

Visualisation de données quantitatives géolocalisées

Laëtitia Viau

▶ **To cite this version:**

Laëtitia Viau. Visualisation de données quantitatives géolocalisées. Analyse de données, Statistiques et Probabilités [physics.data-an]. Université de Montpellier, 2023. Français. ⟨NNT : 2023UMONS034⟩. ⟨tel-04607442⟩

HAL Id: tel-04607442

https://theses.hal.science/tel-04607442v1

Submitted on 10 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En informatique

École doctorale : Information, Structures, Systèmes

Unité de recherche Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), France

Visualisation de données quantitatives géolocalisées

Présentée par Laetitia VIAU
le 20 décembre 2023

Sous la direction de Pascal PONCELET et Arnaud SALLABERRY

Devant le jury composé de

David AUBER, Professeur des universités, Université de Bordeaux	Rapporteur
Sidonie CHRISTOPHE, Directrice de recherche, Université Gustave Eiffel	Rapportrice
Isabelle MOUGENOT, Professeure des universités, Université de Montpellier	Présidente
Jean-François GIRRES, Maître de conférences, Université Paul-Valéry Montpellier 3	Examinateur
Pascal PONCELET, Professeur des universités, Université de Montpellier	Directeur
Arnaud SALLABERRY, Maître de conférences HDR, Université Paul-Valéry Montpellier 3	Co-Directeur
Nancy RODRIGUEZ, Maître de conférences, Université de Montpellier	Encadrante

UNIVERSITÉ DE MONTPELLIER

Résumé

Les données quantitatives géolocalisées, i.e. les données quantitatives associées à des entités géographiques, sont très couramment analysées pour comprendre, expliquer et prédire des phénomènes de différentes natures. La visualisation d'informations est quant à elle un support très efficace pour analyser des données. Elle permet non seulement de les synthétiser mais également de les explorer en vue de formuler de nouvelles hypothèses.

Dans ce mémoire, nous nous intéressons à la visualisation de données quantitatives associées à des entités géographiques, comme par exemple des pays, des régions... Nous distinguons deux cas de figure. Dans le premier, les données sont associées à des entités géographiques indépendantes, c'est-à-dire que chaque valeur des données ne concerne qu'une seule entité. Par exemple, le PIB ou la population des pays rentrent dans ce cas. Dans le second cas, les données sont associées à des relations entre des entités géographiques. Cette fois-ci, une valeur concerne une paire d'entités géographiques. Par exemple, les quantités de marchandises échangées entre pays rentrent dans ce cas.

Ces deux cas de figure nous ont permis de proposer trois contributions. La première concerne la conception d'une interface visuelle d'exploration de données dédiée à la comparaison, par juxtaposition, de variables géolocalisées. Sa réalisation dans le cadre d'un projet en épidémiologie a donné lieu au déploiement d'une plateforme nommée Epid Data Explorer. La seconde contribution est une étude utilisateur mise en place dans le but de comparer différentes techniques de cartographie bivariée, i.e. montrant deux variables quantitatives. Ces deux premières contributions concernent la visualisation de données associées à des entités géographiques indépendantes. La troisième contribution est l'élaboration et l'implémentation d'un algorithme de distorsion de fond de carte d'une carte de flux pour en faciliter la lecture. Elle porte donc sur la visualisation de données associées à des relations entre entités géographiques.

Mots-clés - Données quantitatives, visualisation d'information, géovisualisation, sémiologie graphique, cartographie

Abstract

Geographically referenced quantitative data, i.e., quantitative data associated with geographical entities, are commonly analyzed to understand, explain, and predict various phenomena. Information visualization, on the other hand, is a highly effective tool for data analysis. It not only helps in summarizing data but also facilitates exploration with the aim of generating new hypotheses.

In this thesis, we focus on the visualization of quantitative data associated with geographical entities, such as countries, regions, ... We distinguish between two scenarios. In the first scenario, data is linked to independent geographical entities, meaning that each data value pertains to a single entity. For example, GDP or population data falls into this category. In the second scenario, data is associated with relationships between geographical entities. In this case, a value pertains to a pair of geographical entities. For instance, quantities of traded goods between countries fit into this category.

These two scenarios have allowed us to make three contributions. The first one concerns the design of a dedicated data exploration visual interface for the comparison, by juxtaposition, of geolocated variables. Its implementation in the context of an epidemiology project led to the deployment of a platform called Epid Data Explorer. The second contribution is a user study established to compare various bivariate mapping techniques, i.e., those displaying two quantitative variables. These first two contributions pertain to the visualization of data associated with independent geographical entities. The third contribution is the development and implementation of a map background distortion algorithm for a flow map to facilitate its readability. Thus, it focuses on the visualization of data associated with relationships between geographical entities.

Keywords - Quantitative data, information visualization, geovisualization, semiology of graphics, cartography.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au bon déroulement de cette thèse et à la rédaction de ce mémoire.

Tout d'abord, je souhaite remercier les membres du jury. Sidonie Christophe et David Auber pour leur lecture du manuscrit, évaluation de mon travail et pour toutes leurs remarques pertinentes. Merci également à Isabelle Mougenot et Jean-François Girres d'avoir accepté de faire partie des membres du jury.

Je tiens à exprimer toute ma reconnaissance auprès de mes encadrants Pascal Poncelet, Nancy Rodriguez et Arnaud Sallaberry. Je les remercie pour tout ce temps qu'ils m'ont régulièrement accordé, leurs conseils ainsi que leur aide ayant grandement contribué à l'avancement de mes travaux. Ils ont fait preuve de beaucoup de patience et ont toujours eu confiance en moi. Je leur adresse un grand merci pour leur encadrement et soutien.

Je remercie également tous les membres de mon équipe, permanents, post-doctorants, doctorants et stagiaires pour tous ces échanges scientifiques et leurs conseils mais également pour tous ces moments plus conviviaux passés ensemble.

Un grand merci également à ma mentore, Christine Coudray, d'avoir participé à un programme de mentorat à mes côtés. Je la remercie d'avoir parfaitement rempli ses engagements, sa vision externe m'a beaucoup aidé et a été précieuse dans le bon déroulement de cette thèse.

Enfin, je remercie mes amis qui ont toujours été présents ainsi que les membres de mon club de Taekwondo. Chacun m'a soutenu d'une manière ou d'une autre. Un grand merci aux membres de ma famille pour leur soutien inestimable depuis toujours, leurs encouragements, et d'être fières de la personne que je suis aujourd'hui. Pour terminer, je remercie bien évidemment la personne avec qui je partage ma vie. Merci de m'avoir encouragé à me lancer dans l'aventure d'une thèse, de m'avoir soutenue à toute épreuve et d'avoir toujours cru en moi.

TABLE DES MATIÈRES

1 Introduction **1**
- 1.1 Contexte . . . 1
- 1.2 Contributions . . . 3
- 1.3 Organisation du mémoire . . . 5

2 Données associées à des entités géographiques **7**
- 2.1 Introduction . . . 9
- 2.2 État de l'art . . . 9
  - 2.2.1 Visualisation . . . 10
  - 2.2.2 Comparaison . . . 21
- 2.3 Comparaison par juxtaposition . . . 26
  - 2.3.1 Approche proposée . . . 27
  - 2.3.2 Application avec Epid Data Explorer . . . 31
- 2.4 Comparaison par superposition . . . 40
  - 2.4.1 Comparer des représentations cartographiques bivariées . . . 40
  - 2.4.2 Présentation de l'expérimentation . . . 45
  - 2.4.3 Résultats de l'expérimentation . . . 50
- 2.5 Discussion . . . 58
- 2.6 Conclusion . . . 60

3 Données associées à des relations entre des entités géographiques **61**
- 3.1 Introduction . . . 62

3.2 État de l'art 63
3.2.1 Visualisation 64
3.2.2 Cartes de flux 67
3.2.3 Distorsion de carte 72
3.3 Algorithme de distorsion FMD (Flow Map Distortion) 74
3.3.1 Première distortion 75
3.3.2 Courbes 79
3.3.3 Deuxième distorsion 79
3.4 Création de carte de flux un-à-plusieurs 81
3.4.1 Création du graphe initial 81
3.4.2 Définition de l'arbre 83
3.4.3 Simplification de l'arbre 83
3.4.4 Dessin des flux 86
3.5 Discussion 87
3.5.1 Jeux de données 87
3.5.2 Temps de calcul 87
3.5.3 Résultats 88
3.6 Conclusion 93

4 Conclusion et perspectives **95**
4.1 Synthèse des contributions 95
4.2 Perspectives 96
4.2.1 Conception de nouvelles approches de comparaison . . . 96
4.2.2 Visualisation de flux par distorsion de fond de carte . . . 98
4.2.3 Comparaison des flux à partir d'une même carte de flux . 99

CHAPITRE

1

INTRODUCTION

La visualisation de données consiste à représenter visuellement des données afin de mieux les comprendre et les appréhender. Une donnée est, par définition, une représentation conventionnelle d'une information. Il existe une grande variété de types de données (données textuelles, géographiques, qualitatives, quantitatives, relationnelles...). Dans cette thèse, nous nous intéressons particulièrement à deux d'entre elles : les données quantitatives et les données géographiques. Les données quantitatives renvoient à des valeurs numériques. Les données géographiques se présentent sous la forme d'un ensemble de points, polylignes ou polygones qui représentent des entités géographiques, i.e. des zones délimitées à la surface de la Terre (des routes, des communes, des départements, des régions, des continents...). Ces deux types de données peuvent se combiner dans des jeux de données contenant des variables quantitatives associées à des entités géographiques, ou associées à des relations entre plusieurs entités. L'objectif de cette thèse est d'étudier les visualisations qui permettent de les explorer et les analyser en vue d'en extraire des connaissances.

Cette thèse s'est déroulée au sein de l'équipe ADVANSE [1], ADVanced Analytics for data SciencE, au LIRMM [2], Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier. Elle a été effectuée dans le cadre du projet européen MOOD [3], MOnitoring Outbreak events for Disease surveillance in a data science context.

1.1 Contexte

La cartographie constitue le moyen privilégié pour la représentation visuelle de données quantitatives liées à des entités géographiques. Une carte est définie comme étant *une image simplifiée et codifiée de l'espace géographique, qui représente ses caractéristiques et/ou son organisation* [55]. Une des plus anciennes, représentant un plan de ville, est la peinture murale de Çatal Huyük, en Turquie, vers 6200 av. J-C. Le cadastre de Bedolina, une gravure sur roche

1. https://www.lirmm.fr/equipes/advanse/
2. https://www.lirmm.fr/
3. https://mood-h2020.eu/

en Italie actuelle, représente des routes, des maisons et des champs mais également des hommes et des animaux domestiques. Cette carte daterait de l'âge du Bronze ancien, vers 2000 av. J-C. Une première carte du monde est proposée, vers 550 av. J-C, par Anaximandre et Hécatée. Cette carte est de forme circulaire centrée sur la Méditerranée et la terre est entourée d'eau. Durant l'époque romaine, les romains auraient élaboré des cartes routières pour leur itinéraires terrestres. La table de Peutinger, une copie de la carte originale recouvrant tout l'empire romain et les conquêtes d'Alexandre le Grand, illustre comment les romains représentaient le monde. Elle est considérée comme l'ancêtre des cartes routières. À la suite de ces premières cartes, du Moyen-Age à notre époque, les cartes n'ont cessé d'évoluer et font partie intégrante de l'histoire de l'humanité.

Les premières cartes représentant des données quantitatives remontent, quant à elles, qu'au XIX^{ème} siècle. On peut par exemple citer Charles François Dupin (1784-1873) et Charles Joseph Minard (1781-1870). Le premier, Dupin, est l'auteur d'une des toutes premières *cartes statistiques modernes* en 1826 (cf. Figure 1.1). Plus précisément, cette carte représente des données quantitatives associées à des entités géographiques. Les données représentées concernent le taux de scolarisation des garçons dans chaque département de la France métropolitaine. Afin d'illustrer le niveau d'éducation en France, Dupin représente les valeurs à partir de la technique choroplèthe : les départements ayant un faible taux ont une couleur proche du blanc, et plus la valeur est élevée plus on se rapproche du noir. Le second, Minard, est quant

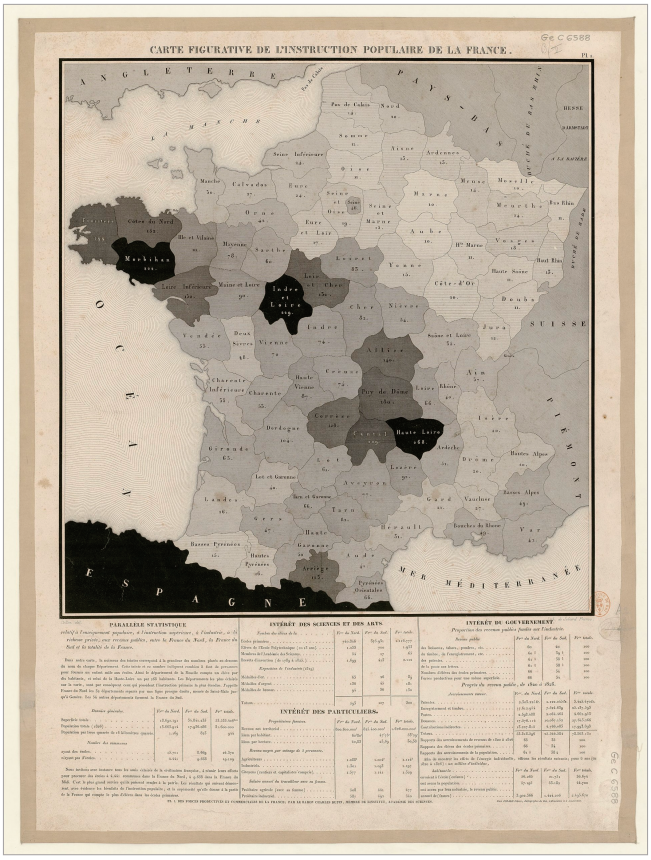


FIGURE 1.1 – Carte choroplèthe conçue en 1826 par Charles Dupin. Le taux de scolarisation des garçons est représenté dans chaque département. (source : site de la Bibliothèque nationale de France (BNF) [dernière consultation le 04/08/2023])

à lui particulièrement connu pour ses représentations visuelles grâce à la qualité de leur conception et leur précision. Elles sont d'ailleurs considérées comme des chefs d'œuvre de l'âge d'or de la représentation visuelle statistique (seconde moitié du XIX^{ème} siècle) comme

l'indique Sandra Rendgen dans son ouvrage dédié à ses représentations [75]. Une de ses cartes les plus connues est la carte de flux qu'il a conçu en 1869 pour représenter le nombre d'hommes dans l'armée française durant la campagne de Russie de 1812 (cf. Figure 1.2). Cette dernière, contrairement à la carte de Dupin, ne représente pas des données quantitatives associées à des entités géographiques mais à des relations entre des entités géographiques.

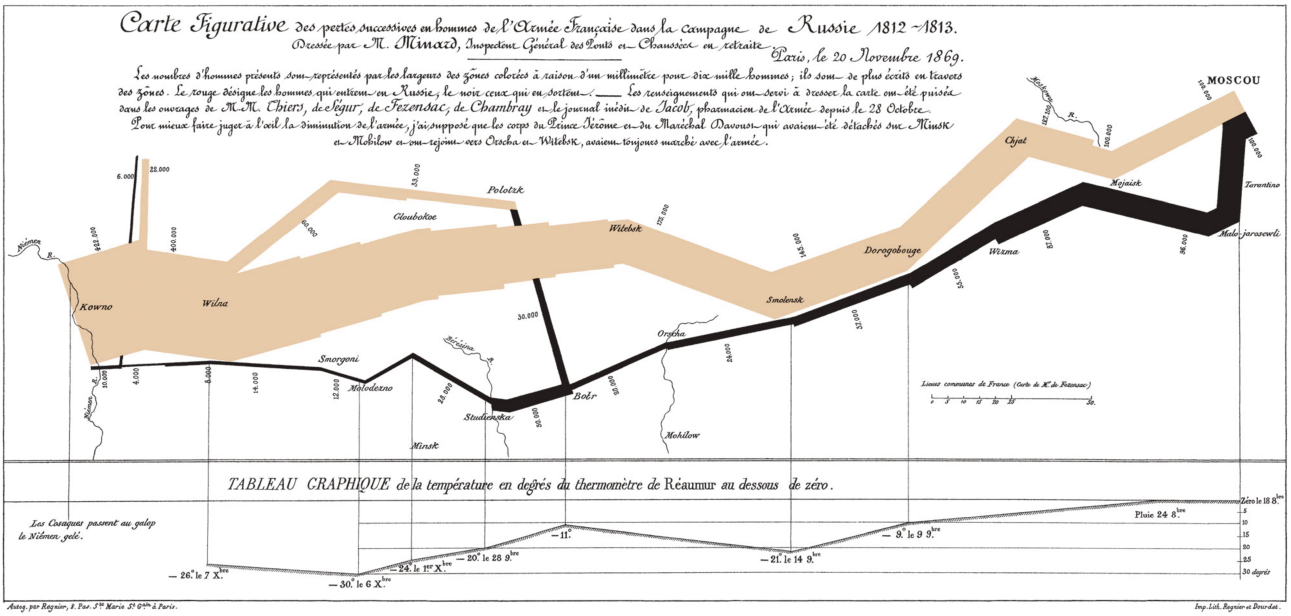


FIGURE 1.2 – Carte de Charles Joseph Minard représentant le nombre d'hommes dans l'armée française durant la campagne de Russie en 1812 [75].

