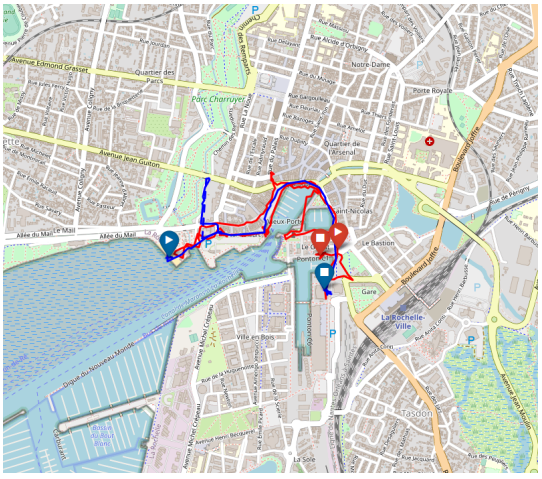


Figure E.7 – Paire n°7 : trajectoire n°90 (en rouge) et trajectoire n°93 (en bleu)

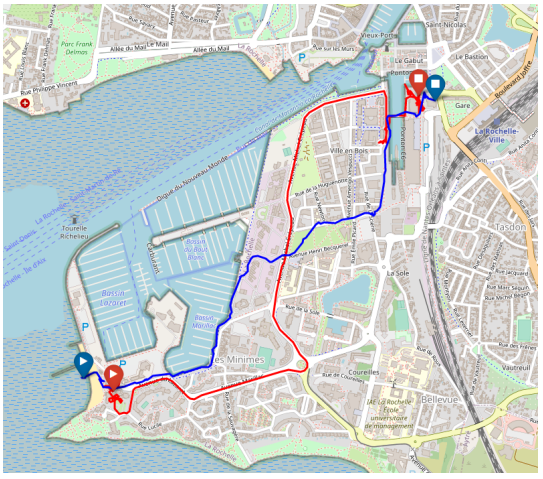


Figure E.8 – Paire n°8 : trajectoire n°61 (en rouge) et trajectoire n°92 (en bleu)

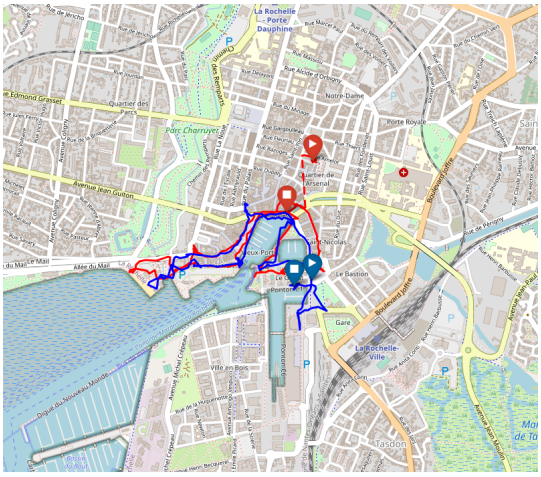


Figure E.9 – Paire n°9 : trajectoire n°23 (en rouge) et trajectoire n°90 (en bleu)

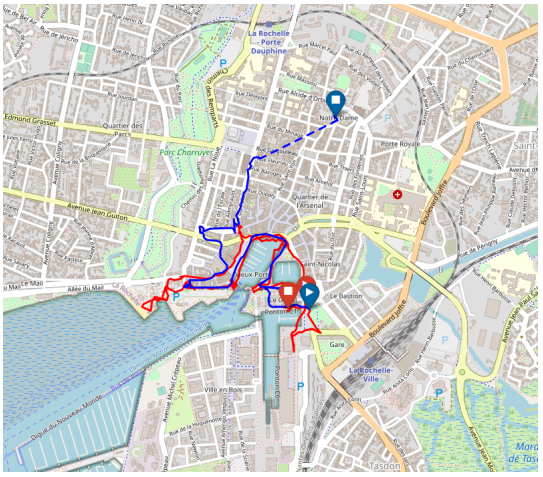


Figure E.10 – Paire n°10 : trajectoire n°90 (en rouge) et trajectoire n°136 (en bleu)

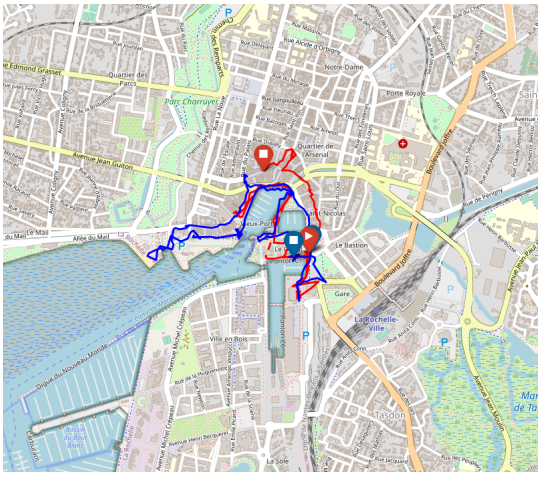


Figure E.11 – Paire n°11 : trajectoire n°71 (en rouge) et trajectoire n°90 (en bleu)

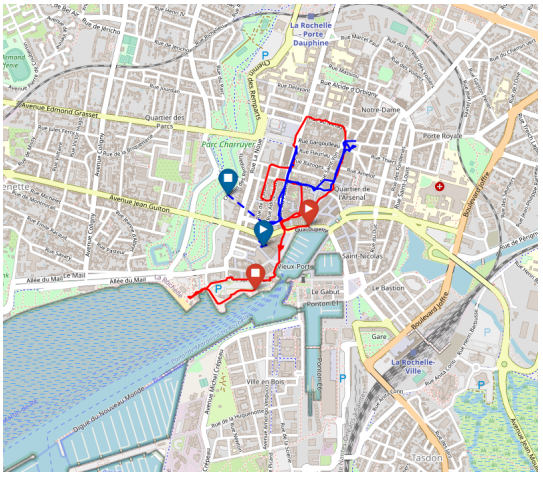


Figure E.12 – Paire n°12 : trajectoire n°115 (en rouge) et trajectoire n°147 (en bleu)

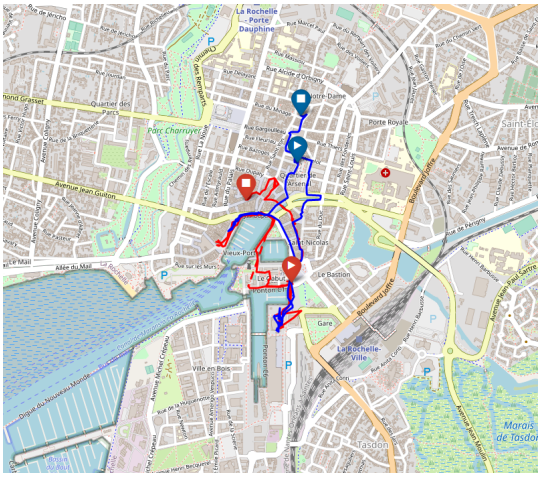


Figure E.13 – Paire n°13 : trajectoire n°71 (en rouge) et trajectoire n°107 (en bleu)

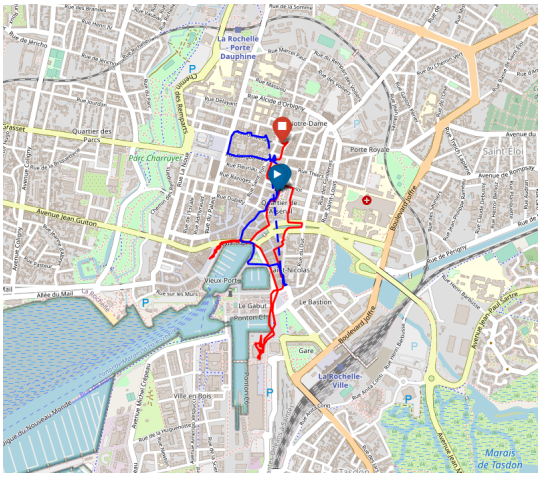


Figure E.14 – Paire n°14 : trajectoire n°107 (en rouge) et trajectoire n°162 (en bleu)

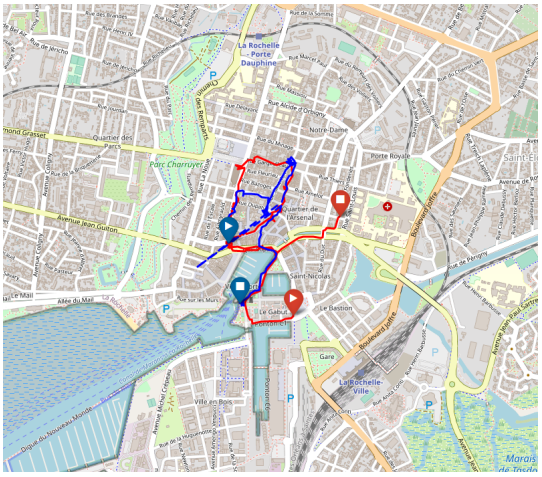


Figure E.15 – Paire n°15 : trajectoire n°21 (en rouge) et trajectoire n°113 (en bleu)

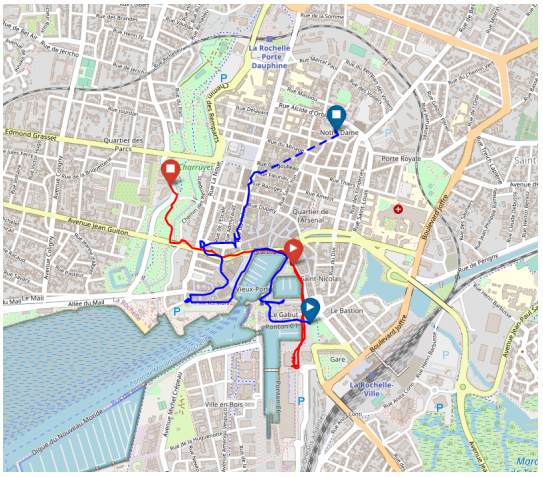


Figure E.16 – Paire n°16 : trajectoire n°109 (en rouge) et trajectoire n°136 (en bleu)

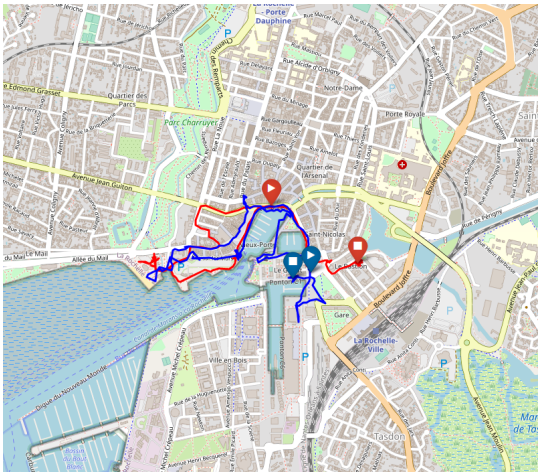


Figure E.17 – Paire n°17 : trajectoire n°69 (en rouge) et trajectoire n°90 (en bleu)