L'émergence des ordinateurs et des premiers concepts de géographie quantitative et informatique dans les années 1960 ont donnée naissance aux premiers systèmes d'information géographique (SIG) ; les premiers logiciels informatiques de cartographie ont alors été créés et améliorés [28]. Avec les progrès de l'informatique et des SIG, les visualisations de données quantitatives sont, par la suite, devenues de plus en plus variées, élaborées, sophistiquées et détaillées. Face à cette variété et à cette complexité, et au regard des nouvelles possibilités offertes par les technologies actuelles, notre but est d'étudier et de comparer les techniques de visualisation de données liées à des entités géographiques existantes, mais également de proposer de nouvelles approches et de nouveaux algorithmes.

Comme illustré par les deux cartes présentées précédemment citées (cf. Figure 1.1 et Figure 1.2), nous distinguerons tout au long de ce mémoire deux types de cartographies de données quantitatives : (1) les visualisations de données associées à des entités géographiques, comme celle de Dupin, et (2) les visualisations de données associées à des relations entre des entités géographiques, comme celle de Minard.

1.2 Contributions

Pour traiter la question de représentations de données associées à des entités géographiques indépendantes ou en relation entre elles, nos travaux se traduisent par trois axes majeurs : (1) développement d'une approche pour faire des comparaisons par juxtaposition, (2) étude utilisateur comparant différentes techniques de cartographie bivariée, (3) élaboration et implémentation d'un algorithme de distorsion de formes géographiques pour

améliorer la lisibilité des cartes de flux.

Dans le cadre du projet MOOD, il était nécessaire de développer un outil de visualisation de données épidémiologiques. Les besoins étaient nombreux : les utilisateurs finaux voulaient non seulement voir et explorer les données, mais aussi les enregistrer, les télécharger, les requêter, les comparer et les protéger. Pour la comparaison, nous avons proposé une approche par juxtaposition de deux cartes. Cette approche est basée sur un système de synchronisation entre ces deux cartes au niveaux de la date (ou de la période), de la variable et de la zone géographique. Elle a été implémentée pour fournir une plateforme nommée *Epid Data Explorer* [4]. La Figure 1.3 donne un aperçu de l'interface.

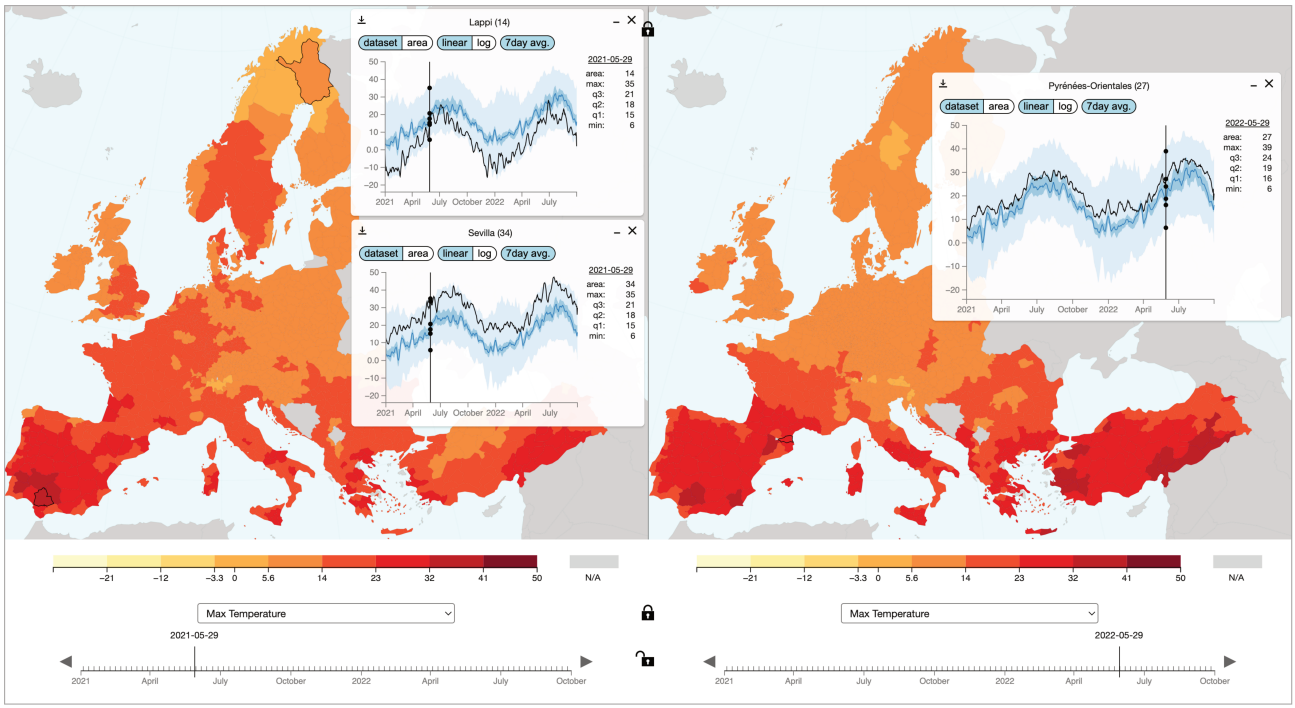


FIGURE 1.3 – Aperçu de l'approche de comparaison par juxtaposition basée sur un système de synchronisation entre les deux vues que l'on propose.

Suite à ces travaux, nous nous sommes penchés sur la comparaison de variables quantitatives par superposition. Nous avons comparé différentes techniques de conception cartographique bivariée. Autrement dit, nous avons évalué différentes cartes représentant chacune simultanément deux variables à comparer. Afin de procéder à cette évaluation, nous avons mis en place une expérimentation. Les participants n'avaient pas de connaissance approfondies en cartographie afin d'éviter au maximum des disparités d'évaluations (certaines cartes sont plus fréquemment utilisées que d'autres). Nous avons sélectionné 5 types de cartes bivariées facilement reproductibles pour notre étude. Un exemple pour chacune d'entre elles est présenté dans la Figure 1.4.

Contrairement aux deux précédentes qui portent sur des données quantitatives associées à des entités géographiques, la dernière contribution porte sur des données associées à des relations entre entités. La technique de visualisation la plus couramment utilisée pour cela est la carte de flux. Nous avons cependant pu constater que, malgré leur ancienneté et les progrès numériques, les cartes de flux actuelles peuvent présenter des problèmes d'encombrement visuel. Différentes solutions sont proposées dans la littérature pour y remédier mais les liens entre les entités géographiques, dont l'épaisseur varie selon la valeur associée,

4. https://advanse.lirmm.net/EDE/epid-data-explorer/

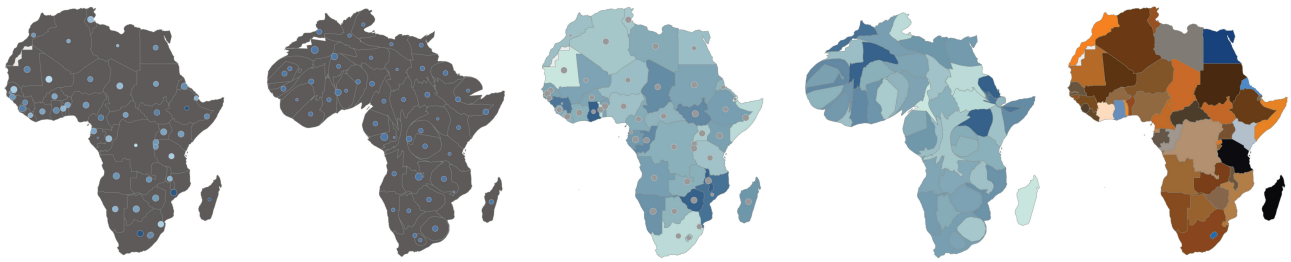
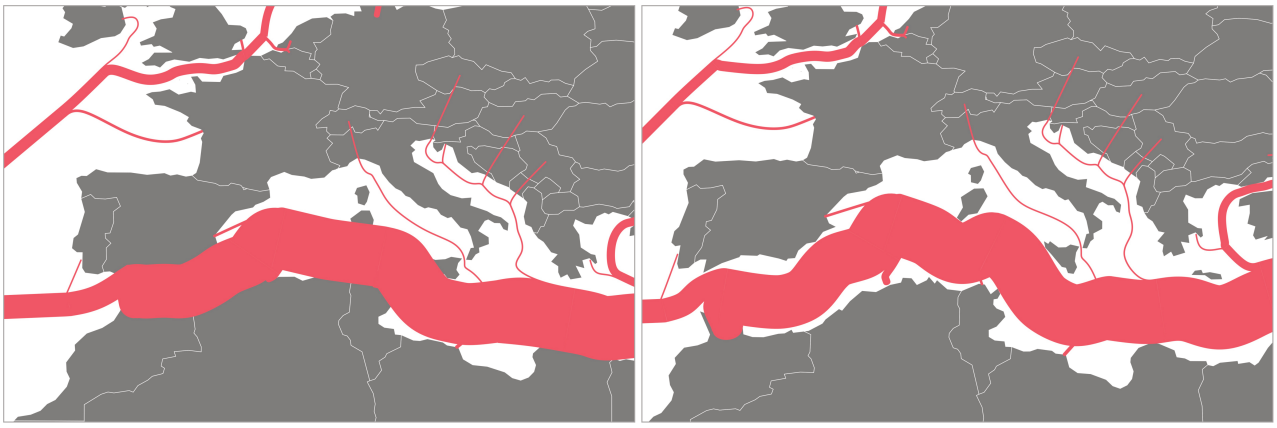


FIGURE 1.4 – Cartes bivariées que nous avons évalué. Un exemple pour chaque technique est proposé à partir d'un même jeu de données.

peuvent générer des problèmes d'occlusion visuelle. Pour répondre à cette problématique, nous nous sommes inspirés des cartes de flux de Minard. En effet, dans ses cartes, on ne rencontre aucune superposition car les entités géographiques sont déformées de façon à ce que les liens ne viennent pas les cacher. L'objectif a alors été de faire de même en automatisant cette idée. Nous avons ainsi proposé et implémenté un algorithme de distorsion de cartes. La Figure 1.5 montre un exemple de résultat avant et après déformation obtenu avec notre solution.



(a) Sans déformation, les entités géographiques sont occultées.

(b) Avec déformation, toutes les superpositions ont été supprimées.

FIGURE 1.5 – Carte de flux obtenue à partir de l'algorithme que nous proposons : aperçu sans et avec déformation des entités géographiques.

1.3 Organisation du mémoire

Le mémoire est organisé de la manière suivante. Les travaux relatifs aux approches de comparaison de données associées à des entités géographiques sont décrits dans le Chapitre 2. L'approche permettant de créer des cartes de flux avec distorsion du fond de carte est exposée dans le Chapitre 3. Ces deux chapitres sont organisés de manière similaire : tout d'abord la problématique est présentée puis un état de l'art est rapporté. Ensuite, les contributions concernées sont décrites et, pour clôturer le chapitre, nous proposons une discussion avant de conclure. Pour terminer ce mémoire, nous concluons et énonçons les perspectives associées aux travaux de recherche dans le Chapitre 4.

CHAPITRE

2

DONNÉES ASSOCIÉES À DES ENTITÉS GÉOGRAPHIQUES

Sommaire

2.1 Introduction . . . 9
2.2 État de l'art . . . 9
2.2.1 Visualisation . . . 10
2.2.1.1 Historique . . . 10
2.2.1.2 Sémiologie graphique . . . 12
2.2.1.3 Types de cartes . . . 13
2.2.1.4 Perception graphique et représentation cartographique . . . 16
2.2.2 Comparaison . . . 21
2.2.2.1 Juxtaposition . . . 21
2.2.2.2 Superposition . . . 23
2.3 Comparaison par juxtaposition . . . 26
2.3.1 Approche proposée . . . 27
2.3.1.1 Besoins . . . 27
2.3.1.2 Visualisation avec deux vues . . . 28
2.3.1.3 Tooltips . . . 29
2.3.2 Application avec Epid Data Explorer . . . 31
2.3.2.1 Contexte et besoins additionnels . . . 32
2.3.2.2 Présentation . . . 33
2.3.2.3 Cas d'usage . . . 37
2.4 Comparaison par superposition . . . 40
2.4.1 Comparer des représentations cartographiques bivariées . . . 40
2.4.1.1 Techniques de visualisation . . . 42
2.4.1.2 Tâches . . . 44

2.4.2 Présentation de l'expérimentation 45
2.4.2.1 Jeux de données 45
2.4.2.2 Matériel et participants 45
2.4.2.3 Procédure 46
2.4.2.4 Hypothèses 49
2.4.3 Résultats de l'expérimentation 50
2.4.3.1 Précision des réponses 50
2.4.3.2 Temps de réponse 54
2.4.3.3 Retours des participants 54
2.4.3.4 Discussion sur les résultats 56
2.5 Discussion 58
2.6 Conclusion 60

2.1 Introduction

La visualisation de données quantitatives associées à des entités géographiques est une étape non négligeable en analyse de données. En effet, prendre connaissance des données permet de formuler des hypothèses mais également de définir la méthode d'analyse de données la plus adaptée. Pour cela, la visualisation est une alliée de taille, malheureusement encore bien trop souvent sous-estimée. Confronter les données, les comparer et observer d'éventuelles corrélations doit être possible grâce à la visualisation. Cette dernière doit également tirer parti des informations spatiales des données.

La cartographie est alors parfaitement adaptée à cette problématique. En plus de proposer en une image une synthèse des données, elle montre leur contexte géographique et permet d'observer des phénomènes pouvant être en lien avec la localisation. Cependant, la comparaison de deux cartes présentant chacune une variable différente peut très vite devenir complexe si elles sont affichées de manière indépendante, dans des fenêtres distinctes. Concevoir et élaborer des outils de visualisation permettant de faire des comparaisons de manière intuitive constitue la problématique de ce chapitre. Deux approches majeures sont envisageables : la première consiste à juxtaposer les cartes, comme illustré dans la Figure 2.1a, tandis que la seconde implique une superposition, comme le montre la Figure 2.1b. La principale différence entre ces deux approches réside dans le fond de carte :

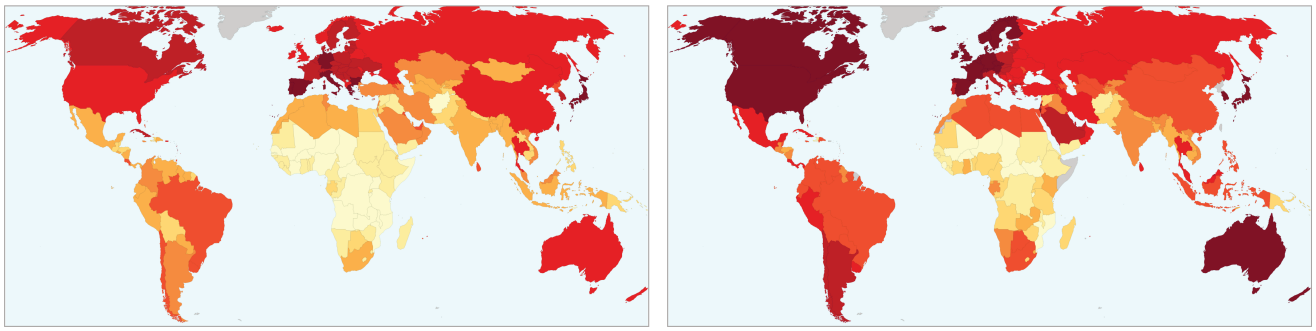
- Dans le cas de la juxtaposition, on dispose de différents fonds de carte, chacun correspondant à une variable distincte, placés côte à côte.
- Dans le cas de la superposition, un seul fond de carte est utilisé pour toutes les variables.

Les travaux abordés dans ce chapitre se déclinent en deux aspects : (1) la conception et le développement d'une plateforme interactive basée sur la juxtaposition, et (2) l'expérimentation visant à comparer différentes techniques de conception de cartes par superposition.

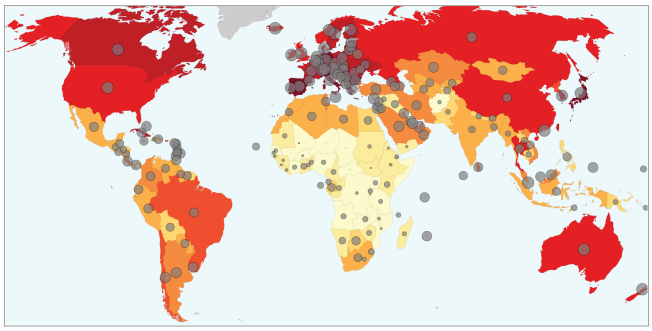
Ce chapitre est organisé de la manière suivante : nous débutons avec un état de l'art sur la visualisation de données quantitatives associées à des entités géographiques et les approches de comparaisons existantes dans la Section 2.2. L'approche de comparaison par juxtaposition que nous proposons est ensuite exposée dans la Section 2.3. Ensuite, nous présentons l'étude utilisateur menée pour comparer différents types de cartes par superposition dans la Section 2.4. Une discussion à propos des travaux menés est soumise dans la Section 2.5 avant de conclure dans la Section 2.6.

2.2 État de l'art

Étant donné que ce chapitre se consacre à la visualisation de données quantitatives associées à des entités géographiques, les types de visualisations appropriés sont les cartes thématiques. Une carte thématique est une représentation cartographique qui affiche des informations spécifiques. Elle revêt une grande utilité dans le processus de prise de décision et doit être conçue de manière à être aisément compréhensible pour être accessible à un large public. Il en existe différents types. C'est pourquoi, comprendre les avantages et les inconvénients de chacune est primordial avant d'envisager une quelconque comparaison. En effet, différentes informations ont souvent besoin d'être confrontées entre elles, dans ce cas plusieurs cartes thématiques entrent en jeu. Pour les comparer nous avons deux possibilités : les juxtaposer ou les superposer.



(a) Comparaison par juxtaposition : deux cartes sont positionnées l'une à côté de l'autre. Les deux cartes ont la même technique de représentation (cartes choroplèthes). L'âge médian de la population est représenté sur la carte de gauche. La carte de droite correspond à l'indice de développement humain.



(b) Comparaison par superposition : le fond de carte est le même mais les techniques de représentation sont différentes (carte choroplèthe pour l'âge médian et carte à symboles proportionnels pour l'indice de développement humain).

FIGURE 2.1 – Deux types de comparaisons : juxtaposition et superposition. Les deux visualisations représentent les mêmes données (âge médian de la population et indice de développement humain).

Nous étudierons donc les méthodes de conception de cartes thématiques pour visualiser des données quantitatives et passerons ensuite à la présentation des approches majeures pour comparer ces visualisations.

2.2.1 Visualisation

Cette première partie concerne les principales cartes thématiques pour visualiser des données quantitatives associées à des entités géographiques. Dans un premier temps, nous proposons un bref historique sur les cartes représentant des données quantitatives. Nous verrons ensuite la sémiologie graphique constituant les règles de base en conception cartographique, puis les types de cartes qui en découlent. Nous finirons avec une discussion à propos des travaux menés sur la perception graphique et la configuration des cartes.

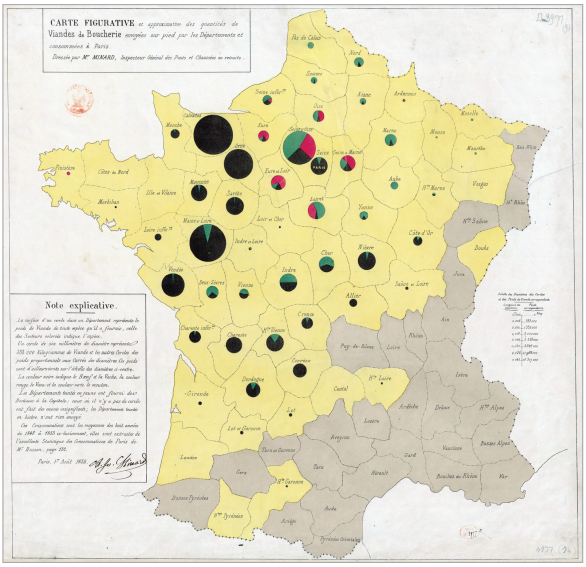
2.2.1.1 Historique

Les cartes utilisées comme communication visuelle sur des données quantitatives ont une histoire, dont le début est situé au XIX^{ème} siècle, difficile à appréhender, il s'agit d'un sujet bien plus complexe et riche que ce que nous pouvons penser. Les fonctions cartographiques de représentation générale, topographique ou hydrographique auraient été prééminentes jusqu'au XVIIIème siècle comme le précise Gilles Palsky dans son ouvrage *Des chiffres*

et des cartes : naissance et développement de la cartographie quantitative française au XIXe siècle [72]. Les valeurs quantitatives, ou plutôt les chiffres si on reprend la notion utilisée auparavant, auraient été intégrées aux cartes à partir du XIXème siècle. Cette évolution a, d'ailleurs, été désignée comme une *"langue nouvelle qui donne aux chiffres la vie qui leur manque"* par Louis Léger Vauthier, ingénieur en chef des ponts et chaussées, en 1890. Charles Dupin (1784-1673) est l'auteur d'une des toutes premières *cartes statistiques modernes* en 1826. Parmi les premiers développements de la cartographie statistique, nous pouvons aussi citer Henry Drury Harness (cf. Figure 2.2a), qui avait élaboré ses cartes afin d'aider la Commission Irlandaise des chemins de fer en 1837. Les ingénieurs des Ponts et Chaussées pour l'implantation des premières lignes de fer en France ont aussi participé au développement de la cartographie de valeurs numériques. Plus connu, Charles Joseph Minard, un ingénieur français, a conçu de nombreuses visualisations statistiques dont justement des cartes figuratives [75] comme celle présentée à la Figure 2.2b.



(a) Carte élaborée par Henry Drury Harness en 1837 représentant les densités de population en Irlande. (source : Atlas accompagnant le 2e rapport des commissaires aux chemins de fer [dernière consultation le 04/08/2023])



(b) Carte figurative de Minard en 1858 [75] : la taille d'un cercle correspond au poids de viande qu'un département a fourni et la couleur représente l'espèce.

FIGURE 2.2 – Premières cartes thématiques représentant des "chiffres", valeurs quantitatives.

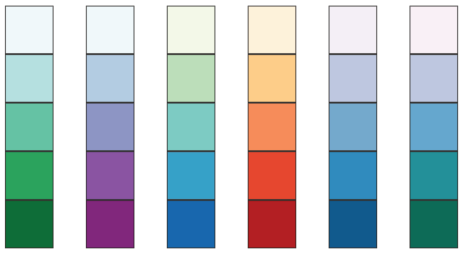
Le terme *carte thématique* s'est largement diffusé en France durant les années 1960. L'émergence de l'informatique a révolutionné un bien grand nombre de domaines dont la cartographie avec l'émergence des Systèmes d'Information Géographique (SIG). Un de ces précurseurs est le géographe britannique Roger Tomlinson. Au fil des années, les SIG ont commencé à être commercialisés, pour finalement devenir largement accessibles. Aujourd'hui, la création d'une carte est à la portée de n'importe quel utilisateur grâce à divers outils. Elle peut même être automatisée, comme l'a souligné Anne Le Fur [56], qui a déclaré que *"la carte n'est plus un but mais un outil d'exploration des hypothèses et des données"*. En effet, élaborer des cartes thématiques de données quantitatives permet désormais de prendre connaissance des

données en amont afin d'émettre des hypothèses et effectuer des analyses de données plus approfondies.

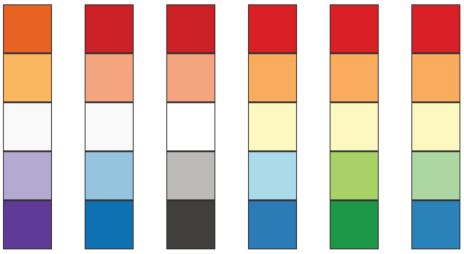
2.2.1.2 Sémiologie graphique

La représentation cartographique est la transcription d'une information complexe en une image graphique à l'aide de signes. La compréhension de l'image cartographique repose sur des règles de construction de la symbolique : c'est la sémiologie.

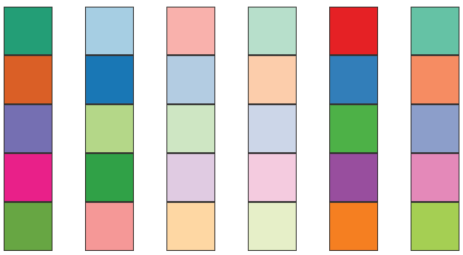
La sémiologie graphique a été introduite par Jacques Bertin en 1967 [7] pour représenter l'information selon un ensemble de codes et de règles. Cette sémiologie graphique a depuis largement été utilisée pour créer des cartes. Parmi les paramètres fondamentaux à identifier figure le choix de l'implantation et de la variable visuelle. Les variables visuelles, définies par Bertin et désignées comme "variables rétiniennes", sont les suivantes : taille, valeur, grain, couleur, orientation et la forme. Dans le cadre de la visualisation de données quantitatives, seules la taille et la valeur sont adaptées. La variable visuelle valeur renvoyait initialement uniquement à des variations de luminosité ■■■■■■ ou de saturation ■■■■■■ pour une même couleur. De nos jours, cette variable visuelle renvoie également à des variations de teintes qui restent tout de même légères ; comme par exemple du jaune au rouge (cf. Figure 2.3a). Dans le cas d'une échelle divergente, c'est-à-dire où il y a une valeur pivot, les variations de teintes peuvent être plus conséquentes (Figure 2.3b). Pour une échelle divergente on peut prendre comme exemple les valeurs du pH avec la valeur 7 et où les valeurs inférieures à cette valeur pivot correspondent à l'acidité. À l'inverse, les valeurs supérieures à 7 qui tendent vers 14 sont de plus en plus basiques. La distinction entre la variable visuelle valeur et la variable couleur est que la variable visuelle couleur consiste à utiliser différentes couleurs n'ayant aucun ordre de grandeur (cf. Figure 2.3c).



(a) Échelles de couleurs pour visualiser des données quantitatives séquentielles.



(b) Échelles de couleurs pour visualiser des données quantitatives divergentes.

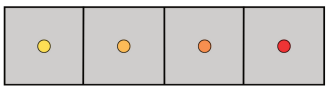
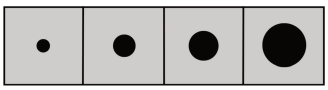
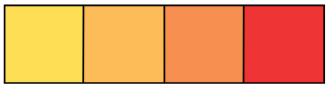
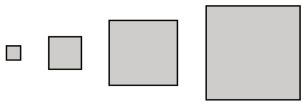


(c) Échelles de couleurs pour visualiser des données qualitatives.

FIGURE 2.3 – Exemples d'échelles de couleurs issues de ColorBrewer [42] (https://colorbrewer2.org/ [dernière consultation le 04/08/2023]).

Pour résumer, la variable visuelle couleur ■ ■ ■ ■ ■ ■ est utilisée pour des données qualitatives ; nous nous concentrons donc sur la variable visuelle valeur ■ ■ ■ ■ ■ ■ et la taille • • ● ● puisqu'elles sont les plus adaptées à la représentation de données quantitatives. Pour rappel, ce qui nous intéresse dans ce chapitre est la visualisation de données quantitatives **sur** des entités géographiques, l'implantation concernée est donc uniquement ponctuelle et zonale. Il existe cependant des implantations linéaires que nous traiterons ultérieurement pour la visualisation de données associées à des relations **entre** des entités géographiques (cf. Chapitre 3). Le Tableau 2.1 résume les différentes combinaisons possibles selon l'implantation et la variable visuelle utilisée. Pour fluidifier la suite de la lecture, maintenant que la distinction entre la variable valeur et la variable couleur a bien été formulée, nous utiliserons le terme couleur pour désigner la variable visuelle valeur. À partir des implantations ponctuelles et zonales, ainsi que des variables visuelles telles que la couleur et la taille, on peut alors distinguer trois principaux types de cartes : les cartes à symboles (*Sy*), les cartes choroplèthes (*Ch*) et les cartogrammes (*Ca*).

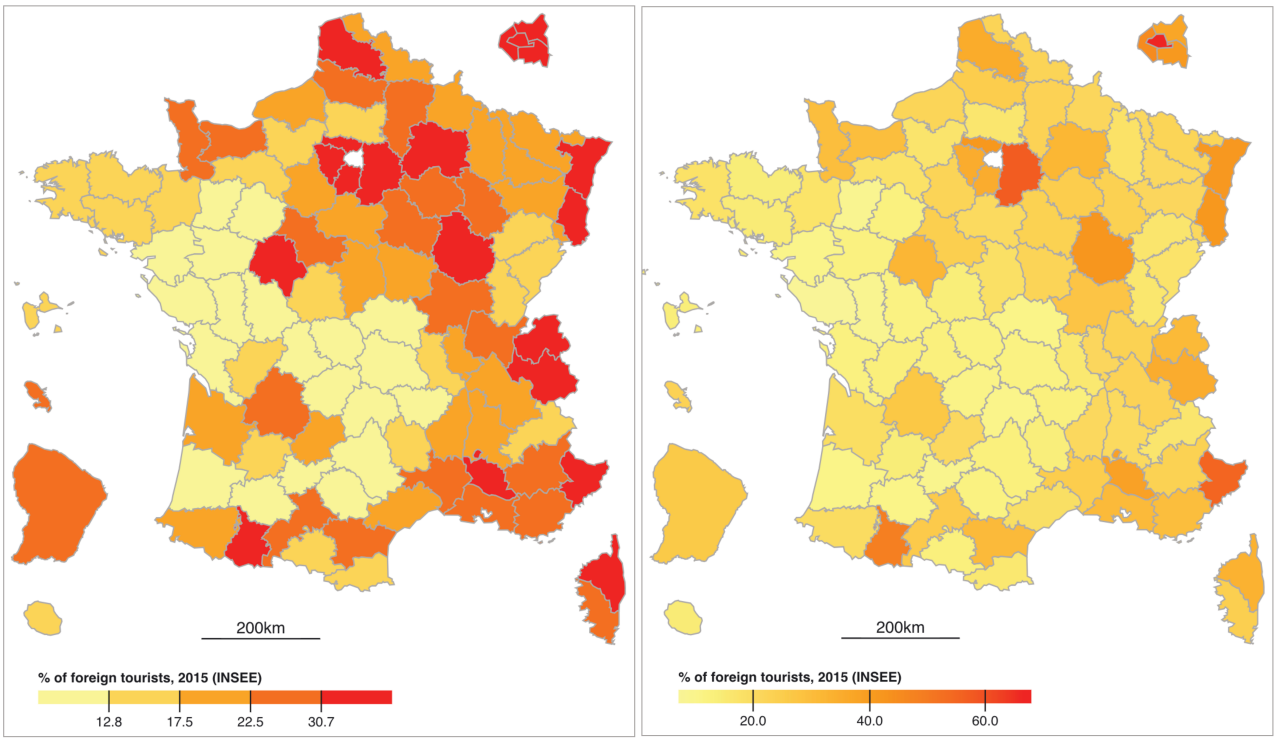
Tableau 2.1 – Différents types de cartes possibles selon les implantations et les variables visuelles définies par Bertin pour visualiser des données quantitatives.