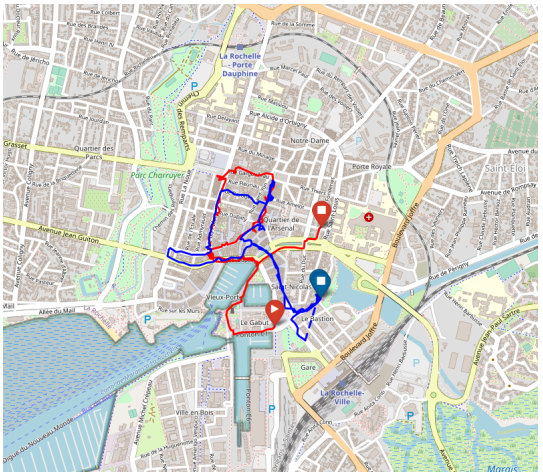


Figure E.18 – Paire n°18 : trajectoire n°21 (en rouge) et trajectoire n°27 (en bleu)

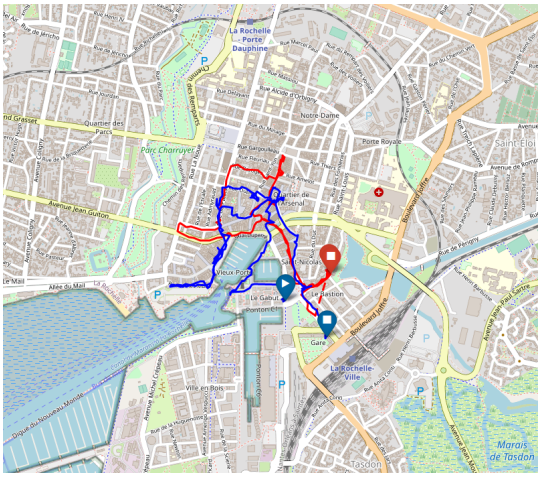


Figure E.19 – Paire n°19 : trajectoire n°27 (en rouge) et trajectoire n°68 (en bleu)

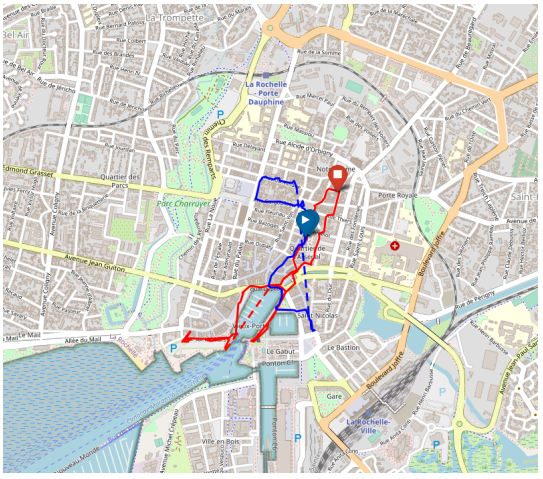


Figure E.20 – Paire n°20 : trajectoire n°17 (en rouge) et trajectoire n°162 (en bleu)

Figure E.21 – Paire n°21 : trajectoire n°93 (en rouge) et trajectoire n°92 (en bleu)

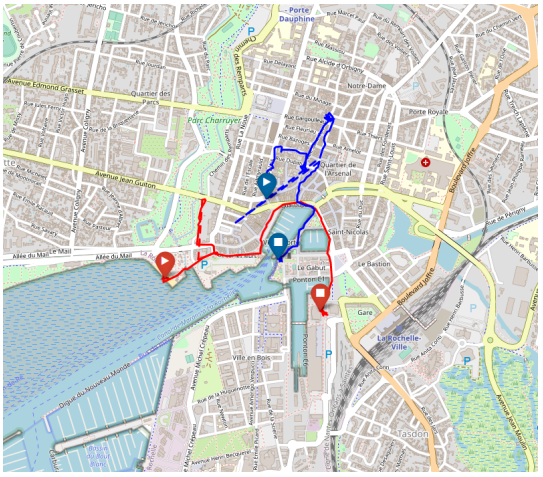


Figure E.22 – Paire n°22 : trajectoire n°93 (en rouge) et trajectoire n°113 (en bleu)

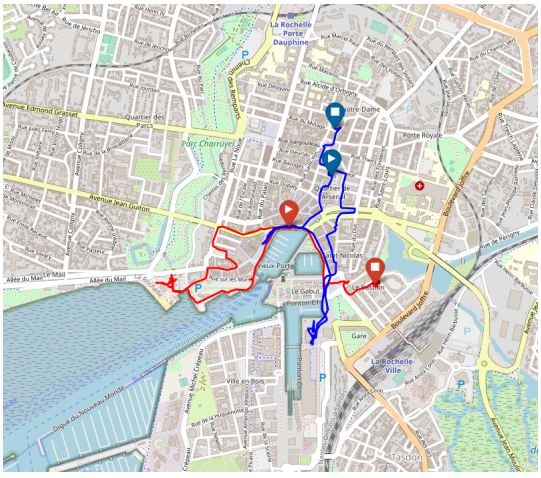


Figure E.23 – Paire n°23 : trajectoire n°69 (en rouge) et trajectoire n°107 (en bleu)

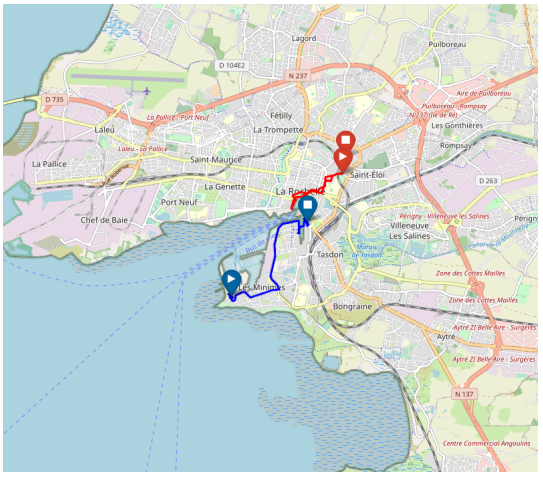


Figure E.24 – Paire n°24 : trajectoire n°6 (en rouge) et trajectoire n°61 (en bleu)

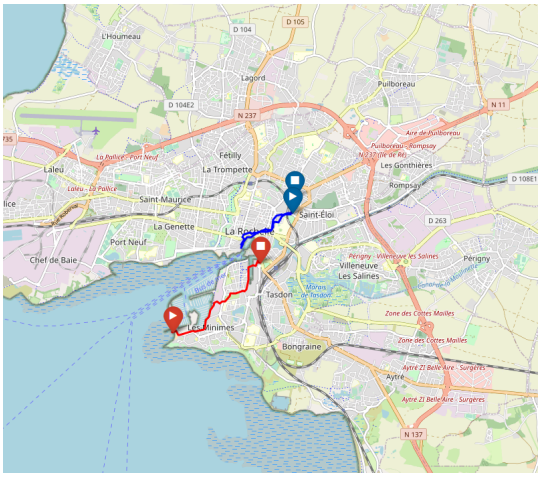


Figure E.25 – Paire n°25 : trajectoire n°92 (en rouge) et trajectoire n°5 (en bleu)

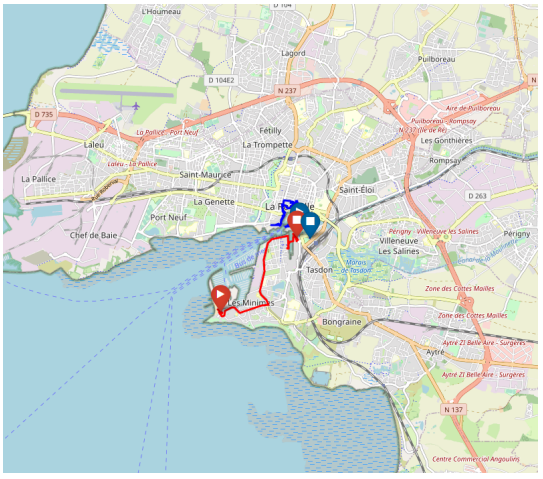


Figure E.26 – Paire n°26 : trajectoire n°61 (en rouge) et trajectoire n°68 (en bleu)

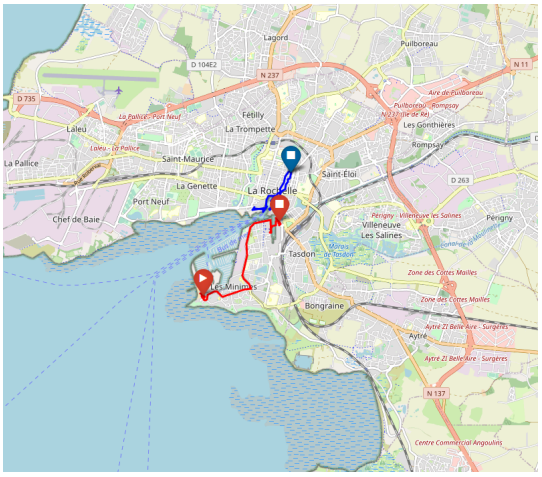


Figure E.27 – Paire n°27 : trajectoire n°61 (en rouge) et trajectoire n°17 (en bleu)

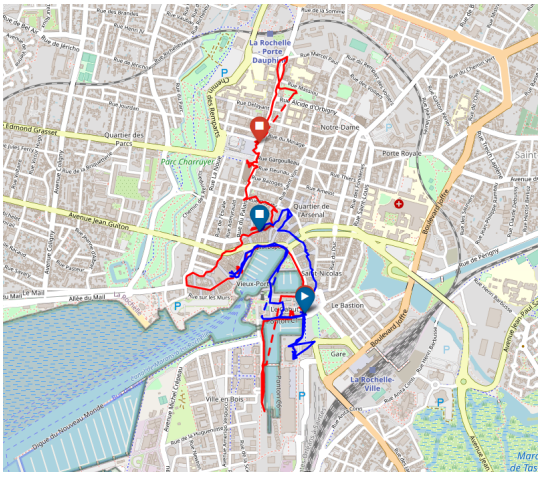


Figure E.28 – Paire n°28 : trajectoire n°65 (en rouge) et trajectoire n°71 (en bleu)

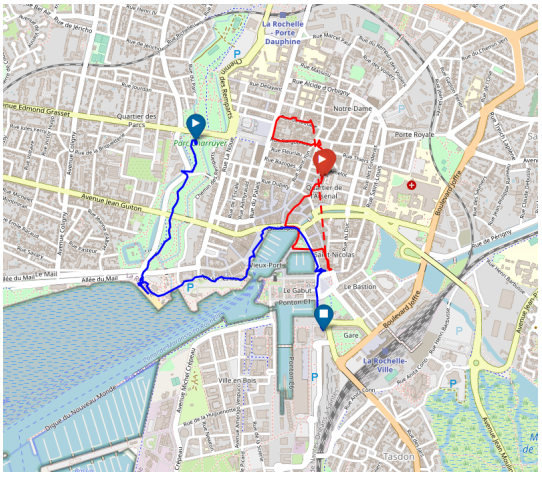


Figure E.29 – Paire n°29 : trajectoire n°162 (en rouge) et trajectoire n°103 (en bleu)