	Couleur	**Taille**
Ponctuelle	 Carte à symboles (*Sy*)	 Carte à symboles (*Sy*)
Zonale	 Carte choroplèthe (*Ch*)	 Cartogramme (*Ca*)

2.2.1.3 Types de cartes

La carte choroplèthe est un type de carte thématique consistant à représenter des valeurs par des couleurs de remplissage à l'intérieur des entités géographiques comme illustré par la Figure 2.6c. Il est également possible de trouver des motifs tels que des hachures (Slocum et al. [84]). Torguson [93] affirme qu'elles sont très simples à créer, particulièrement à partir d'un SIG et très aisément compréhensibles. Une des premières cartes représentant des valeurs quantitatives était d'ailleurs une carte choroplèthe Ces cartes nommées initialement *cartes teintées* ont rapidement été diffusées et ont gagné en popularité. Ce n'est que dans les années 1930 que le terme carte choroplèthe a été introduit. On distingue les choroplèthes classées et les non classées. On parle de carte choroplèthe classée si les données sont regroupées en classes pour simplifier le modèle de distribution (cf. Figure 2.4a). Les techniques de classification telles que les intervalles égaux ou bien les quantiles produisent des représentations différentes et parfois même trompeuses comme l'indique Kelly [52]. Tobler a introduit les cartes choroplèthes non classées [90] qu'il détermine comme des représentations plus précises des donnés originales (cf. Figure 2.4b).

Cependant, le regroupement des données en classes reste courant. Il a été démontré que les logiciels SIG fournissent des schémas de couleurs pour lesquels il est impossible de créer une carte choroplèthe réellement non classée mais le nombre de classes peut être suffisamment conséquent pour s'en rapprocher : jusqu'à 999 avec QGIS et 1000 dans ArcMap (Kelly [52]).



(a) Carte choroplèthe avec 5 classes.

(b) Carte choroplèthe continue, non classée.

FIGURE 2.4 – Cartes choroplèthes représentant les mêmes données, issues de l'INSEE sur le pourcentage de touristes étrangers en France (https://observablehq.com/@ljegou/ttt-continuous-choropleth?collection=@ljegou/ttt#cell-312 [dernière consultation le 04/08/2023]).

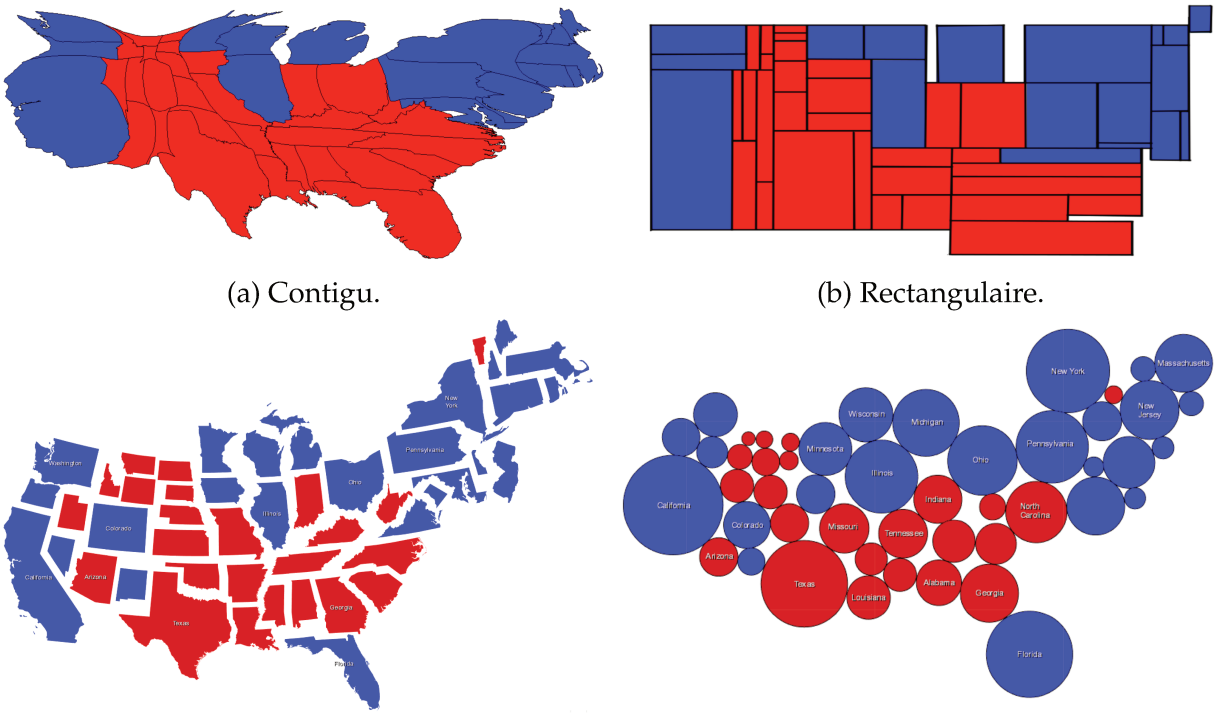
Concernant les cartes à symboles, dans le Tableau 2.1, les symboles sont sous forme de cercle mais il est possible d'utiliser bien d'autres formes en 2D ou même en 3D telles que des sphères ou bien des cubes par exemple. Un ensemble d'exemples des nombreuses formes de symboles ponctuels a été présenté par Brewer et Campbell [9]. Les cartes à symboles proportionnels, c'est-à-dire les cartes où l'implantation est ponctuelle et la variable visuelle est la taille, sont utilisées depuis bien longtemps et font parties des premières cartes à représenter des valeurs quantitatives avec les cartes choroplèthes. En effet, Minard avait utilisé, en 1858, cette méthode pour représenter les quantités de viandes fournies par les départements français [75] (cf. Figure 2.2b). Comme on peut le voir sur la Figure 2.6b, les cartes à symboles proportionnels présentent parfois des chevauchements. Cela se produit souvent dans le cas où plusieurs symboles sont grands et/ou lorsque la densité est conséquente. Si les chevauchements sont importants, la lecture de la carte peut alors être erronée et les entités géographiques non visibles. La solution serait de limiter la taille maximale mais si celle-ci est trop réduite alors le jugement de la taille devient trop complexe. L'échelle doit donc être suffisamment grande mais ne doit pas produire des chevauchements rendant la carte illisible (Cabello et al. [13]).

Les cartogrammes, quant à aux, comme définis par Nusrat et Kobourov [70], combinent des informations statistiques et géographiques dans des cartes thématiques où les entités géographiques sont mises à l'échelle proportionnellement à certaines valeurs comme par exemple la population. Autrement dit, les cartogrammes consistent à modifier les surfaces des entités géographiques afin qu'elles correspondent à la valeur associée à l'entité. Initialement, le terme cartogramme était utilisé pour désigner différents types de cartes tels que les cartes choroplèthes comme l'indique Waldo Tobler [92]. Bertin désignait quant à lui les

cartogrammes tels que nous les connaissons actuellement sous le nom d'anamorphose cartographique [7]. Il utilisait le terme cartogramme pour désigner des cartes intégrant des graphiques.

Différents types de cartogrammes existent, ils sont répertoriés dans l'état de l'art de Nusrat et Kobourov [70]. La Figure 2.5 présente les principaux types de cartogrammes identifiés : contigu, rectangulaire, non contigu et Dorling.

Plusieurs algorithmes ont été proposés afin de générer des cartogrammes contigus. Parmi les plus populaires nous pouvons citer celui proposé par Dougenik et al. [23] en 1985 (le cartogramme de la Figure 2.6d a été générée à partir de cet algorithme). Il y a également la méthode basée sur la diffusion de Gastner et Newman [34] en 2004 qui s'appuie sur des idées issues de la physique élémentaire. Elle a été améliorée par Gastner et al. [35] en 2018 en accélérant le processus de production.

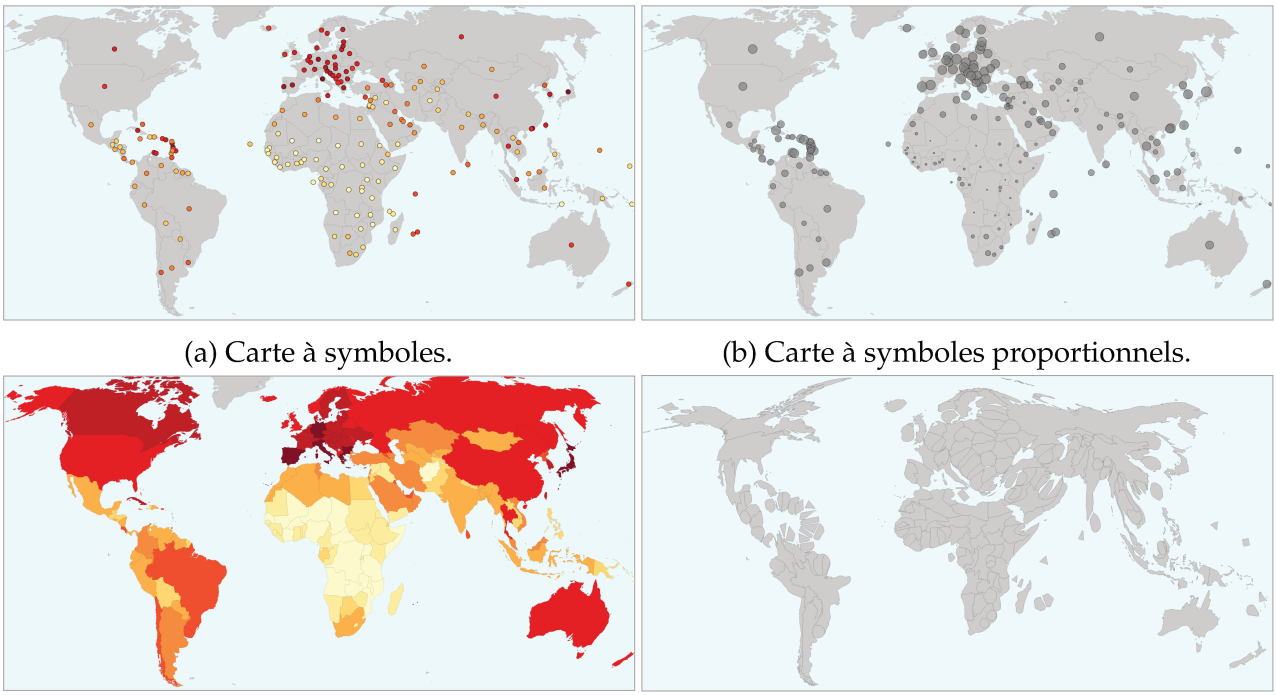


(a) Contigu. (b) Rectangulaire.

(c) Non contigu. (d) Dorling.

FIGURE 2.5 – Principaux types de cartogrammes. (a) et (b) représentent les élections présidentielles des États-Unis de 2004 tandis que (c) et (d) celles de 2012.

Nous avons présenté les trois principaux types de cartes. Il en existent d'autres qui sont des variantes. Une nouvelle technique présentée comme solution alternative aux cartogrammes a été introduite par Roth et al. [79] : la carte valeur-par-alpha (*value-by-alpha*). Elle consiste à incorporer un masque d'opacité dans une carte choroplèthe : plus la valeur est élevée, plus la couleur est opaque. Son principal avantage est que, contrairement aux cartogrammes, les formes géographiques sont préservées. L'ouvrage *Thematic cartography and geovisualization* de Slocum et al. [83] fait un état des lieux, dans la seconde partie (Mapping Techniques), des différents types de cartes. En complément des trois principales, nous avons notamment la carte dasymétrique, carte isarithmique et la carte à points. Les deux premières sont dérivées des cartes choroplèthes. Crampton [18] présente la carte dasymétrique comme une bonne alternative à la carte choroplèthe car celle-ci prend en compte les variations au sein d'une entité géographique. En effet, les unités ne sont pas définies avant la collecte des données mais selon les données. Crampton indique également que la distribution des popu-



(a) Carte à symboles.

(b) Carte à symboles proportionnels.

(c) Carte choroplèthe.

(d) Cartogramme.

FIGURE 2.6 – Différents types de cartes représentant les mêmes données (âge médian de la population).

lations suivent rarement des frontières politiques rendant ainsi les choroplèthes trompeuses. Une connaissance spécifique préalable du lieu à cartographier est cependant nécessaire. La carte isarithmique est assez similaire mais utilise des lignes pour relier des lieux ayant des valeurs similaires. Elle est particulièrement utilisée pour représenter des températures. La carte à points est quant à elle une variante de la carte à symboles. Elle utilise comme symboles des points et chacun d'eux peut représenter l'emplacement réel d'un phénomène individuel. La carte très connue de John Snow à propos de l'épidémie de choléra à Londres en 1854 en est un bon exemple.

2.2.1.4 Perception graphique et représentation cartographique

La perception graphique est, comme le défini Saket et al. [80], la capacité d'interpréter avec précision les valeurs des données à partir de visualisations. Colin Ware indique dans son ouvrage sur la perception au service de la conception de la visualisation d'information [96] que le même système visuel de l'être humain est adapté pour recevoir des représentations de données de certaines manières. De manière générale, l'être humain est confronté à des phénomènes sensoriels, tels que les illusions présentées à la Figure 2.7, qui persistent malgré le fait qu'on en ait conscience. Dès l'instant où de telles illusions apparaissent dans des représentations visuelles, elles sont alors susceptibles d'induire en erreur. Une très bonne compréhension du fonctionnement de la perception de l'œil humain et du fonctionnement du cerveau est donc essentielle pour concevoir des visualisations efficaces et pertinentes. Dans son ouvrage sur la sémiologie graphique [7], Bertin énumère lui aussi l'importance de la prise en considération de la perception du lecteur : «*Pour qu'une construction graphique soit "utile" il faut connaître le processus de perception visuelle. Il détermine des règles qui conduisent à l'efficacité*». Comme défini par Bertin, nous considérons une visualisation plus efficace qu'une autre si pour obtenir une interprétation correcte elle requiert un temps

d'observation plus court.

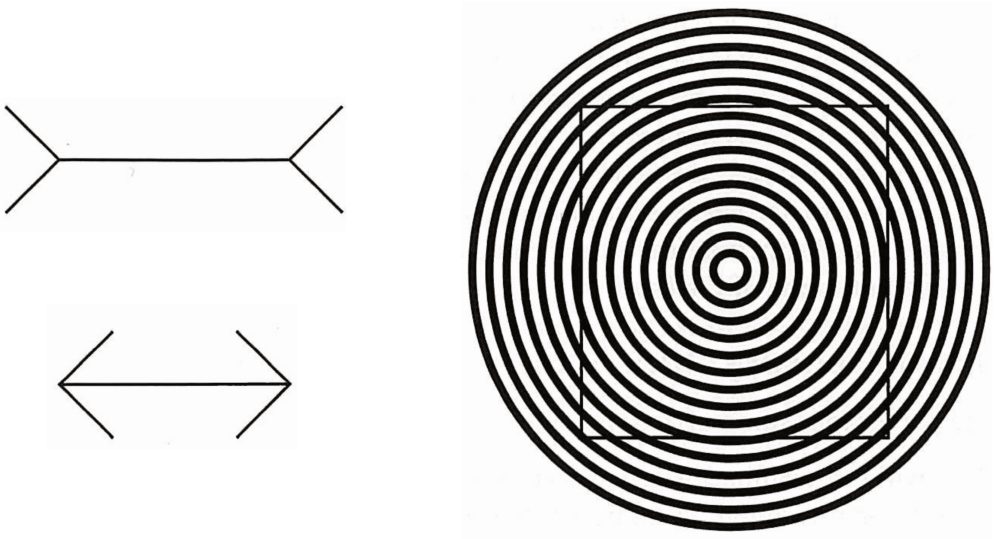


FIGURE 2.7 – Dans l'illusion de Muller-Lyer, à gauche, la ligne horizontale du haut semble plus longue que la même ligne du bas. À droite, le rectangle semble déformé (Ware [96]).

L'étude de Cleveland et McGill en 1984 [17], très largement reconnue, présente la perception graphique comme étant le décodage visuel de l'information codée sur des graphiques. Dans cette étude, 10 tâches perceptives élémentaires que les gens utilisent pour extraire des informations quantitatives des graphiques ont été identifiées et classées en fonction de la précision avec laquelle les personnes les exécutaient. Une tâche perceptive élémentaire est considérée comme plus précise qu'une autre si elle conduit à des jugements humains qui se rapprochent des quantités réelles encodées. Les 10 tâches des plus précises au moins précises sont (1) la position sur une échelle commune, (2) la position le long d'échelles non alignées, (3) la longueur, direction et angle, (4) la surface, (5) le volume et courbure, (6) l'ombrage et la saturation des couleurs. Plus récemment, en 2014 Munzner [64] propose également un classement de ces tâches qu'elle nomme canaux visuels. Son classement des canaux visuels est très semblable à celui de Cleveland et McGill mais, à différence de ce dernier, la couleur et l'ombrage/luminosité ne sont pas en dernière position et se classent juste avant les courbes suivies du volume en dernière position (cf. Figure 2.8). Concernant les cartes, si l'on se tient à ces classements, les cartes à symboles proportionnels et les cartogrammes seraient plus effectifs que les cartes choroplèthes et cartes à symboles utilisant la variable visuelle couleur. Cleveland et McGill précisent également que les cartes avec des ombrages ou saturation des couleurs, soit donc les cartes choroplèthes, sont les plus couramment utilisées à tort.

Ware [96] précise que les couleurs sont interprétées différemment selon les cultures en donnant l'exemple d'une étudiante chinoise qui représentait la suppression d'entités en vert et l'ajout de nouvelles en rouge. En effet, en Chine la mort serait représentée en vert tandis que la chance en rouge ; chose qui ne semble pas naturelle pour une personne occidentale. De ce point de vue, représenter des valeurs à partir des surfaces ou longueurs par exemple serait finalement plus "neutre". L'étude utilisateur de McNabb et al. [63] met en évidence les limites des cartes choroplèthes. Le sujet de cette étude est l'impact de la taille d'une entité géographique sur la précision de la perception. La taille minimale d'une zone pour qu'elle soit perceptiblement utile est proposée. En effet, plus la zone est petite, plus le temps de lecture et la précision seront moins bons. Cependant, comme le démontre Rittschof et al. [76], une mauvaise connaissance des zones géographiques limite l'efficacité des cartogrammes puisque les changements de surface peuvent être très mal interprétés contrairement aux

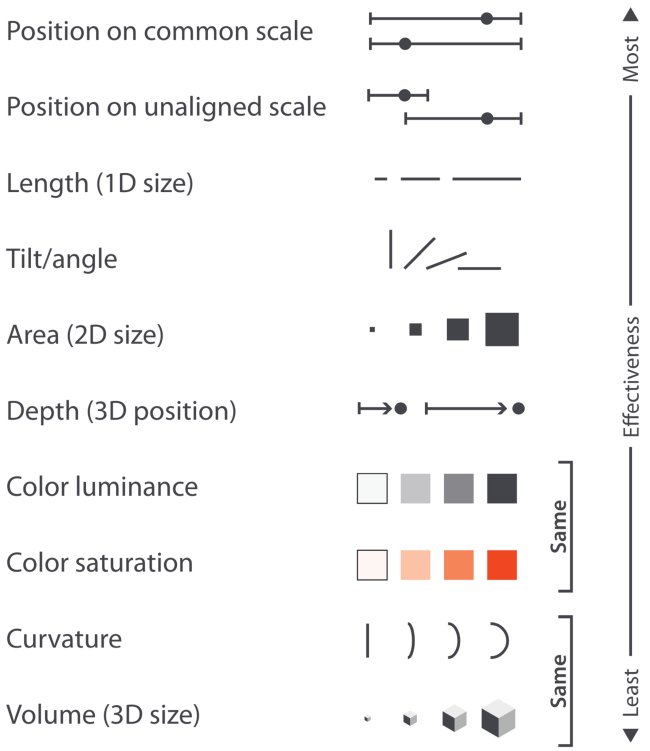
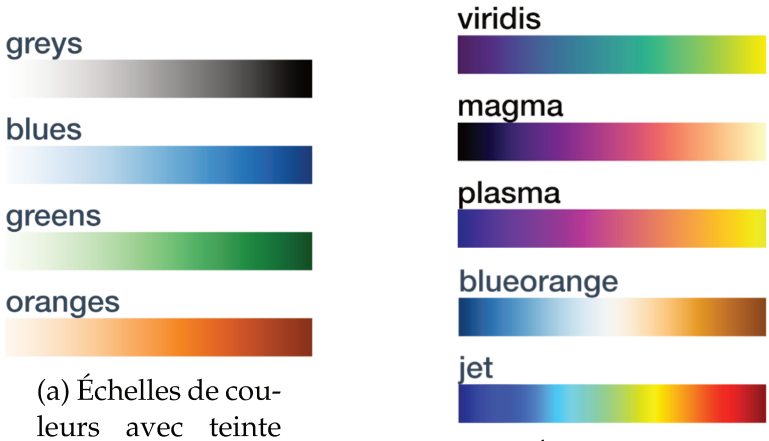


FIGURE 2.8 – Classement proposé par Munzner [64] des canaux visuels, du plus effectif au moins effectif (du haut vers le bas).

cartes choroplèthes. Des études comme celle de Sun et al. [86] ont d'ailleurs montré que les cartes choroplèthes ont de meilleurs résultats que les cartogrammes pour des questions complexes. Il en est de même avec les cartes à symboles proportionnels avec notamment les travaux de Rui Li [58] dont la motivation était, dans le contexte de Covid-19, de visualiser les informations à propos de la pandémie à un niveau géographique fin. Pour cela, il a conçu une carte choroplèthe affichant les cas confirmés de Covid-19 au niveau des comtés de l'État de New York. Pour l'évaluer, cette carte a été comparée à deux autres cartes, toutes deux utilisant des symboles proportionnels. Les résultats de cette étude montrent que les participants ont eu une meilleure compréhension avec la carte choroplèthe. L'étude de Brewer et al. [10] qui s'est concentré sur la représentation de données de santé avec des cartes choroplèthes montre que la couleur permet une plus grande précision dans la lecture des cartes. Les cartes choroplèthes, malgré certaines critiques, restent finalement très largement utilisées étant donné que leur compréhension et création sont très aisées.

Des recherches ont d'ailleurs été menées concernant le choix des échelles de couleurs. Les travaux de Brewer et al. [10] en 1997 montrent que les participants ont préféré des teintes spectrales violettes/vertes et que les schémas divergents ont amélioré l'efficacité des cartes. Ware [97] indique que l'une des exigences les plus évoquées d'une bonne carte choroplèthe est qu'elle soit perceptuellement uniforme. Les travaux de Liu et Heer [59] de 2018 sont une référence pour les échelles de couleurs quantitatives. Ils confirment l'inefficacité des cartes arc-en-ciel des travaux antérieurs et déconseillent fortement l'utilisation de cette échelle de couleurs (cf. "jet" de la Figure 2.9b). Les cartes de couleurs multi-teintes comme viridis (cf. Figure 2.9b) seraient plus performantes. Cette échelle de couleurs se rapproche fortement des teintes les plus appréciées de l'étude de Brewer et al. [10] présentée précédemment. Les échelles de couleurs à teinte unique (cf Figure 2.9a) ont également donné de bons résultats sur de grandes échelles mais moins pour les plus petits intervalles de données.



(a) Échelles de couleurs avec teinte unique.

(b) Échelles de couleurs multi-teintes.

FIGURE 2.9 – Échelles de couleurs comparées par Liu et Heer [59].

Brewer et Pickle [11] de leur côté se sont penchés sur les méthodes de classification de cartes choroplèthes les plus adaptées. Pour cela, plusieurs méthodes ont été comparées (cf. Figure 2.10). Les quantiles (2B dans la Figure 2.10) se sont montrés comme les meilleures méthodes pour faciliter la comparaison et la lecture générale des cartes choroplèthes.



FIGURE 2.10 – Aperçu des méthodes de classification de cartes choroplèthe comparées par l'étude de Brewer et Pickle [11].

Concernant les cartogrammes, non abordés par Cleveland et McGill, ils ont fait l'objet de nombreuses études et discussions. Leur principal avantage est d'assurer une meilleure cohérence entre la saillance visuelle via la surface et l'importance de la valeur des données. Cependant, son plus grand atout peut aussi finalement être son point faible puisque

la lecture est complexifiée. En effet, la reconnaissance et l'identification d'une entité géographique est perturbée et l'estimation de la valeur biaisée à cause des connaissances préalables du lecteur. En effet, durant l'étude utilisateur menée par Dent [21], les participants ont trouvé les cartogrammes déroutants et difficiles à lire mais en même temps intéressants et novateurs. Les travaux de Sun et al. [86] en 2010 évaluent la pertinence des cartogrammes pour représenter des données qualitatives et quantitatives. Pour cela, deux expérimentations ont été effectuées, une première consistait à comparer des cartogrammes avec d'autres cartes thématiques (cartes à symboles proportionnels et choroplèthes). La seconde compare des types de cartogrammes. Ce qui en résulte c'est que pour des données quantitatives, les cartogrammes contigus sont moins probants que les autres cartes thématiques. Parmi les types de cartogrammes c'est le pseudo-cartogramme (avec grille en fond) qui s'est montré le plus efficace et le cartogramme de Dorling (cf. Figure 2.5d) le moins convainquant. D'autres études ont été menées comme celle de Nusrat et al. [68] en 2016 qui évalue et compare les 4 principaux types de cartogrammes : contigu, non contigu, rectangulaire et Dorling. Le cartogramme contigu s'est avéré le plus efficace pour plusieurs tâches : localisation, recherche de la région ayant la plus grande valeur et détection d'un changement. L'étude de Nusrat et al. [67] de 2020 compare la capacité de mémorisation de deux types de cartogrammes : contigu et Dorling. Les cartogrammes de Dorling seraient les mieux adaptés pour montrer les grands modèles et les tendances.

Très récemment, en 2023, l'expérimentation de Fung et al. [30] évalue l'utilité des légendes des valeurs par rapport aux surfaces et les lignes de quadrillage pour aider les lecteurs à mieux évaluer les surfaces (cf. Figure 2.11). Les résultats de l'étude ont révélé que les légendes et les lignes de quadrillage ralentissaient l'estimation des valeurs sans en améliorer la précision. Les estimations sont cependant moins variables. La légende sélectionnable donne plus de confiance aux utilisateurs dans leur évaluation des données présentées.

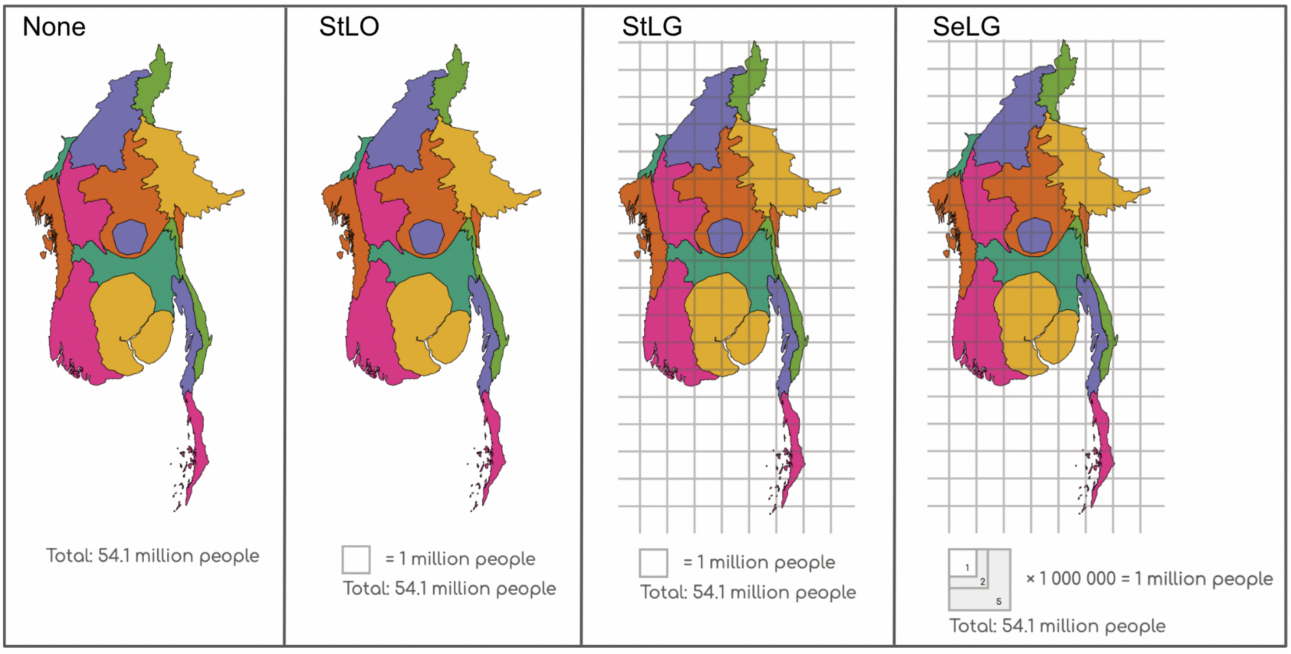


FIGURE 2.11 – Illustration des quatre traitements étudiés dans l'étude de Fung, Perrault et Gastner [30] en utilisant le cartogramme de la population de la Birmanie comme exemple. (1) None : ni légende ni lignes de grille. (2) StLO : légende statique uniquement. (3) StLG : légende statique et lignes de la grille. (4) SeLG : légende et lignes de grille sélectionables.

Bien d'autres études utilisateurs cherchant à mieux comprendre et définir l'efficacité des

cartes existent. Les plus récentes, comme celle de Candela et al. [14], font appel au eye-tracking afin de mieux connaître le mécanisme de lecture d'une carte.

Finalement, le choix des variables visuelles, implantations, échelles de couleurs, et autres paramètres doit être adapté au besoin, contexte, problématique et aux données. En plus du choix de la représentation cartographique vient aussi très souvent le besoin de faire des comparaisons. En effet, pour mieux appréhender des phénomènes et mieux les comprendre avant de faire des analyses plus approfondies, il est nécessaire de visualiser plusieurs variables. Deux solutions majeures sont alors possibles, soit de juxtaposer les visualisations soit de les superposer.

2.2.2 Comparaison

Dans cette section, nous nous focalisons sur les approches permettant de comparer des visualisations de données quantitatives associées à des entités géographiques. On distingue deux approche majeures : la juxtaposition et la superposition. Ces dernières sont toutes les deux présentées dans cette section.

2.2.2.1 Juxtaposition

Comme l'annonce Munzner [64] la visualisation de données constituée de plusieurs vues juxtaposées implique de nombreux choix quant à la manière de les coordonner pour créer des vues liées. Parmi les choix à faire, on en distingue quatre principaux à savoir : (1) les vues partagent-elles le même codage visuel ou utilisent-elles des codages différents ? (2) la mise en évidence est-elle liée entre les vues ? (3) les vues présentent-elles les mêmes données, l'une présente-t-elle un sous-ensemble de ce qui se trouve dans l'autre, ou les vues présentent-elles un partitionnement disjoint où chacune présente un ensemble différent ? (4) la navigation est-elle synchronisée entre les vues ?

La juxtaposition de plusieurs cartes, nommée *small-multiple maps* dans la littérature, est l'une des techniques les plus utilisées pour visualiser les évolutions. Dans ses travaux Slocum et al. [83] fournit une évaluation utilisateur de mapTime, un logiciel d'exploration de données spatio-temporelles. Ce logiciel propose des cartes animées et aussi des cartes juxtaposées. Chaque carte est associée à une date. Les utilisateurs ont notamment apprécié le fait d'avoir une vue d'ensemble qui permet de simplement identifier des dates clés et connaitre l'évolution globale des données représentées. Le principal avantage de ces cartes multiples est de ne pas avoir besoin de mémoriser contrairement aux cartes animées.

La visualisation élaborée pour présenter les travaux de Eisenstein [24] constitue un bon exemple de visualisation par juxtaposition (cf. Figure 2.12). Cette représentation permet de constater l'évolution de l'utilisation de certains mots au fil des semaines en un coup d'œil mais aussi de comparer les différents mots.

La visualisation de Johnson et al. [49] est un autre bon exemple de visualisation par juxtaposition (cf. Figure 2.13). Cette visualisation permet d'observer la corrélation entre l'adoption d'Internet, le PIB et la population au fil des années. Cette visualisation fait également de la superposition puisque chaque carte représente deux variables : une première avec des couleurs et une seconde selon la superficie qui est déformée.