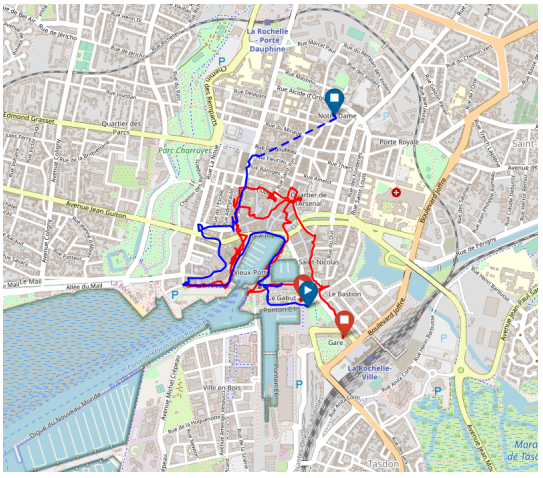


Figure E.30 – Paire n°30 : trajectoire n°68 (en rouge) et trajectoire n°136 (en bleu)

E.0.2 Dimension thématique et temporelle

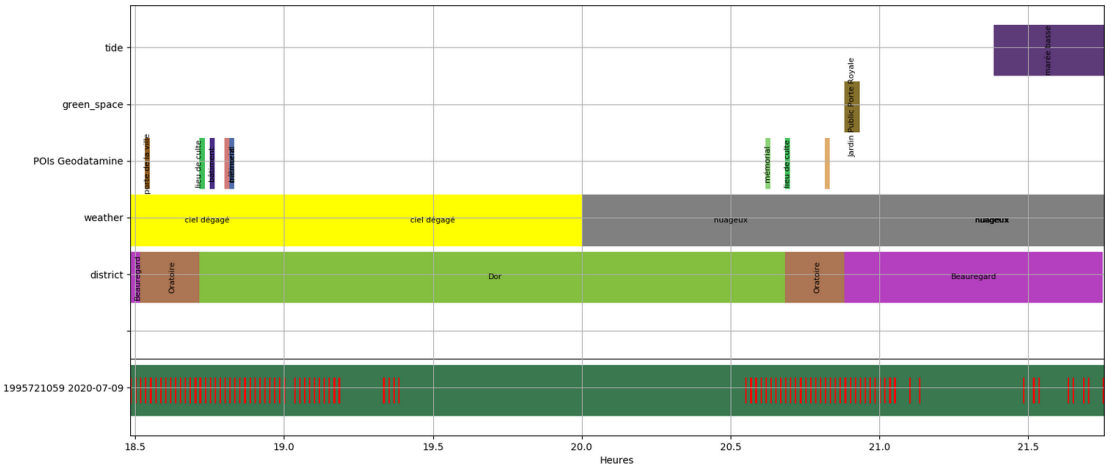


Figure E.31 – Dimension thématique de la trajectoire n°5

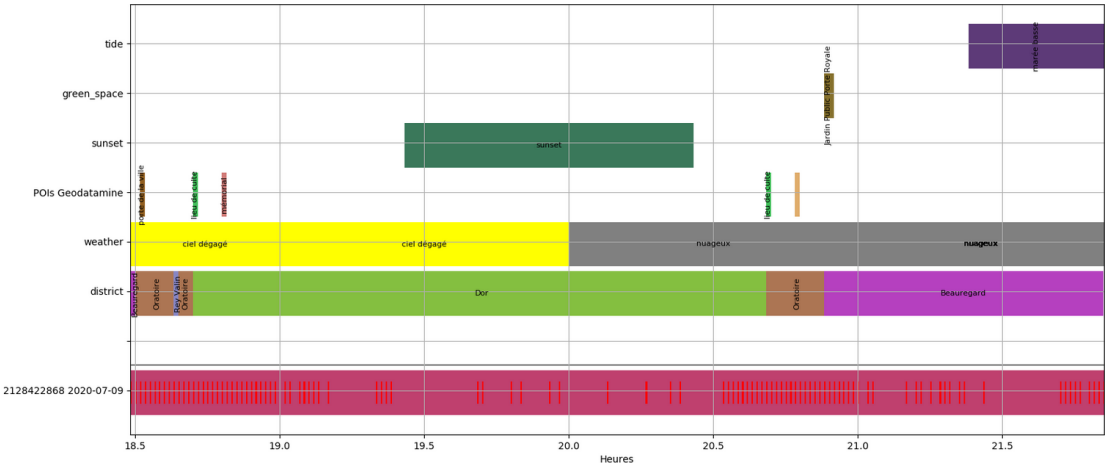


Figure E.32 – Dimension thématique de la trajectoire n°6

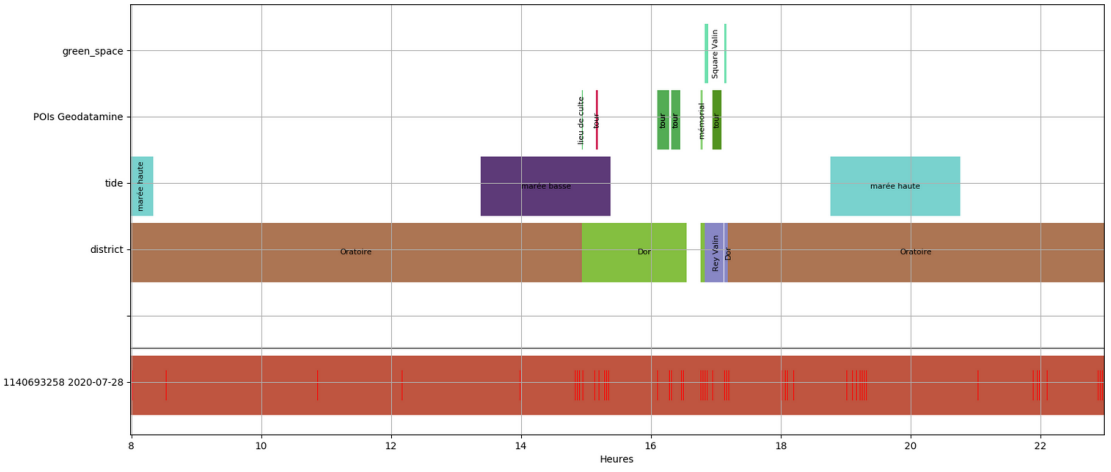


Figure E.33 – Dimension thématique de la trajectoire n°17

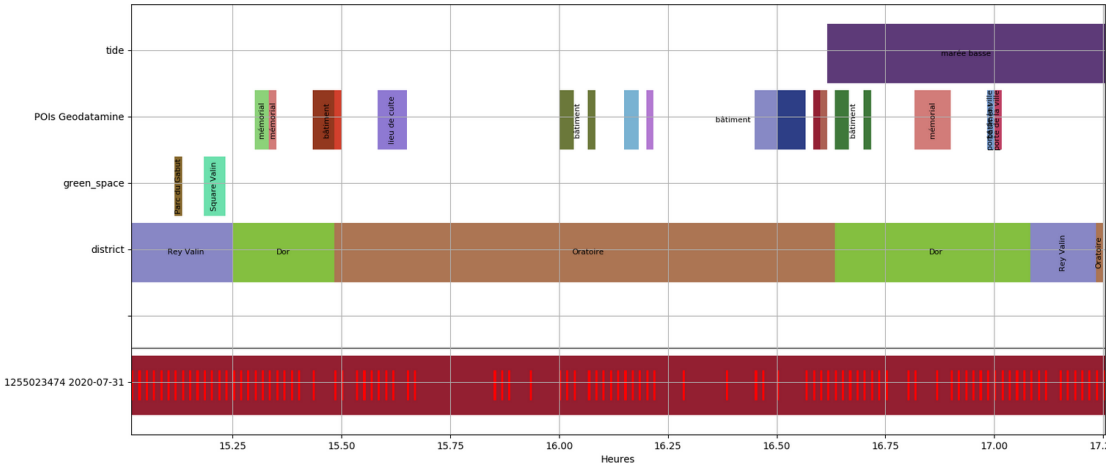


Figure E.34 – Dimension thématique de la trajectoire n°21

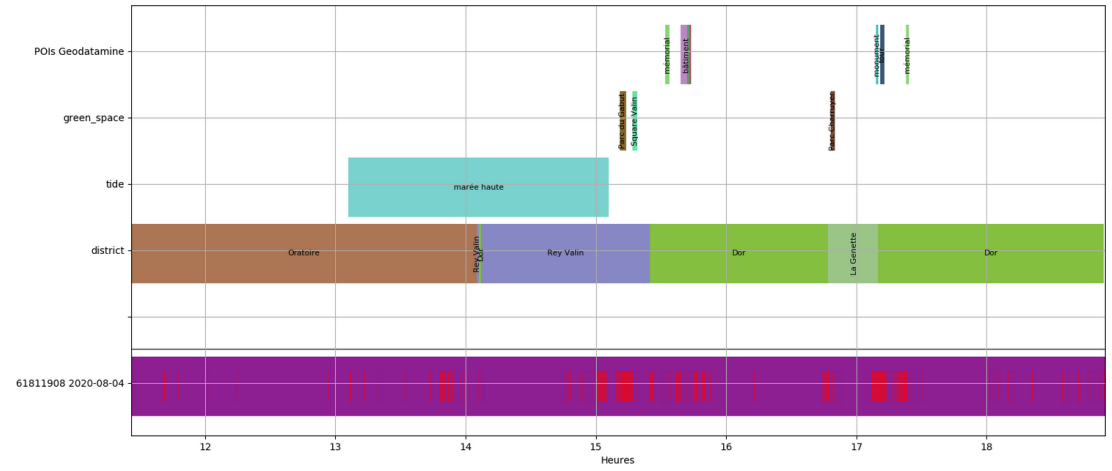


Figure E.35 – Dimension thématique de la trajectoire n°23

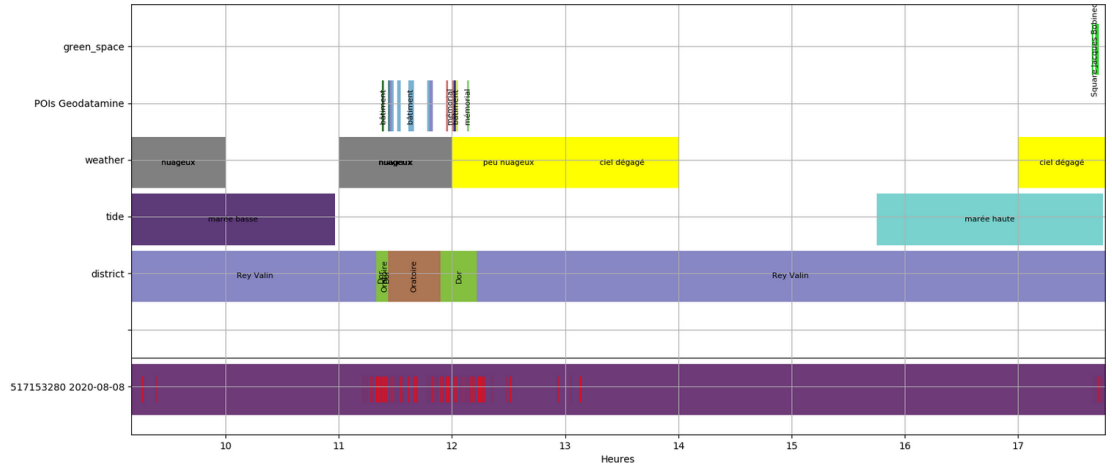


Figure E.36 – Dimension thématique de la trajectoire n°27

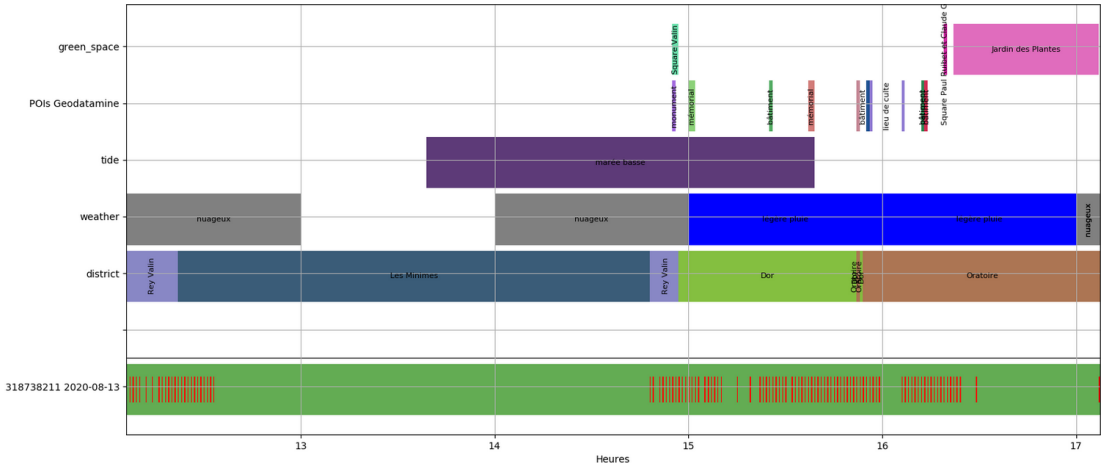


Figure E.37 – Dimension thématique de la trajectoire n°31

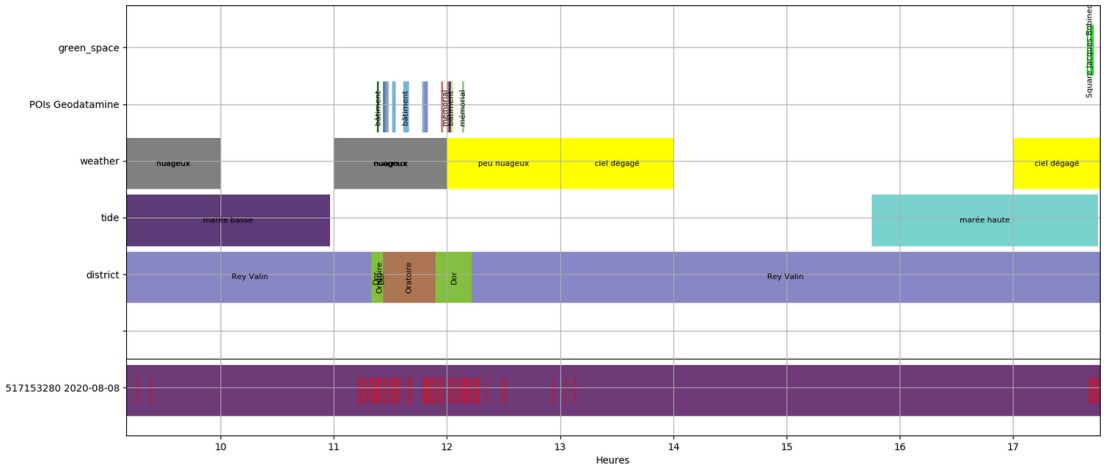


Figure E.38 – Dimension thématique de la trajectoire n°61

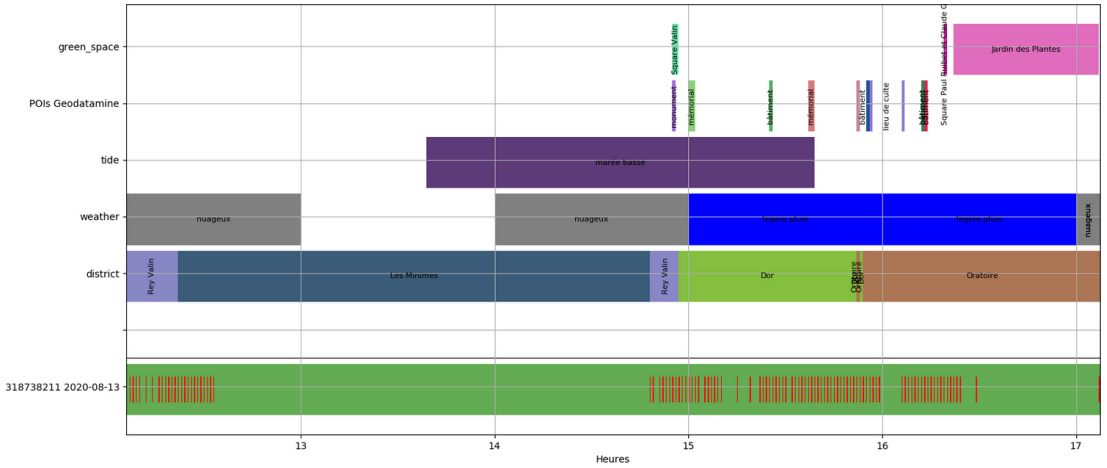


Figure E.39 – Dimension thématique de la trajectoire n°65

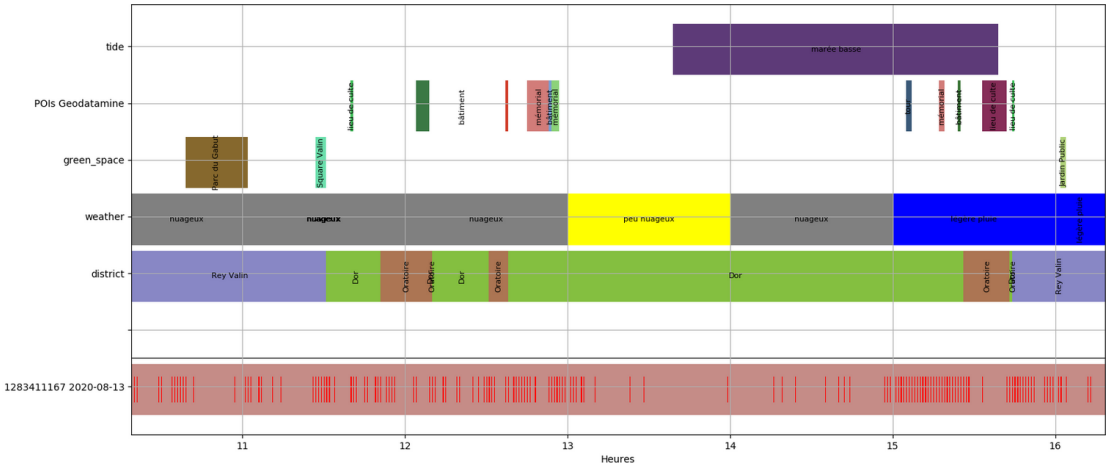


Figure E.40 – Dimension thématique de la trajectoire n°68

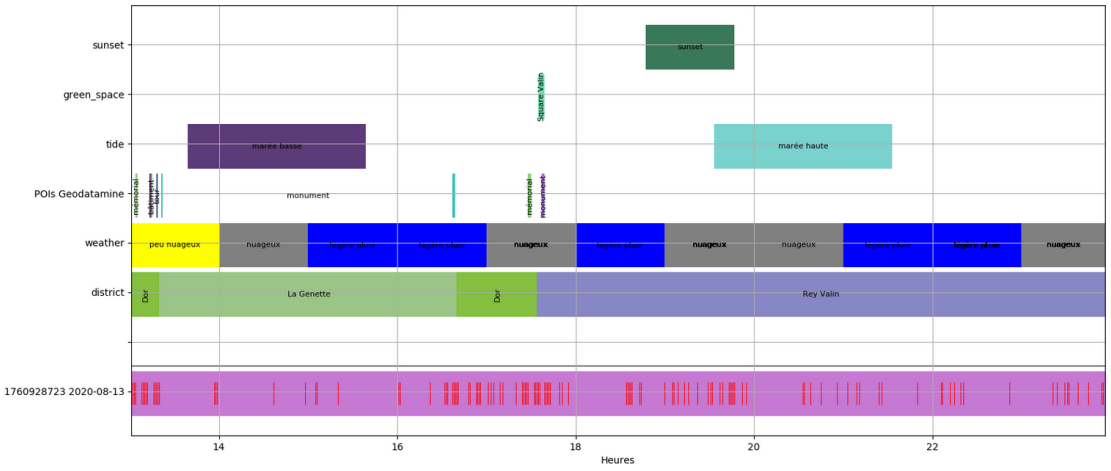


Figure E.41 – Dimension thématique de la trajectoire n°69

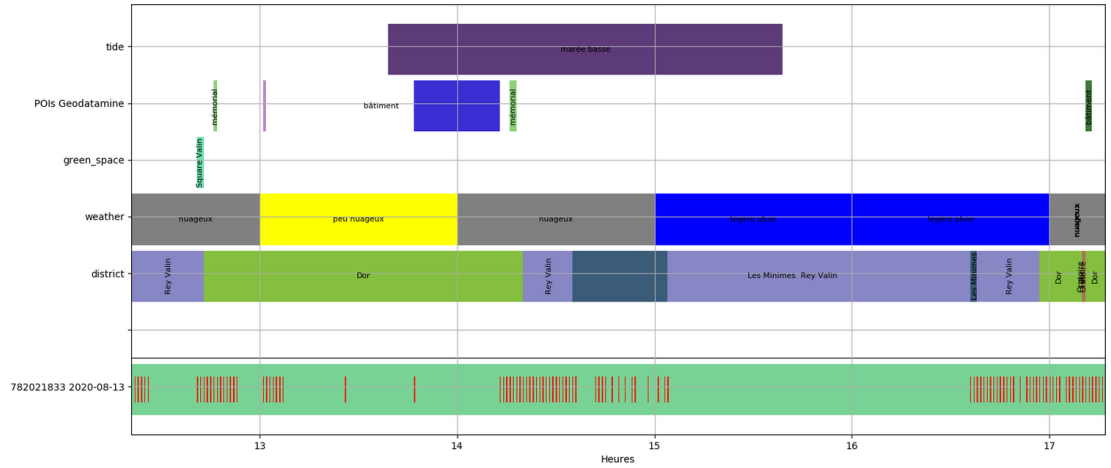


Figure E.42 – Dimension thématique de la trajectoire n°71

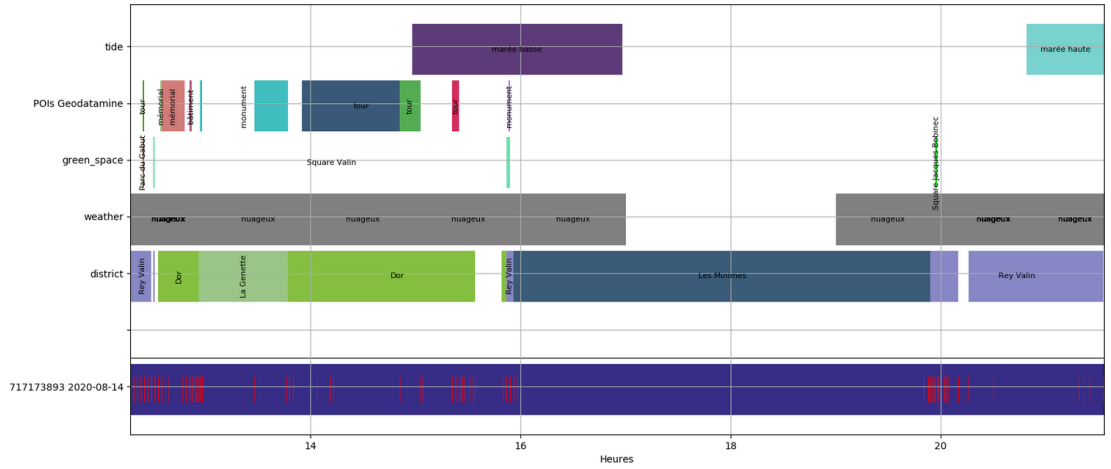


Figure E.43 – Dimension thématique de la trajectoire n°90

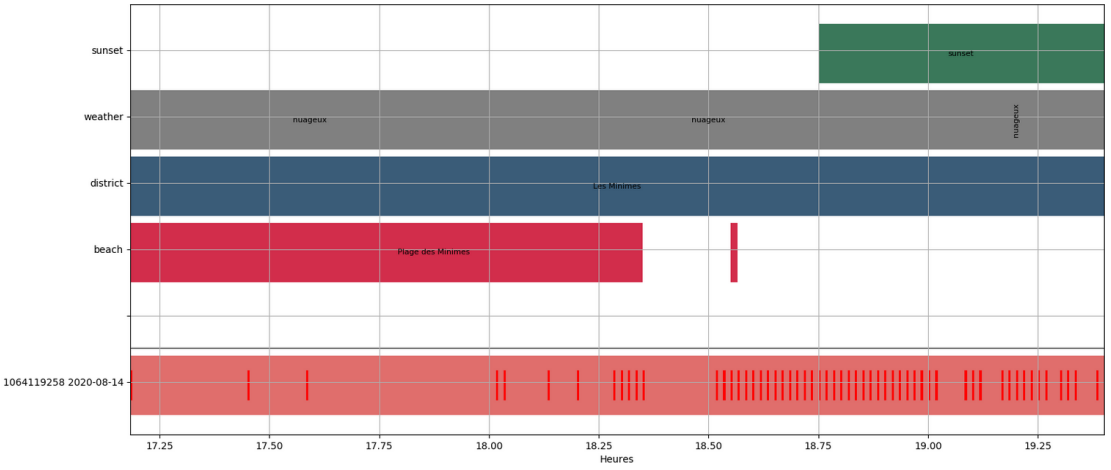


Figure E.44 – Dimension thématique de la trajectoire n°92

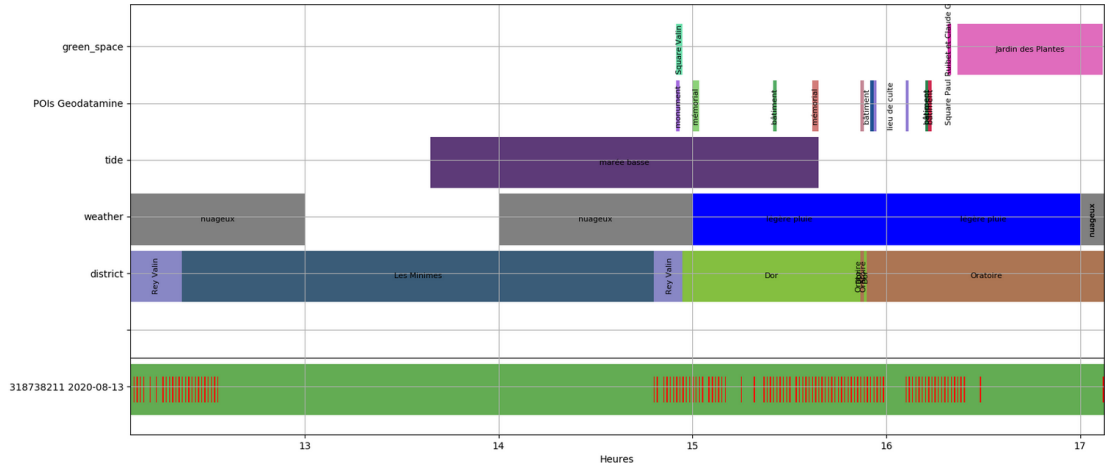


Figure E.45 – Dimension thématique de la trajectoire n°93

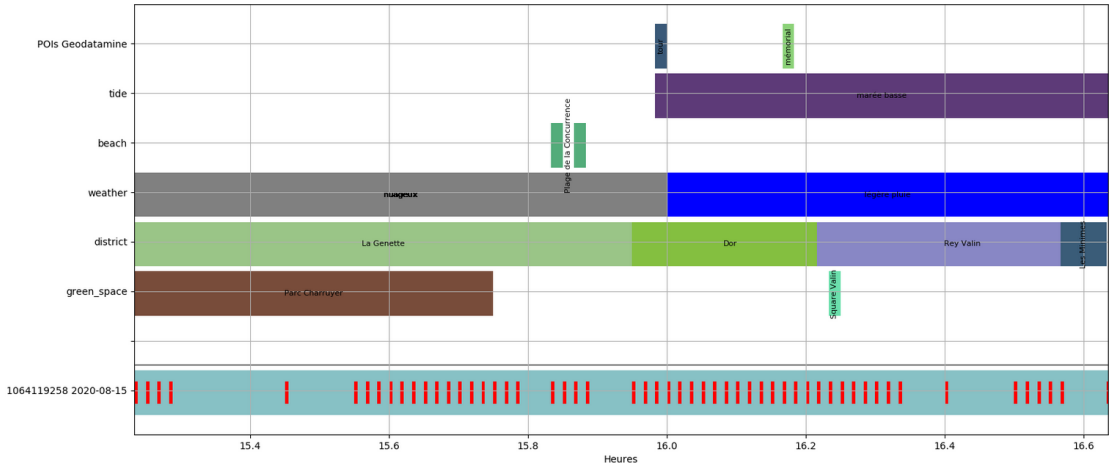


Figure E.46 – Dimension thématique de la trajectoire n°103

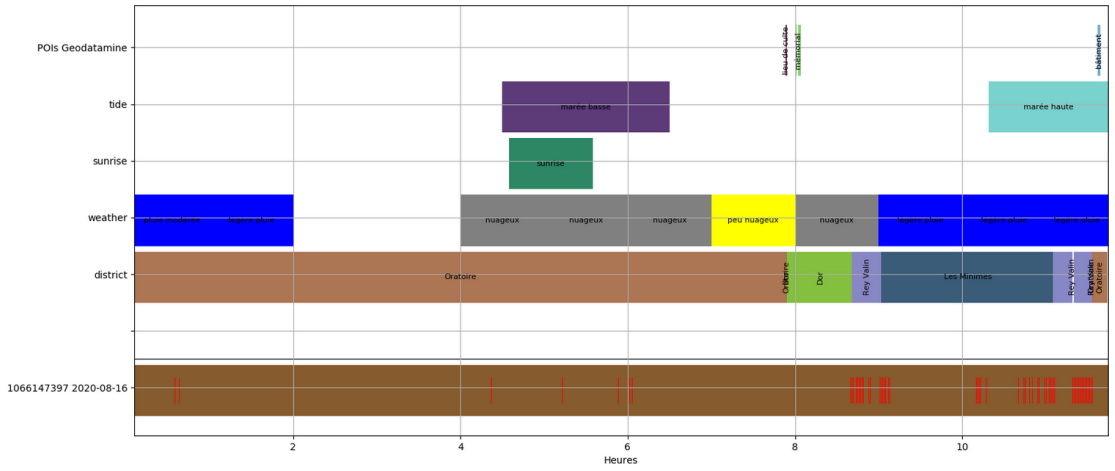


Figure E.47 – Dimension thématique de la trajectoire n°107

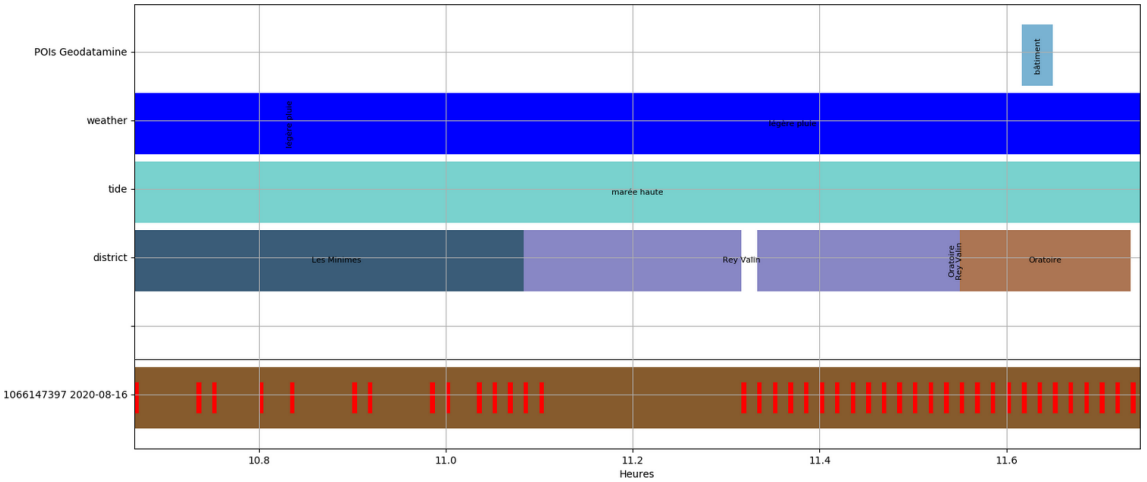


Figure E.48 – Marée haute et poi de la trajectoire n°107

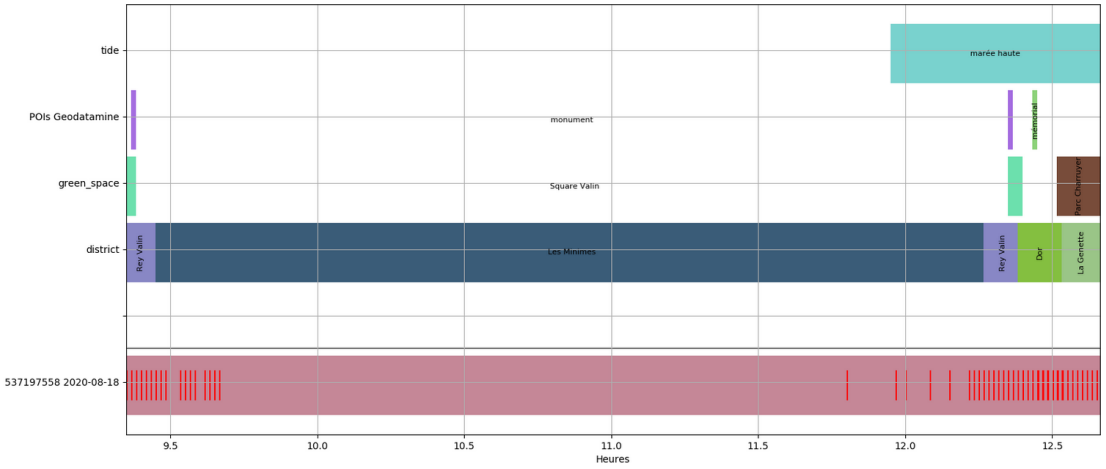


Figure E.49 – Dimension thématique de la trajectoire n°109

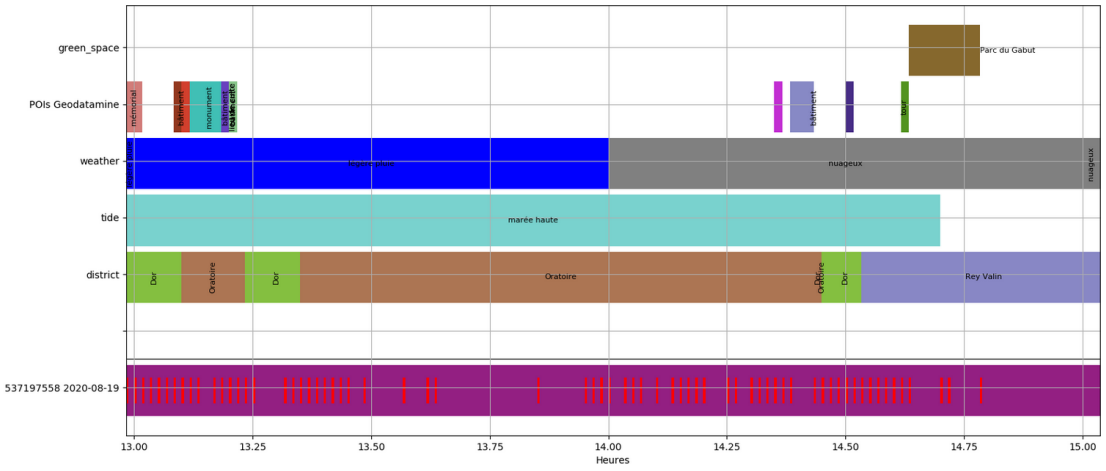


Figure E.50 – Dimension thématique de la trajectoire n°113