Peña-Araya et al. [73] ont étudié l'efficacité de plusieurs techniques de visualisation pour communiquer la corrélation entre deux variables qui évoluent dans l'espace et dans le temps. En effet, ces travaux prennent également en compte la dimension temporelle pour visualiser des données multivariées. Afin d'extraire des modèles et des tendances, il est

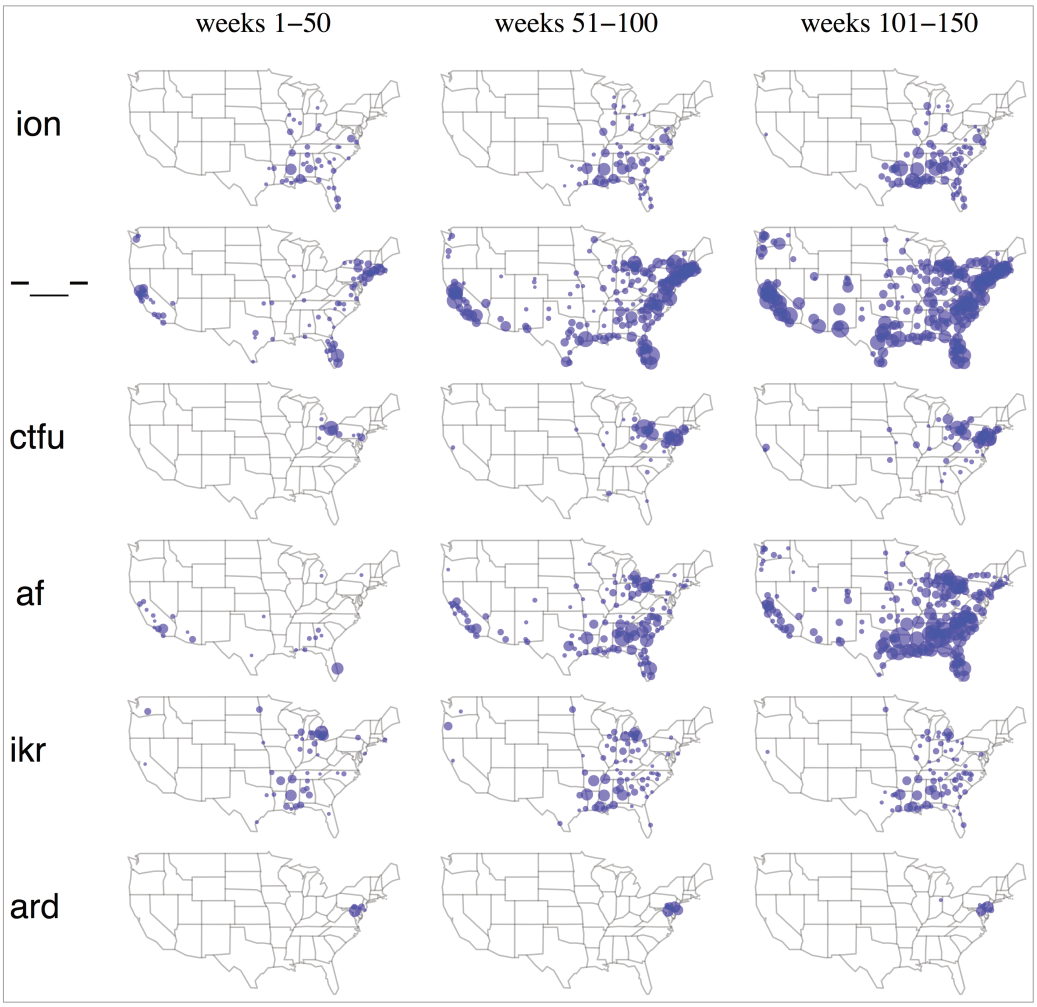
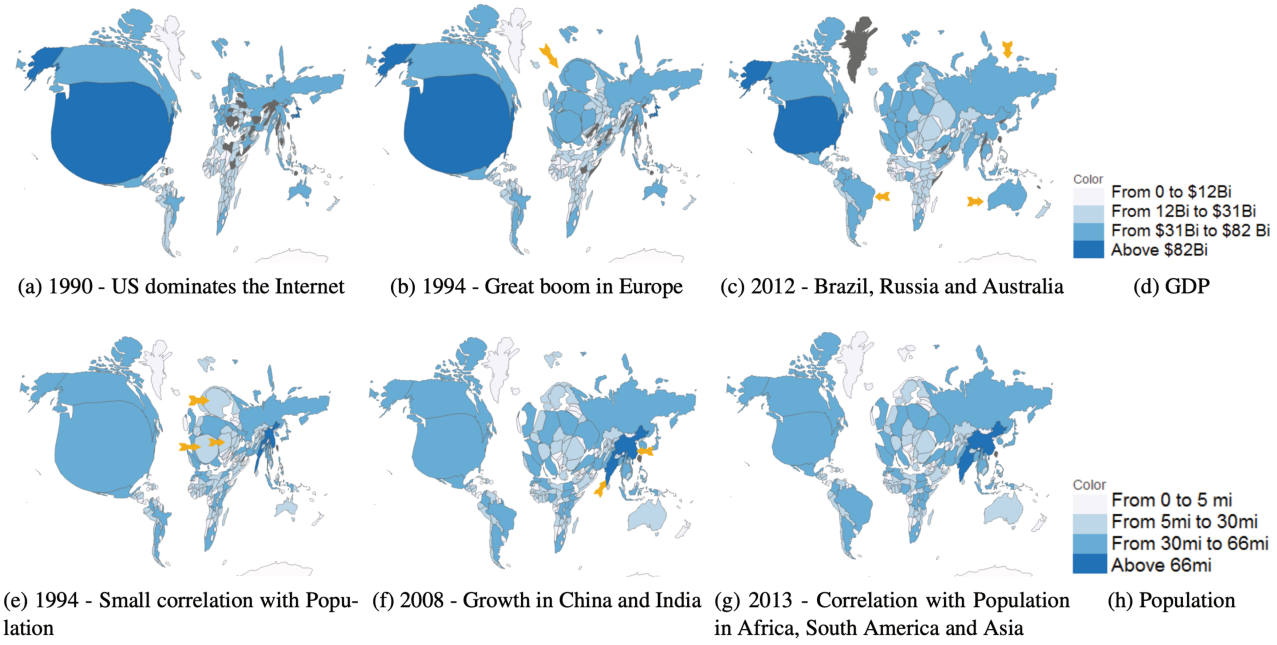


FIGURE 2.12 – Exemple d'une visualisation par juxtaposition [24]. Elle représente les changements de fréquence pour six mots : les cercles indiquent les villes où, en moyenne, au moins 0,1% des utilisateurs utilisent le mot au cours d'une semaine. La surface d'un cercle est proportionnelle à la probabilité du mot.

essentiel de choisir une représentation adéquate. Pour cela, cette étude compare trois stratégies de visualisation de données géo-temporelles multivariées : (1) la juxtaposition de tous les pas de temps en utilisant comme technique de visualisation des cartogrammes de Dorling (cf. Figure 2.16a), (2) la même juxtaposition mais cette fois-ci avec des carte à symboles (cf. Figure 2.16b), (3) une carte avec des diagrammes à barres où chaque barre correspond à une date (cf. Figure 2.16c). La première variable, qui est le taux d'alphabétisation, est respectivement représentée à partir de la surface des entités, de la taille des symboles et de la hauteur des barres. La seconde variable, le nombre d'heures de travail hebdomadaire, est représentée par la variation de saturation de la teinte bleue. Les résultats montrent que l'efficacité d'une visualisation dépend fortement de la tâche à accomplir. En effet, pour identifier des corrélations à un moment précis, avec la carte unique composée de diagrammes à barres les temps de réponses ont été presque deux fois plus long car les participants devaient faire abstraction des dates non concernées tandis qu'avec les deux autres stratégies de juxtaposition il leur suffisait de se concentrer sur la carte correspondant à la date d'intérêt. En revanche, pour identifier les corrélations et l'évolution temporelle sur tous les pas de temps, il est préférable d'utiliser une carte unique avec des diagrammes à barres.



(a) 1990 - US dominates the Internet (b) 1994 - Great boom in Europe (c) 2012 - Brazil, Russia and Australia (d) GDP

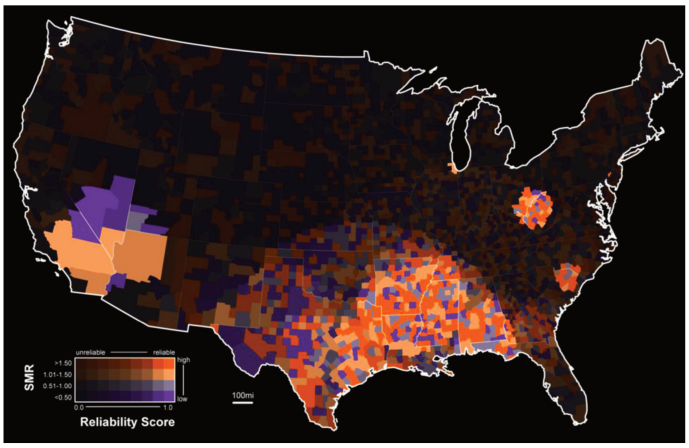
(e) 1994 - Small correlation with Population (f) 2008 - Growth in China and India (g) 2013 - Correlation with Population in Africa, South America and Asia (h) Population

FIGURE 2.13 – La superficie des pays correspond au nombre de AS (système autonome : ensemble de réseaux informatiques IP intégrés à Internet). La première ligne représente le PIB par pays et la seconde ligne la population [49].

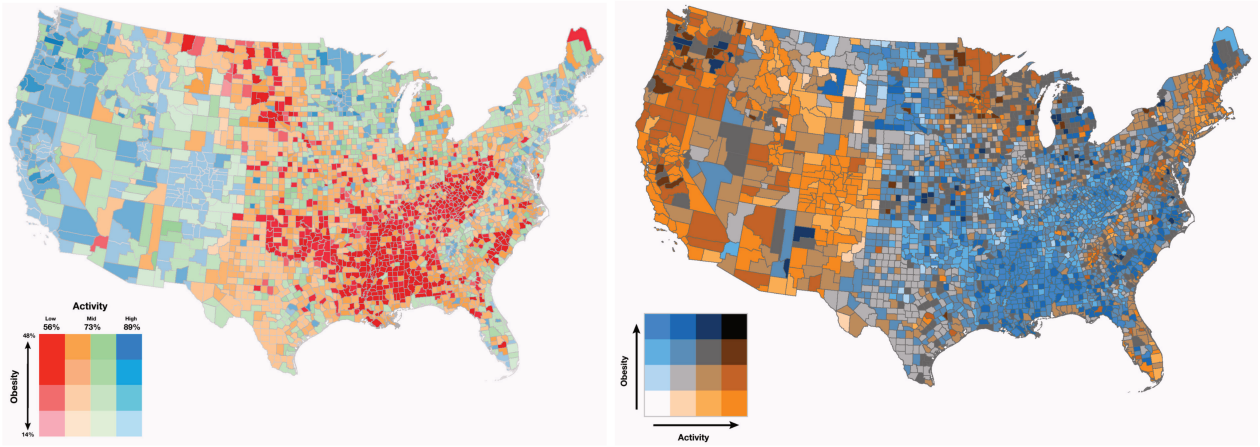
2.2.2.2 Superposition

Pour rappel, dans ce manuscrit nous parlons de superposition lorsqu'on a une unique carte pour représenter deux variables. Les méthodes de cartographie univariées de variables quantitatives ont été présentées dans la Section 2.2.1.3 : cartes à symboles (*Sy*), cartes Choroplèthes (*Ch*) et cartogrammes (*Ca*). Comme mentionné par Strode et al. [85], les cartes univariées sont à la fois standard et répandues. Pour comparer deux variables et observer des corrélations, la superposition peut-être une bonne alternative à la juxtaposition. En effet, les cartes bivariées répondent à ce besoin en combinant deux cartes univariées.

Des cartes bivariées peuvent être créées en utilisant diverses combinaisons de techniques cartographiques. Il est assez courant de combiner la technique de conception de carte choroplèthe avec la technique de conception de cartogramme. Les cartes bivariées ne sont cependant pas nécessairement une combinaison de deux types de cartes. La carte valeur-par-alpha introduite par Roth et al. [79], présentée comme une solution alternative aux cartogrammes, peut représenter deux variables. En effet, si nous avons une échelle de couleur pour une première variable et un masque d'opacité pour une seconde, alors la carte est bivariée comme dans la Figure 2.14a. Leonowicz [57] affirme que la juxtaposition des cartes choroplèthes n'est pas une bonne solution, le processus de lecture séparée et de comparaison entre les cartes serait biaisé par la vision humaine. Comme alternative, les cartes bivariées choroplèthes sont proposées. Les cartes choroplèthes à deux variables, similaires aux cartes de valeur-par-alpha, sont une autre méthode de représentation de deux variables sur la même carte. Trumbo en 1981 [94] se concentre sur leur conception et plus récemment Strode et al. en 2019 [85] en s'appuyant sur le travail de Trumbo. La conception des cartes choroplèthes bivariées repose particulièrement sur le choix des palettes de couleurs (cf. Figure 2.14b). Une carte choroplèthe à deux variables peut afficher trois relations spatiales : des relations inverses, une plage d'une variable dans une autre et des relations directes. La principale différence avec les cartes valeur par alpha est que ces dernières sont basées sur l'opacité des couleurs, tandis que les choroplèthes sont basées sur la saturation.



(a) Carte valeur-par-alpha bivariée [79].



(b) Cartes choroplèthes bivariées représentant les même données avec des échelles de couleurs différentes [85] .

FIGURE 2.14 – [73]

Une technique pour visualiser deux variables sur un seul cartogramme a été proposée par Nusrat et al. [69]. Les variables sont codées en zones. La relation entre deux variables est représentée par deux couleurs complémentaires (cf. Figure 2.15). Cette technique de représentation est très efficace pour trouver des valeurs aberrantes et peut être appliquée à différents types de cartogrammes. Pour valider cette nouvelle technique de cartographie, une étude utilisateur a été réalisée pour comparer cette méthode avec deux cartogrammes juxtaposés et des cartogrammes associés à la technique choroplèthe. Cependant, seuls les cartogrammes de Dorling ont été évalués et la nature cartographique de cette méthode peut potentiellement compliquer la reconnaissance des zones géographiques. De plus, les deux variables doivent être de même nature, dans la Figure 2.15 une variable correspond à la distribution de McDonald's pour 100 000 résidents et l'autre à la distribution de Starbucks pour 100 000 résidents. On ne pourrait pas avoir comme seconde variable la population par exemple.

Les travaux de Nelson en 1999 [65] fournissent un point de départ empirique pour sélectionner efficacement des combinaisons de variables graphiques pour la conception de cartes bivariées. Ses travaux s'appuient sur la théorie de la sélection attentive. Nelson soutient que la conception de symboles bivariés et multivariés peut être guidée par cette théorie. La théorie de la sélection attentive affirme que la capacité à interpréter la variable visuelle d'un élément est affectée par les autres variables visuelles présentes dans le même sym-

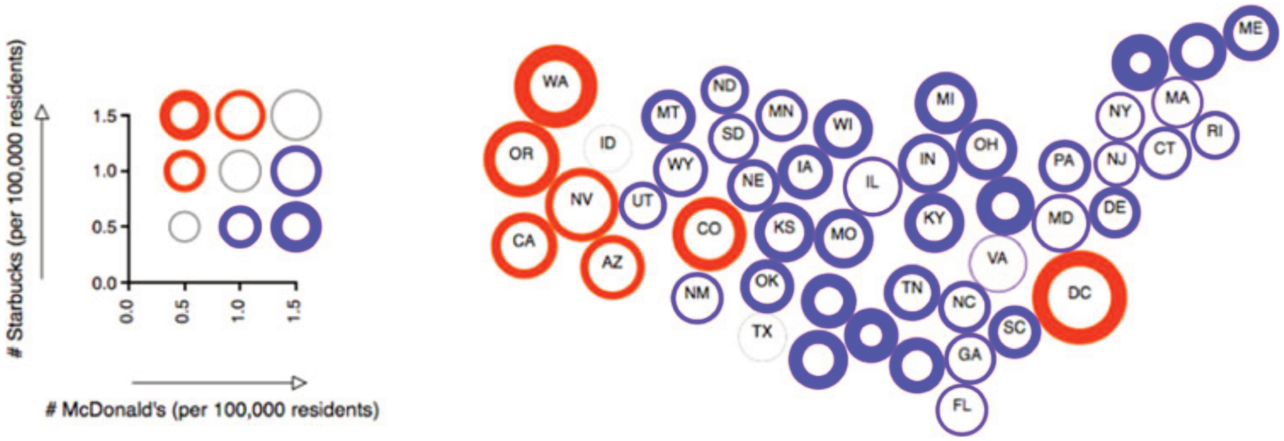


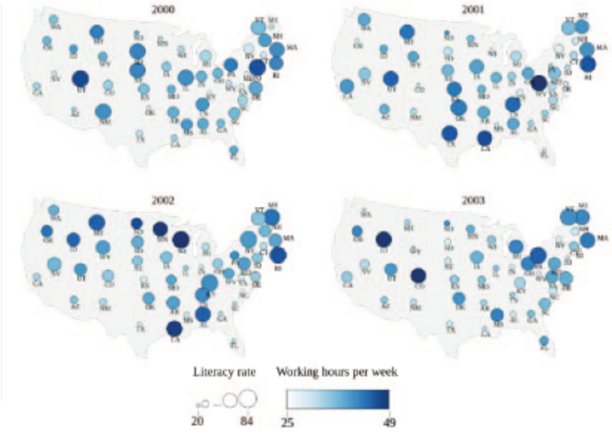
FIGURE 2.15 – Cartogramme bivarié (*bivariate cartogram*) utilisant comme type de cartogramme le Dorling [69].

bole. Les recherches en psychologie suggèrent que certaines combinaisons de variables visuelles peuvent améliorer ou restreindre l'attention sélective. Ces combinaisons de variables peuvent être (1) séparables, c'est-à-dire lorsque les variables peuvent êtres prises en compte indépendamment ; (2) intégrales, si elles ne peuvent être traitées sans interférence ; (3) configurables, si des caractéristiques de séparabilité et à la fois d'intégralité sont présentées ; ou (4) asymétriques, dans ce cas des effets d'interaction entre les deux variables visuelles sont représentés. Chacune de ces configurations présente des avantages et inconvénients selon les conditions dans lesquelles la carte est utilisée. L'étude de Nelson l'année suivante en 2000 [66] fournit une preuve supplémentaire de la force de la théorie de l'attention sélective pour aider à la conception de cartes thématiques bivariées. La capacité des participants à extraire des informations en corrélation ou individuelles a été testée à l'aide de plusieurs modèles de symboles bivariés. Ces modèles ont été définis sur la base de recherches antérieures.

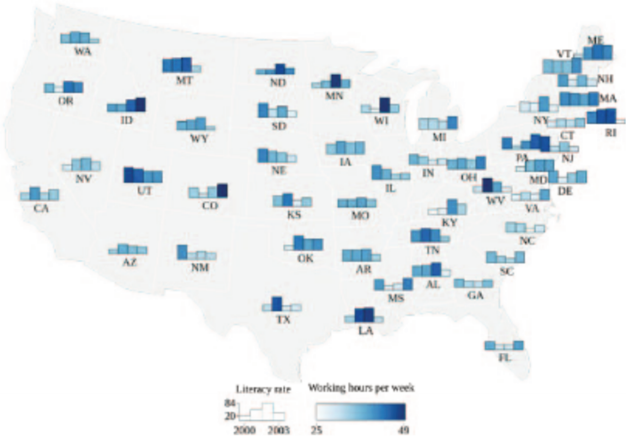
Finalement, comme le précise Elmer [25], il existe de nombreuses possibilités de conception de cartes bivariées, bien plus que pour représenter une seule variable. De nombreux types de cartes bivariées sont déjà couramment utilisés et de nouvelles solutions continuent d'être proposées. Brewer et al. [9] donne des exemples de symboles bivariés, et décrit également leur variété et formes diverses. Malgré le grand nombre de possibilités de conception de cartes bivariées, à notre connaissance, peu d'études similaires à celles de Nelson [65, 66] ont été menées pour comparer des techniques de représentation cartographique bivariée. Nous pouvons, cependant, en citer quelques-unes comme l'étude de Gao et al. [33] où la carte valeur-par-alpha a été comparée avec des cartogrammes non contigus et des cartes à symboles proportionnels. Selon cette étude, les cartes de symboles proportionnels fonctionnent mieux dans l'ensemble. Une autre étude, de Peña et al. [73], que nous avons déjà évoqué dans la section précédente concernant la juxtaposition, compare 3 visualisations de cartes multivariées. Pour rappel, les visualisations de la Figure 2.16a et de la Figure 2.16b juxtaposent des cartes qui elles-mêmes superposent deux techniques, respectivement cartogramme de Dorling et symboles proportionnels avec la variable visuelle couleur. La visualisation de la Figure 2.16c quant à elle est une carte à barres proportionnelles, une barre correspond à une date. Lorsqu'un seul point dans le temps était pris en compte, les cartogrammes obtenaient les moins bons résultats. Nous avons également pu en noter une autre étude, de Kaspar et al. [51], qui compare l'efficacité de deux cartes bivariées. La première carte est un cartogramme combiné à une carte choroplèthe. La seconde est une carte choroplèthe combinée à des cercles gradués. Les résultats de cette étude ont montré que dans l'ensemble, la carte la plus efficace et efficiente pour l'analyse des données du recensement



(a) Juxtaposition de cartogrammes de Dorling. Chaque carte correspond à une date.



(b) Juxtaposition de carte à symboles proportionnels. Chaque carte correspond à une date.



(c) Carte constituée de diagrammes à barres. Chaque barre correspond à une date.

FIGURE 2.16 – Visualisations comparées dans l'expérimentation de Peña-Araya et al. [73].

de la population est la carte choroplèthe avec des cercles gradués. Les deux correspondent à des interactions dimensionnelles séparables, il n'y a donc pas de comparaison entre les différentes interactions dimensionnelles. L'étude d'Elmer [25] compare également ces deux types de cartes bivariées avec 6 autres types de cartes bivariées. Cette étude est plus complète, comptant 2 types de cartes bivariées par configuration d'interaction dimensionnelle (séparable, intégrale, configurale et asymétrique) comme le montre la Figure 2.17. Le but de sa recherche est d'examiner si et comment les caractéristiques perceptives des types de cartes bivariées affectent la capacité à extraire des informations.

2.3 Comparaison par juxtaposition

Dans le cadre de la visualisation de données associées à des entités géographiques, la comparaison par juxtaposition consiste à agencer plusieurs cartes sur une même visualisation. Différentes configurations sont possibles : chaque carte peut représenter les valeurs d'une ou plusieurs variables pour une date/période précise. Il est également possible de représenter sur une même carte une variable à différentes dates, chaque carte correspondant alors à une variable. À l'inverse, une comparaison par superposition consiste à visualiser

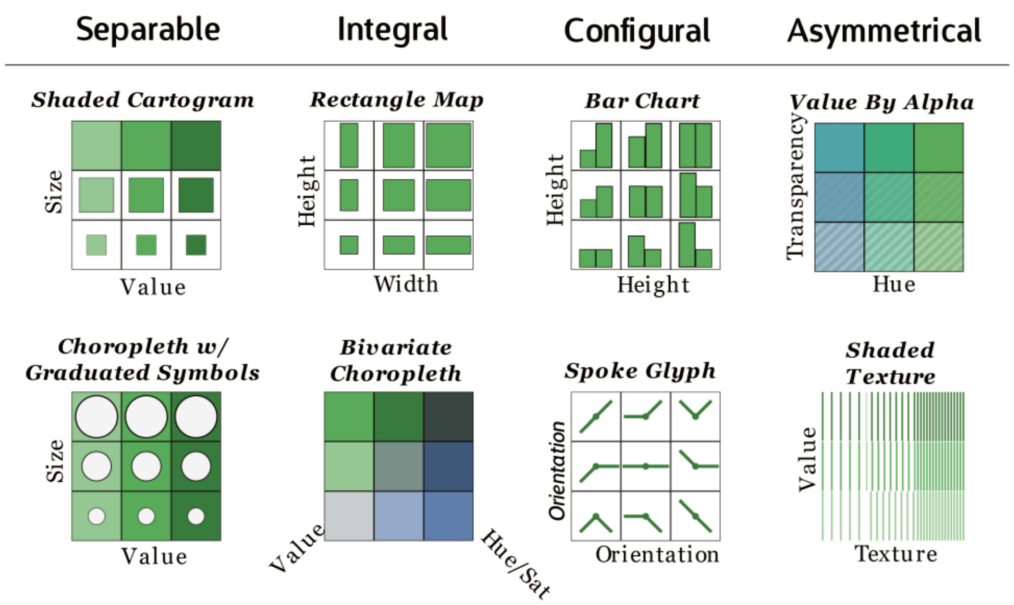


FIGURE 2.17 – Cartes bivariées étudiées par Elmer, elles sont classées selon les différentes conditions de sélectivité [25].

plusieurs variables et/ou plusieurs dates sur un même fond de carte.

Agencer un ensemble de cartes en les juxtaposant présente l'avantage d'avoir une vue d'ensemble. Cependant, la visualisation est vite surchargée et on accède plus difficilement à des informations plus précises lorsqu'on veut aller au-delà d'une vue générale. C'est pour cette raison que nous proposons une interface dynamique où il est possible de modifier la date et la variable à visualiser. Dans l'interface, deux cartes sont juxtaposées pour explorer et comparer des données spatio-temporelles.

Dans cette section, l'approche de comparaison par juxtaposition que nous proposons est d'abord exposée. Un cas d'application de cette approche avec la plateforme Epid Data Explorer, que nous avons développé dans un contexte de visualisation de données épidémiologiques, est ensuite présenté.

2.3.1 Approche proposée

Pour comparer des cartes, les juxtaposer peut-être suffisant si l'on a besoin de comparer au maximum deux éléments bien définis. En revanche, si l'on a besoin de naviguer sur les cartes, de visualiser les valeurs de plusieurs variables à différentes dates, cela peut devenir laborieux. Par exemple, si on souhaite comparer deux régions différentes en visualisant les valeurs d'une variable pour une même date, on doit alors modifier sur chaque carte la date et la variable à visualiser à chaque fois. À partir de ces constats, des besoins ont été définis et pour y répondre nous avons mis en place une approche spécifique accompagnée de fonctionnalités supplémentaires.

2.3.1.1 Besoins

On souhaite visualiser les valeurs quantitatives de variables associées à des entités géographiques et faire des comparaisons sur une unique visualisation, sans avoir besoin d'ouvrir de multiples fenêtres. Les comparaisons sont réalisées à partir de 3 dimensions : (1) variable (la valeur de la variable visualisée), (2) temporelle (la date ou période correspondante)

et (3) spatiale (la zone géographique). On doit pouvoir comparer les deux cartes juxtaposées mais aussi comparer les tendances des variables pour chaque entité géographique. Nous avons alors distingué deux besoins :

Besoin B1. L'ensemble des combinaisons possibles entre les trois dimensions pour faire des comparaisons doit être pris en charge sur une unique visualisation. Par exemple, on doit pouvoir comparer pour une même zone géographique et à une même date deux variables différentes, comparer pour une même variable deux zones géographiques différentes à une même date ou bien encore comparer pour une même zone géographique une variable à deux dates différentes.

Besoin B2. Il doit être possible pour un utilisateur de visualiser l'évolution d'une variable afin de détecter d'éventuelles tendances. La visualisation de l'évolution d'une variable doit être observable pour chaque entité géographique de manière indépendante. Une comparaison de l'évolution d'une variable pour plusieurs entités géographiques doit être également possible tout en restant toujours sur la même visualisation.

C'est à partir de ces deux besoins que nous avons élaboré une visualisation avec deux vues, chacune composée d'une carte, proposant des fonctionnalités spécifiques.

2.3.1.2 Visualisation avec deux vues

Si on reprend les choix à faire pour concevoir une visualisation constituée de plusieurs plusieurs vues juxtaposées, déterminés par Munzner [64] et formulés dans la Section 2.2.2.1, la visualisation composée de deux vues juxtaposées que nous proposons est définie comme suit : (1) les vues partagent le même codage visuel, à savoir la technique de cartographie choroplèthe, (2) la mise en évidence n'est pas liée entre les deux vues, chacune est indépendante sur ce point, (3) les deux vues peuvent présenter les mêmes données mais également des données différentes, (4) la navigation est synchronisée entre les deux vues selon le souhait de l'utilisateur. En effet, pour effectuer des comparaisons l'approche que nous proposons se base sur un système de synchronisation des trois dimensions qui sont la variable, la date et la zone géographique. Les éléments clés de la synchronisation entre les deux vues sont des cadenas (cf. Figure 2.18-ⓐ, ⓑ, et ⓒ). Chaque cadenas correspond à une dimension. Dans le cas où il est verrouillé cela signifie que les deux vues sont synchronisées et, à l'inverse, s'il est déverrouillé alors chaque vue est indépendante pour la dimension concernée. Dans l'exemple de la Figure 2.18, les cadenas associés à la dimension spatiale et à la sélection de la variable à visualiser sont verrouillés tandis que celui de la dimension temporelle est déverrouillé. Notons bien que les deux vues fonctionnent exactement de la même manière. Dans la Figure 2.18 seuls les éléments de la vue de gauche sont étiquetés pour plus de lisibilité (ⓓ, ⓔ, et ⓕ). La navigation sur la carte (Figure 2.18-ⓓ) facilite l'exploration des différentes zones géographiques et permet de zoomer vers les zones d'intérêt pour plus de détails. Comme évoqué plus tôt, le cadenas associé à la dimension spatiale est verrouillé. Les deux cartes sont donc synchronisées, le moindre changement sur une des deux cartes (zoom et/ou déplacement) engendrera exactement la même action sur la seconde. Il est en de même pour la variable sélectionnée à partir d'une liste déroulante (Figure 2.18-ⓔ). Le cadenas étant verrouillé, lorsqu'on change de variable sur la vue de gauche ou de droite, l'autre vue aura le même changement. Quant à la dimension temporelle dont la valeur peut être éditée à partir de la timeline (Figure 2.18-ⓕ), elle n'est pas synchronisée entre les deux vues puisque le cadenas est déverrouillé. Chaque vue est indépendante au niveau de la date, dans l'exemple de la Figure 2.18 la date du 29 mai 2021 est comparée à la même date

de l'année 2022. Le verrouillage/déverrouillage se fait par un simple clic sur le cadenas. Par exemple, si l'utilisateur clique sur le cadenas traitant la synchronisation de la navigation sur les cartes (Figure 2.18-ⓐ) alors il sera déverrouillé. Ainsi, il pourra naviguer librement sur une des deux cartes sans engendrer le moindre changement sur la seconde. Ceci peut, par exemple, permettre de comparer deux zones différentes.

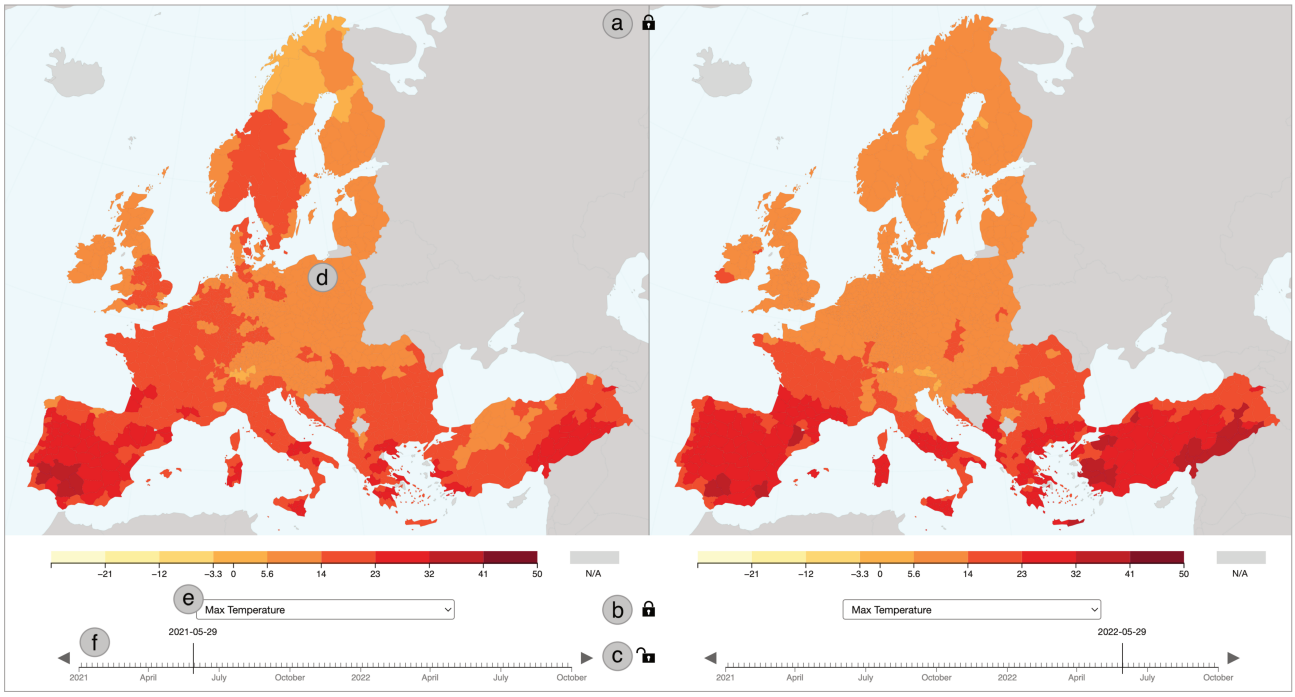


FIGURE 2.18 – Exemple d'une visualisation avec deux vues.