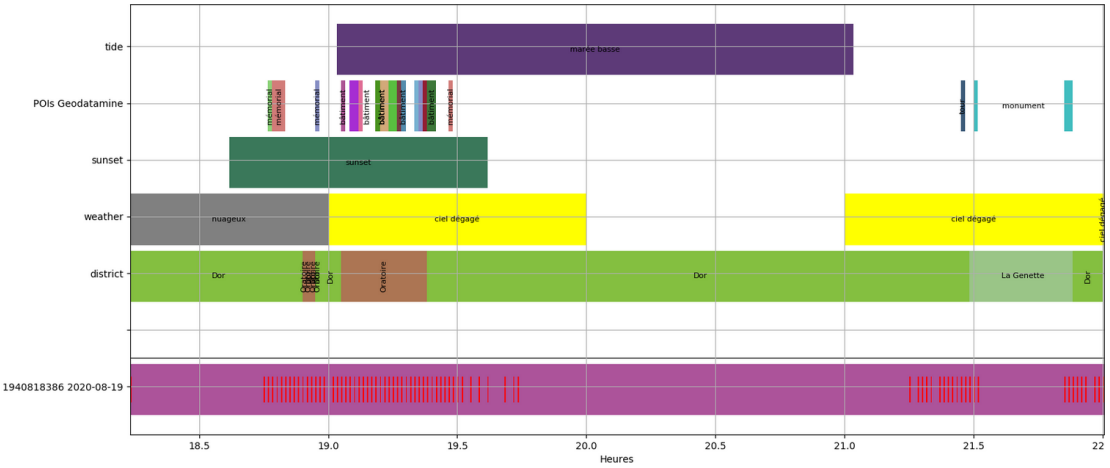


Figure E.51 – Dimension thématique de la trajectoire n°115

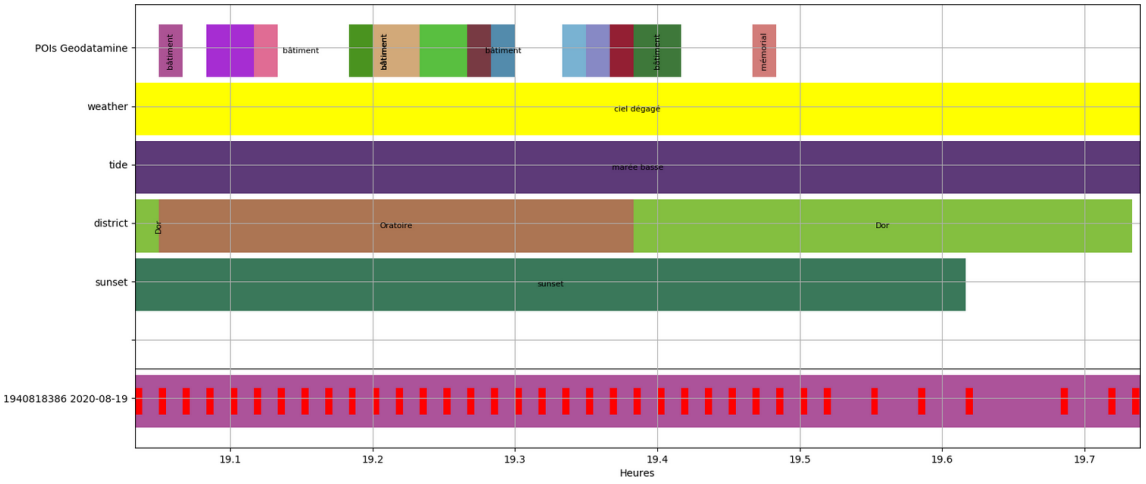


Figure E.52 – Marée haute et poi de la trajectoire n°115

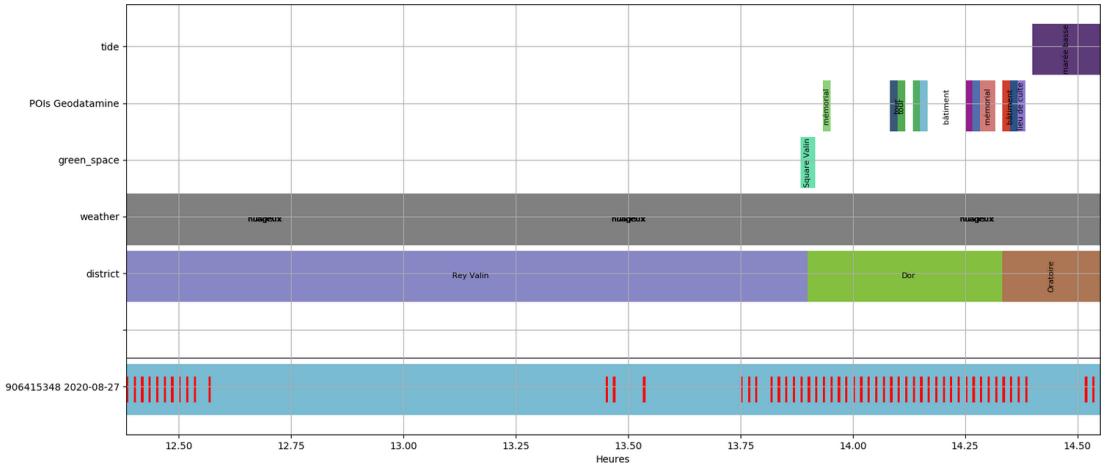


Figure E.53 – Dimension thématique de la trajectoire n°136

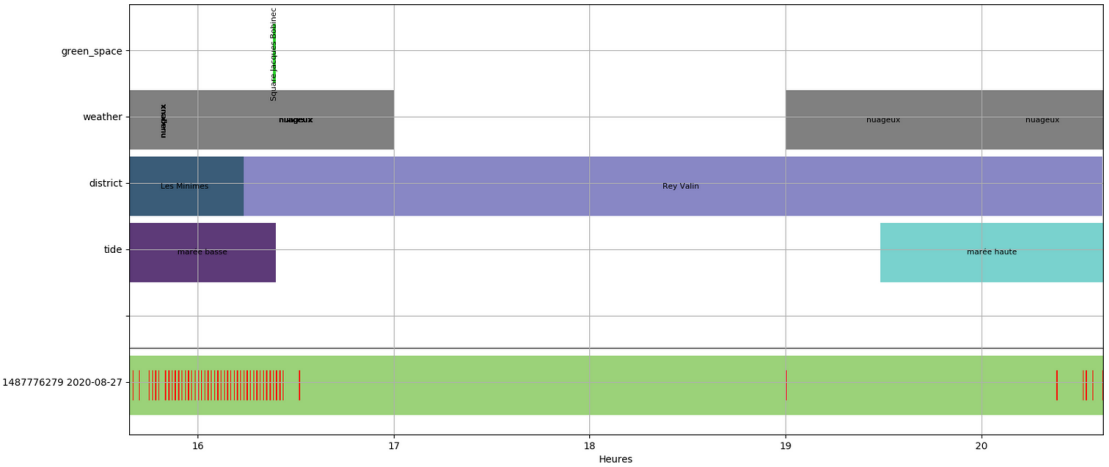


Figure E.54 – Dimension thématique de la trajectoire n°137

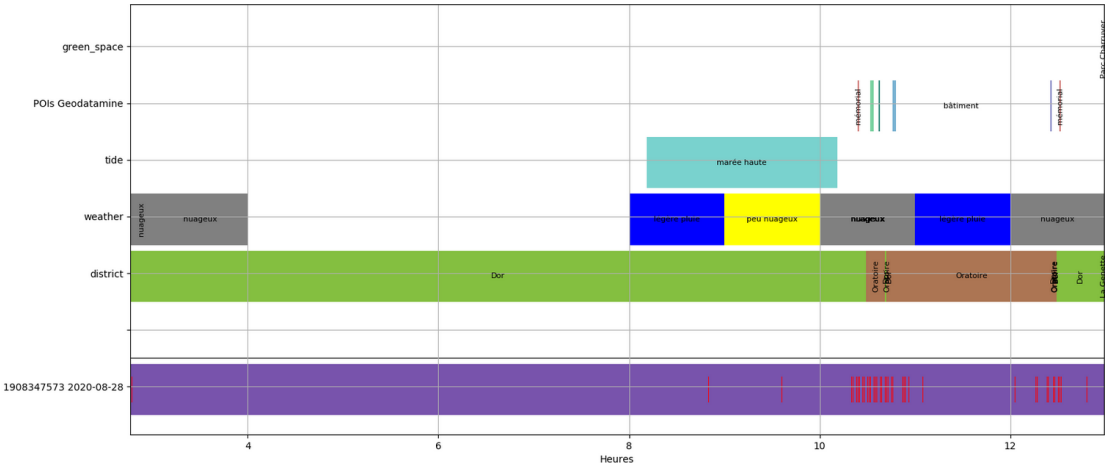


Figure E.55 – Dimension thématique de la trajectoire n°147

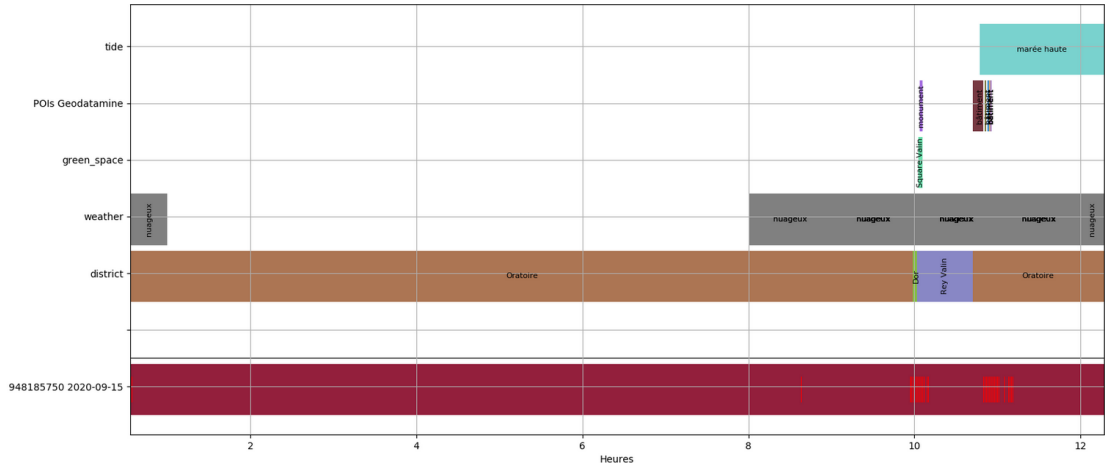


Figure E.56 – Dimension thématique de la trajectoire n°162

Bibliographie

Marco Aiello. A spatial similarity measure based on games : Theory and practice. *Logic Journal of IGPL*, 10, January 2002.

Basma H. Albanna, Ibrahim F. Moawad, Sherin M. Moussa, and Mahmoud A. Sakr. Semantic Trajectories : A Survey from Modeling to Application. In Vasily Popovich, Christophe Claramunt, Manfred Schrenk, Kyrill Korolenko, and Jérôme Gensel, editors, *Information Fusion and Geographic Information Systems (IF&GIS' 2015) : Deep Virtualization for Mobile GIS*, Lecture Notes in Geoinformation and Cartography, pages 59–76. Springer International Publishing, Cham, 2015.

James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11) :832–843, November 1983.

Helmut Alt. The Computational Geometry of Comparing Shapes. In *Efficient Algorithms : Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, pages 235–248. Springer-Verlag, Berlin, Heidelberg, September 2009.

Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, José Antônio Fernandes de Macêdo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, GIS '07, pages 1–8, Seattle, Washington, November 2007. Association for Computing Machinery.

James Andrew. Astronomical and nautical tables. 1805.

Landy Andriamampianina, Franck Ravat, Jiefu Song, and Nathalie Vallès-Parlangeau. Towards an efficient approach to manage graph data evolution : conceptual modelling and experimental assessments. In Samira Cherfi, Anna Perini, and Selmin Nurcan, editors, *Research Challenges in Information Science. RCIS 2021*, volume 415 of *Lecture Notes in Business Information Processing book series (LNBIP)*, pages 471–488, virtual, Cyprus, May 2021.

Muhammad Arslan, Christophe Cruz, and Dominique Ginhac. Understanding Worker Mobility within the Stay Locations using HMMs on Semantic Trajectories. In *2018 14th International Conference on Emerging Technologies (ICET)*, pages 1–6, November 2018.

Winda Astriani and Rina Trisminingsih. Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle. *Procedia Environmental Sciences*, 33 :626–634, January 2016.

Miriam Baglioni, José Antônio Fernandes de Macêdo, Chiara Renso, and Monica Wachowicz. An Ontology-Based Approach for the Semantic Modelling and Reasoning on Trajectories. In Il-Yeol Song, Mario Piattini, Yi-Ping Phoebe Chen, Sven Hartmann, Fabio Grandi, Juan Trujillo, Andreas L. Opdahl, Fernando Ferri, Patrizia Grifoni, Maria Chiara Caschera, Colette Rolland, Carson Woo, Camille Salinesi, Esteban Zimányi, Christophe Claramunt, Flavius Frasincar, Geert-Jan Houben, and Philippe Thiran, editors, *Advances in Conceptual Modeling – Challenges and Opportunities*, Lecture Notes in Computer Science, pages 344–353, Berlin, Heidelberg, 2008. Springer.

Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. CONSTAnT – A Conceptual Data Model for Semantic Trajectories of Moving Objects. *Transactions in GIS*, 18(1) :66–88, 2014.

Roland Bouman and Jos Van Dongen. Pentaho solutions. *Business Intelligence and Data Warehousing with Pentaho and MYSQL*, 2009.

Cécile Cayèré. Plateforme ETL dédiée à l'analyse de la mobilité touristique dans une ville. In *Proceedings of the forum Jeunes Chercheuses Jeunes Chercheurs, INFORSID*, pages 13–16, Dijon, France, April 2020.

Cécile Cayèré, Cyril Faucher, Christian Sallaberry, Marie-Noelle Bessagnet, and Philippe Roose. Tools for processing digital trajectories of tourists. In *Proceedings of the 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 232–233, Versailles, France, June 2020. doi : 10.1109/MDM48529.2020.00049.

Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noelle Bessagnet, and Philippe Roose. Proposition d'un modèle de trajectoires multi-aspects et multi-niveaux appliqué au tourisme. In *Proceedings of the Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA'21)*, pages 56–64, Bordeaux, France, June 2021a.

Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose, Maxime Masson, and Jérémy Richard. Multi-Level and Multiple Aspect Semantic Trajectory Model : Application to the Tourism Domain. *ISPRS International Journal of Geo-Information*, 10(9) :592, September 2021b.

Cécile Cayèré, Christian Sallaberry, Cyril Faucher, Marie-Noëlle Bessagnet, Philippe Roose, and Maxime Masson. Mesure de similarité pour les trajectoires sémantiques : prise en compte de trois niveaux de granularité. In *Proceedings of the 40th INFORSID Congrès*, pages 5–20, Dijon, France, May 2022.

Jaydeep Chakraborty, Aparna Padki, and Srividya K. Bansal. Semantic ETL — State-of-the-Art and Open Research Challenges. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 413–418, January 2017.

Lei Chen and Raymond Ng. On The Marriage of Lp-norms and Edit Distance. pages 792–803, January 2004.

Lei Chen, M. Tamer Özsu, and Vincent Oria. Symbolic representation and retrieval of moving object trajectories. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '04, pages 227–234, New York, NY, USA, October 2004. Association for Computing Machinery.

Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 491–502, January 2005.

Lisi Chen, Shuo Shang, Christian S. Jensen, Bin Yao, and Panos Kalnis. Parallel Semantic Trajectory Similarity Join. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 997–1008, April 2020.

Shihyen Chen, Bin Ma, and Kaizhong Zhang. On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24) :2365–2376, May 2009. ISSN 0304-3975. doi : 10.1016/j.tcs.2009.02.023.

Yueguo Chen, Mario A. Nascimento, Beng Chin Ooi, and Anthony K. H. Tung. SpADe : On Shape-based Pattern Detection in Streaming Time Series. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 786–795, April 2007.

Ian R. Cleasby, Ewan D. Wakefield, Barbara J. Morrissey, Thomas W. Bodey, Steven C. Votier, Stuart Bearhop, and Keith C. Hamer. Using time-series similarity measures to compare animal movement trajectories in ecology. *Behavioral Ecology and Sociobiology*, 73(11) : 151, November 2019.

DATAtourisme. DATAtourisme. `https://www.datatourisme.fr/`, a.

DATAtourisme. DATAtoursime Ontology. `https://framagit.org/datatourisme/ontology`, b.

Urška Dreščček, Mojca Kosmatin Fras, Jernej Tekavec, and Anka Lisec. Spatial ETL for 3D Building Modelling Based on Unmanned Aerial Vehicle Data in Semi-Urban Areas. *Remote Sensing*, 12(12) :1972, January 2020.

Max J. Egenhofer. Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages & Computing*, 8(4) :403–424, August 1997.

Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. *ACM SIGMOD Record*, 23(2) :419–429, May 1994.

Renato Fileto, Marcelo Krüger, Nikos Pelekis, Yannis Theodoridis, and Chiara Renso. Baquara : A Holistic Ontological Framework for Movement Analysis Using Linked Data. In Wilfred Ng, Veda C. Storey, and Juan C. Trujillo, editors, *Conceptual Modeling*, Lecture Notes in Computer Science, pages 342–355, Berlin, Heidelberg, 2013. Springer.

Renato Fileto, Cleto May, Chiara Renso, Nikos Pelekis, Douglas Klein, and Yannis Theodoridis. The Baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98 :104–122, July 2015.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5) :378–382, 1971.

Frédéric Flouvat. *Extraction de motifs spatio-temporels : co-localisations, séquences et graphes dynamiques attribués*. thesis, Université de la Nouvelle-Calédonie, October 2019.

Éric Foley and Manon Guillemette. What is Business Intelligence ? *IJBIR*, 1 :1–28, October 2010.

Ali Frihida, Donia Zheni, Henda Ben Ghezala, and Christophe Claramunt. Modeling Trajectories : A Spatio-Temporal Data Type Approach. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 447–451, August 2009.

Andre Salvaro Furtado, Luis Otavio Campos Alvares, Nikos Pelekis, Yannis Theodoridis, and Vania Bogorny. Unveiling movement uncertainty for robust trajectory similarity analysis. *International Journal of Geographical Information Science*, 32(1) :140–168, January 2018.

Jérôme Gensel, Marlène Villanova-Oliver, Pierre Le Quéau, and David Noël. Un modèle multi points de vue pour représenter les trajectoires de vie. In *CIST2020 - Population, temps, territoires*, pages 173–177, Paris-Aubervilliers, France, November 2020. Collège international des sciences territoriales (CIST).

Google. Google Places. `https://developers.google.com/maps/documentation/places/web-service/overview?hl=fr`.

Markus Hofmann and Ralf Klinkenberg, editors. *RapidMiner : Data Mining Use Cases and Business Analytics Applications*. Chapman and Hall/CRC, 0 edition, April 2016.

Richard Jérémy, Bertet Karell, and Faucher Cyril. Ble Based Indoor Positioning System and Minimal Zone Searching Algorithm (MZS) Applied to Visitor Trajectories within a Museum. *Applied Sciences*, 11(13) :6107, 2021. Publisher : MDPI.

Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3) :358–386, March 2005.

La Rochelle. La Rochelle Open Data. `https://opendata.agglo-larochelle.fr/accueil`.

Juliet S. Lamb, Yvan G. Satgé, and Patrick G. R. Jodice. Influence of density-dependent competition on foraging and migratory behavior of a subtropical colonial seabird. *Ecology and Evolution*, 7(16) :6469–6481, 2017.

Annig Le Parc-Lacayrelle, Mauro Gaio, and Christian Sallaberry. La composante temps dans l'information géographique textuelle. *Document Numérique*, 10(2) :129–148, 2007.

Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering : a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 593–604, New York, NY, USA, June 2007. Association for Computing Machinery.

J.J. Little and Zhe Gu. Video Retrieval by Spatial and Temporal Structure of Trajectories. *Proceedings of SPIE - The International Society for Optical Engineering*, 4315, April 2001.

Eric Hsueh-Chan Lu and Vincent S. Tseng. Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In *2009 Tenth International Conference on Mobile Data Management : Systems, Services and Middleware*, pages 273–278, May 2009.

Nehal Magdy, Mahmoud Sakr, Tamer Abdelkader, and Khaled Elbahnasy. Review on trajectory similarity measures. December 2015.

Pierre-François Marteau. Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2) :306–318, February 2009.

Maxime Masson, Cécile Cayèré, Marie-Noëlle Bessagnet, Christian Sallaberry, Philippe Roose, and Cyril Faucher. An ETL-like platform for the processing of mobility data. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, pages 547–555, New York, NY, USA, April 2022. Association for Computing Machinery. doi : 10.1145/3477314.3507057.

Lucas May Petry, Carlos Ferrero, Luis Alvares, Chiara Renso, and Vania Bogorny. Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23, June 2019.

Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. MASTER : A multiple aspect view on trajectories. *Transactions in GIS*, page tgis.12526, May 2019.

Aline Menin, Sonia Chardonnel, Paule-Annick Davoine, Michael Ortega, Etienne Duble, and Luciana Nedel. eSTIMe : une approche visuelle, interactive et modulable pour l'analyse multi-points de vue des mobilités quotidiennes. In *SAGEO Spatial Analysis and Geomatics*, Clermont-Ferrand, France, November 2019.

Mélanie Mondo. *Traces numériques et dimensions spatiales des pratiques de la ville*. PhD thesis, La Rochelle Université, 2022.

Clément Moreau, Thomas Devogele, and Laurent Etienne. Extraction de motifs de trajectoires sémantiques similaires. In *Spatial Analysis and Geomatics*, Montpellier, France, November 2018.

Tetsuya Nakamura, Keishi Taki, Hiroki Nomiya, Kazuhiro Seki, and Kuniaki Uehara. A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, 16(4) :535–548, November 2013a.

Tetsuya Nakamura, Keishi Taki, Hiroki Nomiya, Kazuhiro Seki, and Kuniaki Uehara. A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, 16(4) :535–548, November 2013b.

Paul Newson and John Krumm. Hidden Markov Map Matching Through Noise and Sparseness. December 2016.

Tales Paiva Nogueira and Hervé Martin. Qualitative Representation of Dynamic Attributes of Trajectories. In *17th AGILE Conference on Geographic Information Science*, Castellón, Spain, June 2014.

Tales Paiva Nogueira and Hervé Martin. Querying Semantic Trajectory Episodes. In *4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems (MobiGIS'15)*, Proceedings of the 4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems (MobiGIS'15), Bellevue, WA, United States, November 2015.

Tales Paiva Nogueira, Reinaldo Braga, and Hervé Martin. An Ontology-Based Approach to Represent Trajectory Characteristics. August 2014.

Tales Paiva Nogueira, Reinaldo B. Braga, Carina T. de Oliveira, and Hervé Martin. FrameSTEP : A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications*, 92 :533–545, February 2018.

Hassan Noureddine, Cyril Ray, and Christophe Claramunt. Semantic Trajectory Modelling in Indoor and Outdoor Spaces. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 131–136, June 2020.

OpenDataFrance. GéoDataMine. `https://geodatamine.fr/`.

OpenWeather. OpenWeatherMap. `https://openweathermap.org/`.

Organisation Mondiale du Tourisme. Thésaurus du Tourisme et des Loisirs. `https://vocabularyserver.com/ttla/fr/`.

Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady L. Andrienko, Natalia V. Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, José Antônio Fernandes de Macêdo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan. Semantic trajectories modeling and analysis. *CSUR*, 2013.

Petar Ristoski, Christian Bizer, and Heiko Paulheim. Mining the Web of Linked Data with RapidMiner. *Journal of Web Semantics*, 35 :142–151, December 2015.

Christian Sallaberry. *Geographical Information Retrieval in Textual Corpora*. FOCUS - Geographical Information Systems Series. Wiley-ISTE, September 2013.

Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, José Antônio Fernandes de Macêdo, Fabio Porto, and Christelle Vangenot. A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1) :126–146, April 2008.

J. Sreemathy, Infant Joseph V., S. Nisha, Chaaru Prabha I., and Gokula Priya R.M. Data Integration in ETL Using TALEND. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1444–1448, Coimbatore, India, March 2020. IEEE.

Han Su, Shuncheng Liu, Bolong Zheng, Xiaofang Zhou, and Kai Zheng. A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, 29(1) :3–32, January 2020.

Yaguang Tao, Alan Both, Rodrigo I. Silveira, Kevin Buchin, Stef Sijben, Ross S. Purves, Patrick Laube, Dongliang Peng, Kevin Toohey, and Matt Duckham. A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing*, 58(5) :643–669, July 2021.

Iraklis Varlamis, Christos Sardianos, Vania Bogorny, Luis Otávio Alvares, Jônata Tyska Carvalho, Chiara Renso, Raffaele Perego, and John Violos. A novel similarity measure for multiple aspect trajectory clustering. In Chih-Cheng Hung, Jiman Hong, Alessio Bechini, and Eunjee Song, editors, *SAC '21 : The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, pages 551–558. ACM, 2021a.

Iraklis Varlamis, Christos Sardianos, Vania Bogorny, Luis Otavio Alvares, Jônata Tyska Carvalho, Chiara Renso, Raffaele Perego, and John Violos. A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, pages 551–558, New York, NY, USA, March 2021b. Association for Computing Machinery.

T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4(1) :52–57, January 1968. ISSN 1573-8337. doi : 10.1007/BF01074755.

Michail Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. In *Proceedings 18th International Conference on Data Engineering*, pages 673–684, San Jose, CA, USA, 2002. IEEE Comput. Soc.

Haozhou Wang, Han Su, Kai Zheng, Shazia Sadiq, and Xiaofang Zhou. An effectiveness study on trajectory similarity measures. pages 13–22, January 2013.

Zhixian Yan, José Antônio Fernandes de Macêdo, Christine Parent, and Stefano Spaccapietra. Trajectory Ontologies and Queries. 2008.

Zhixian. Yan, Christine Parent, Stefano Spaccapietra, and Dipanjan Chakraborty. A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. In *The Semantic Web : Research and Applications*, Lecture Notes in Computer Science, pages 60–75, Berlin, Heidelberg, 2010. Springer.

Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. SeMiTri : A Framework for Semantic Annotation of Heterogeneous Trajectories, 2011.

Jia-Ching Ying, Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent Tseng. Mining user similarity from semantic trajectories. pages 19–26, January 2010.

Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, 5(1) : 5 :1–5 :44, February 2011.

Modélisation de trajectoires sémantiques et calcul de similarité intégrés à un ETL

Résumé : Cette dernière décennie, nous avons pu constater une montée en popularité des applications mobiles basées sur la localisation des téléphones. Ces applications collectent des traces de mobilité qui retracent le déplacement des utilisateurs au cours du temps. Dans le projet régional DA3T, nous faisons l'hypothèse que l'analyse des traces de mobilité de touristes peut aider les aménageurs dans la gestion et la valorisation des territoires touristiques. L'objectif est de concevoir des méthodes et des outils d'aide à l'analyse de ces traces. Cette thèse s'intéresse au traitement des traces de mobilité et propose une plateforme modulaire permettant de créer et d'exécuter des chaînes de traitement sur ces données. Au fil des modules d'une chaîne de traitement, la trace de mobilité brute évolue en trajectoires sémantiques. Les contributions de cette thèse sont : (i) un modèle de trajectoire sémantique multi-niveau et multi-aspect et (ii) deux mesures calculant la similarité entre deux trajectoires sémantiques s'intéressant aux dimensions spatiale, temporelle et thématique. Notre modèle (i) est utilisé comme modèle de transition entre les modules d'une chaîne de traitement. Nous l'avons mis à l'épreuve en instanciant des trajectoires sémantiques issues de différents jeux de données de domaines variés. Nos deux mesures (ii) sont intégrées à notre plateforme comme modules de traitement. Ces mesures présentent des originalités : l'une est la combinaison de sous-mesures, chacune permettant d'évaluer la similarité des trajectoires sur les trois dimensions et selon trois niveaux de granularité différents, l'autre est la combinaison de deux sous-mesures bidimensionnelles centrées autour d'une dimension en particulier. Nous avons évalué nos deux mesures en les comparant à d'autres mesures et à l'avis de géographes.

Mots clés : trace de mobilité, trajectoire sémantique, modèle, mesure de similarité, tourisme

Semantic trajectory modelling and similarity calculation integrated into an ETL

Summary : Over the last decade, we have seen a rise in popularity of mobile applications based on phone location. These applications collect mobility tracks which describe the movement of users over time. In the DA3T regional project, we hypothesise that the analysis of tourists' mobility tracks can help planners in the management and enhancement of tourist areas. The objective is to design methods and tools to help analyse these tracks. This thesis focuses on the processing of mobility tracks and proposes a modular platform for creating and executing processing chains on these data. Throughout the modules of a processing chain, the raw mobility track evolves into semantic trajectories. The contributions of this thesis are : (i) a multi-level and multi-aspect semantic trajectory model and (ii) two measures that compute the similarity between two semantic trajectories along spatial, temporal and thematic dimensions. Our model (i) is used as a transition model between modules of a processing chain. We tested it by instantiating semantic trajectories from different datasets of various domains. Our two measures (ii) are integrated in our platform as processing modules. These measures present originalities : one is the combination of sub-measures, each allowing to evaluate the similarity of trajectories on the three dimensions and according to three different levels of granularity, the other is the combination of two bidimensional sub-measures centred around a particular dimension. We evaluated our two measures by comparing them to other measures and to the opinion of geographers.

Keywords : mobility track, semantic trajectory, model, similarity measure, tourism