Finalement, le système de synchronisation permet bien de répondre au besoin B1 puisque toutes les combinaisons possibles sont bien prises en charge par l'approche que nous proposons. Dans l'exemple sur lequel on s'est appuyé pour présenter les fonctionnalités de la visualisation avec deux vues, les cartes sont de type choroplèthe mais il est bien évidemment possible d'utiliser d'autres types de carte.

2.3.1.3 Tooltips

En naviguant sur une carte, l'utilisateur peut aussi accéder à des informations plus détaillées concernant la variable sélectionnée. Pour cela, il suffit de cliquer sur une entité pour qu'un tooltip s'affiche. Les informations affichées sont relatives à l'entité géographique sélectionnée et concernent l'ensemble de la période couverte par les données de la variable. Il est possible de repositionner et redimensionner un tooltip sur la carte. Pour ajuster sa taille, il suffit de glisser les bords de la fenêtre. La Figure 2.19 montre un exemple où trois tooltips sont ouverts : deux sur la carte de gauche et un autre sur la carte de droite. L'affichage simultané de plusieurs tooltips répond au besoin B2 étant donné qu'on peut visualiser l'évolution des valeurs d'une variable pour une entité géographique et en comparer plusieurs. Dans l'exemple proposé par la Figure 2.19, on peut comparer les valeurs de la température maximale dans trois zones différentes. Notons que si l'utilisateur modifie la variable sur une vue (ou les deux si elles sont synchronisées) les tooltips ouverts se mettent à jour automatiquement.

La Figure 2.20 présente les éléments clés d'un tooltip. Les valeurs de la variable au fil du temps pour l'entité géographique sélectionnée sont représentées par une courbe noire

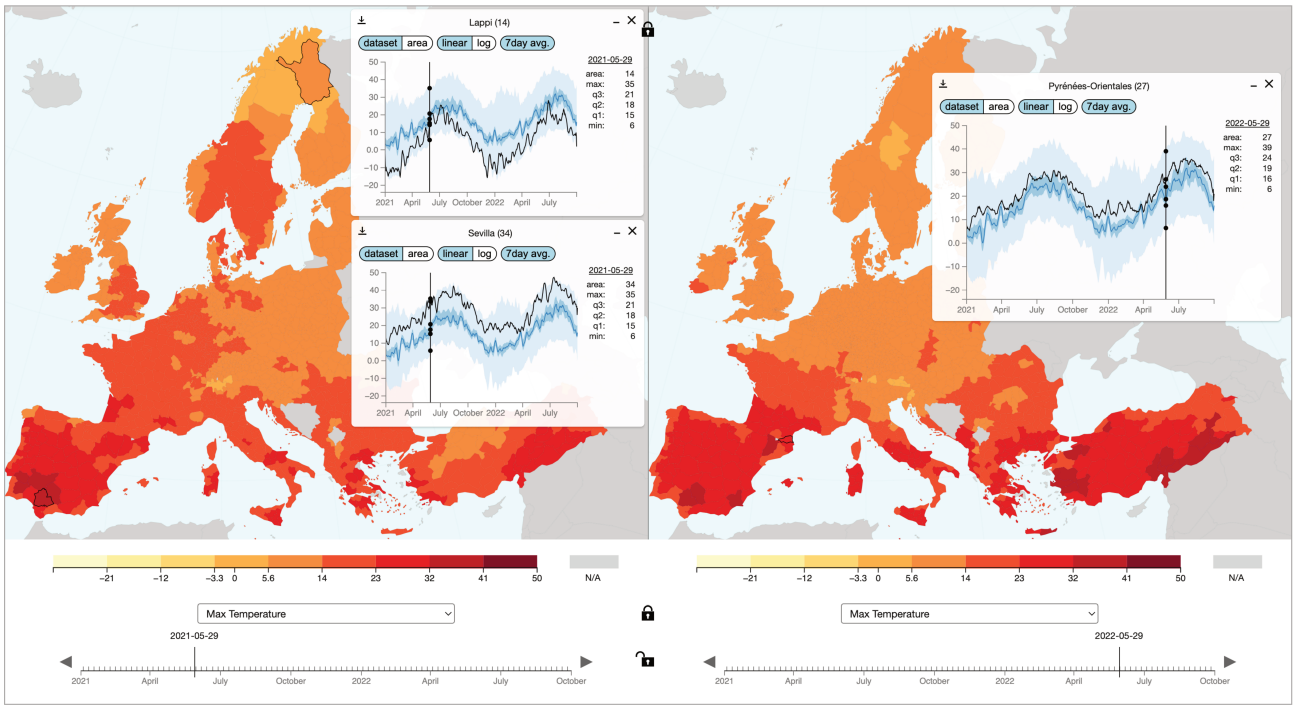


FIGURE 2.19 – Exemple d'une visualisation avec deux vues, dans les deux vues des tooltips sont ouverts.

(Figure 2.20-ⓐ). L'évolution de la valeur médiane, de toutes les entités géographiques, est représentée par une courbe bleue (Figure 2.20-ⓑ). La zone bleu clair (Figure 2.20-ⓒ) couvre l'ensemble des valeurs, les délimitations correspondant aux valeurs minimales et maximales. La zone bleu foncé (Figure 2.20-ⓓ), autour de la courbe de la valeur médiane, correspond aux valeurs entre le premier et le troisième quartile. Ce contexte permet de voir comment l'entité se situe par rapport aux autres. Dans l'exemple fourni, l'entité est quasiment toujours au dessus du troisième quartile à quelques exceptions près, notamment en fin juin 2021 et fin mai 2022. L'utilisateur peut sélectionner différentes options (Figure 2.20-ⓔ), pour adapter le graphique affiché. La première option, *dataset/area*, permet à l'utilisateur de choisir la plage de données affichée sur l'axe des ordonnées, soit celle de l'entité qui coïncide avec le tooltip, soit celle du jeu de données. Avec *linear/log*, l'utilisateur peut choisir entre une échelle linéaire et une échelle logarithmique. La dernière option, *7-day avg.*, permet d'obtenir une courbe lissée en calculant la valeur moyenne sur une fenêtre glissante de 7 jours. Au survol du graphique, une ligne verticale est affichée (Figure 2.20-ⓖ) en conjonction avec l'axe des abscisses pour le temps. Elle permet à l'utilisateur de parcourir en détail les valeurs de la variable sélectionnée au niveau de l'entité géographique. En partant du haut vers le bas, les points correspondent à la valeur maximale, à la valeur de l'entité, au troisième quartile, à la médiane, au premier quartile et à la valeur minimale. En fonction de l'entité sélectionnée, de la variable et de la date, la valeur de l'entité peut évidemment varier et n'est pas nécessairement comprise entre le troisième quartile et la valeur maximale. Pour avoir les valeurs exactes, il suffit de se référer à Figure 2.20-ⓗ. La partie supérieure fait référence à la date sélectionnée dans Figure 2.18-ⓕ correspondant à la date pour laquelle les valeurs affichées sur la carte coïncident. La partie inférieure montre les valeurs de la date sélectionnée lors du survol du tooltip. La ligne verticale Figure 2.20-ⓕ est très similaire à Figure 2.20-ⓖ mais correspond à la date sur la carte. Elle est mise à jour en fonction des modifications apportées à la timeline (Figure 2.18-ⓕ). Si les utilisateurs souhaitent effectuer une analyse plus appro-

fondie des données sélectionnées, ils peuvent facilement les télécharger au format CSV en cliquant sur l'icône de téléchargement en haut à gauche (Figure 2.20-ⓘ).

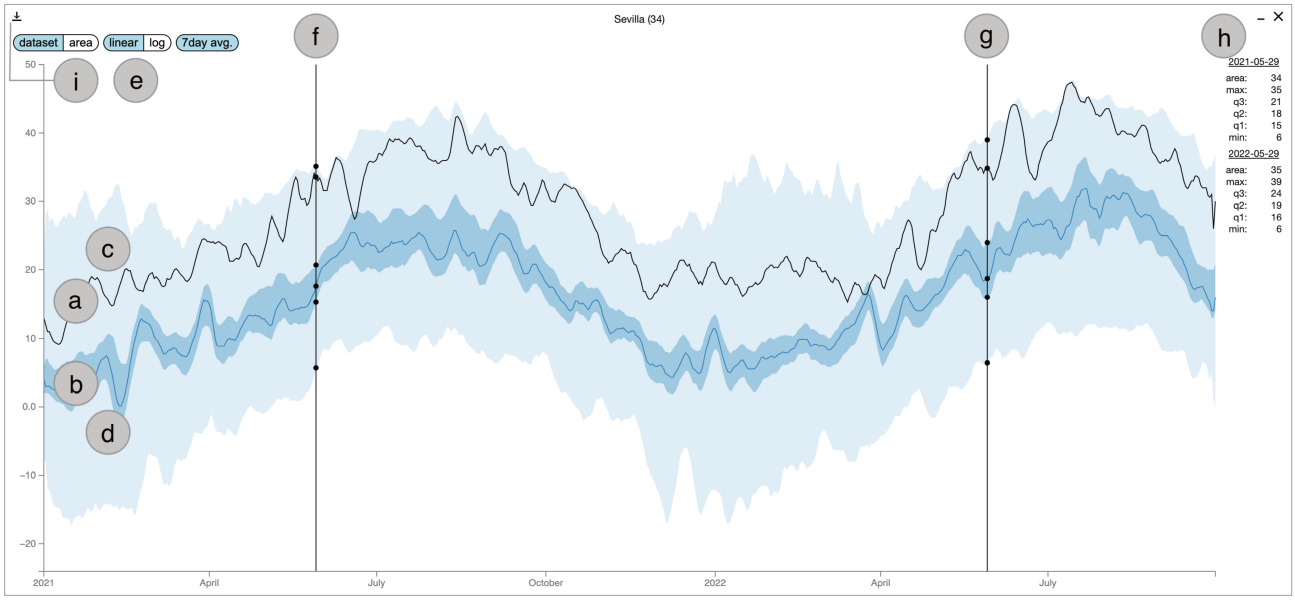


FIGURE 2.20 – Exemple d'un tooltip avec ses principales composantes.

En résumé, nous proposons une approche consistant à juxtaposer deux vues et se basant sur un système de synchronisation. Cette visualisation facilite les démarches de comparaisons de données et, au même titre, d'exploration de données grâce à des fonctionnalités supplémentaires.

2.3.2 Application avec Epid Data Explorer

Dans le cadre du projet européen MOOD, une plateforme interactive de visualisation de données épidémiologiques a été développée. Le projet MOOD est dédié à la détection, évaluation et suivi des maladies infectieuses en Europe dans un contexte de changement des mobilités, populations, environnemental et climatique. L'analyse de données spatio-temporelles est un défi majeur dans la recherche épidémiologique. Différentes approches d'analyse de données sont possibles afin d'extraire des informations utiles. Malgré l'efficacité de ces analyses, il est primordial d'avoir pris connaissance des informations recherchées. C'est là qu'intervient la visualisation et l'importance de développer une plateforme conçue pour cette fonction. L'exploration des données est très propice à l'élaboration de nouvelles hypothèses notamment sur des potentielles corrélations expliquant l'émergence d'une maladie infectieuse. C'est une étape essentielle, mais négligée, avant d'effectuer des analyses afin de les rendre pertinentes ; autant au niveau de l'axe de recherche que du choix de l'approche d'analyse elle-même.

Une discussion avec les utilisateurs finaux a montré que les fonctionnalités de comparaison répondant aux besoins B1 et B2 seraient utiles dans cette plateforme. Elle a également montré que d'autres besoins se présentaient. Nous avons adapté notre approche de visualisation avec deux vues à l'épidémiologie pour créer l'application Epid Data Explorer (EDE). Afin de démontrer l'utilité de EDE, deux cas d'usage sont proposés en fin de section.

2.3.2.1 Contexte et besoins additionnels

Le développement de l'outil Epid Data Explorer est basé sur un ensemble d'exigences visant à répondre aux besoins des utilisateurs finaux. Ces besoins portent à la fois sur la fonctionnalité de la plateforme et sur le système lui-même. Les utilisateurs finaux ont exprimé le souhait de visualiser plusieurs jeux de données au sein d'une plateforme unifiée et d'identifier les corrélations potentielles entre les variables par le biais d'une analyse comparative. Ils souhaitaient également pouvoir suivre la progression des valeurs des variables dans le temps. Pour faciliter ces tâches, la plateforme doit permettre des comparaisons simultanées au sein d'une même vue. Ces exigences correspondent aux besoins B1 et B2. Un autre besoin additionnel a été relevé : les utilisateurs finaux souhaitent avoir accès à différentes granularités spatiales et temporelles mais aussi pouvoir agréger les données dans certains cas. En effet, tous les jeux de données n'ont pas les mêmes granularités spatiales et temporelles mais ils souhaitent tout de même avoir la possibilité de les comparer entre eux. En outre, comme certaines données peuvent être sensibles et nécessiter un stockage confidentiel, l'outil doit garantir de solides mesures de sécurité des données. Cela inclut la possibilité pour les utilisateurs d'importer leurs propres jeux de données tout en maintenant une stricte confidentialité. Sur la base de ces exigences, les besoins supplémentaires suivants ont été identifiés pour l'outil Epid Data Explorer :

Besoin B3. Un utilisateur doit avoir accès à différentes granularités spatiales et temporelles. Certains jeux de données peuvent avoir directement différentes échelles géographiques comme par exemple pays, région ou bien sous-région. Si les données ne sont pas au niveau maximal, il doit être possible d'effectuer des agrégations. Ainsi, si par exemple un jeu de données comporte des valeurs par région, pour chaque pays les valeurs peuvent être agrégées afin que l'utilisateur puisse visualiser les données, par pays et région s'il le souhaite. L'idée est exactement la même pour la granularité temporelle avec comme niveaux par exemple jour, semaine et mois.

Besoin B4. Les utilisateurs doivent pouvoir comparer leurs propres données entre elles et avec les données publiques, si nécessaire.

Besoin B5. La gestion des données publiques et privées à intégrer dans la plateforme doit être sécurisée et simplifiée.

Il existe deux types de plateformes de visualisation des données en épidémiologie. Le premier type concerne les plateformes qui se concentrent sur la surveillance basée sur les événements. Celles-ci utilisent des systèmes automatisés pour visualiser des rapports non structurés sur les maladies associées aux événements signalés. Le texte de ces rapports est traité pour extraire des informations telles que la maladie, l'espèce impliquée, la date et le lieu de l'événement signalé. Le second type de plateformes se rapporte aux plateformes spécialisées dans la surveillance basée sur des indicateurs. Des données statistiques ou d'autres données structurées sont directement visualisées sans qu'il soit nécessaire de les traiter pour en extraire des informations. Nous nous intéressons à ce second type : les plateformes de visualisation de données épidémiologiques basées sur les indicateurs. Il en existe plusieurs [1] mais aucune d'entre elles, à notre connaissance, ne satisfait l'ensemble des besoins. Elles sont répertoriées dans le Tableau 2.2.

1. Epi Visualization (https://vizhub.healthdata.org/epi/), Gapminder (http://www.gapminder.org/world), Empres-i (https://empres-i.apps.fao.org), Surveillance Atlas of Infectious Diseases (http://atlas.ecdc.europa.eu/public/index.aspx), COVID-19 Vaccine Tracker (https://vaccinetracker.ecdc.europa.eu), COVID-19 Dashboard (https://coronavirus.jhu.edu/map.html), COVID-19 Data Explorer (https://ourworldindata.org/explorers/coronavirus-data-explorer), Nextstrain (https://nextstrain.org). [dernière consultation pour l'ensemble des plateformes le 04/08/2023].

Tableau 2.2 – Réponses aux exigences des plateformes basées sur des indicateurs.

	B1	*B2*	*B3*	*B4*	*B5*
Epi visualization	✗	✓	✗	✗	✗
Gapminder [77]	✓	✓	✗	✓	✓
Empres-i [16]	✓	✓	✗	✓	✓
Surveillance Atlas of Infectious Diseases (ECDC)	✗	✓	✗	✗	✗
COVID-19 Vaccine Tracker (ECDC)	✗	✓	✗	✗	✗
COVID-19 Dashboard [22]	✗	✓	✗	✗	✗
COVID-19 Data Explorer (Our World in Data)	✗	✓	✗	✗	✗
Nextstrain [40]	✓	✓	✗	✓	✓
Epid Data Explorer	✓	✓	✓	✓	✓

2.3.2.2 Présentation

La page d'accueil de EDE propose une liste de jeux de données accessibles [2]. Avant d'accéder aux visualisations, l'utilisateur doit sélectionner les jeux de données qui l'intéressent. Pour ce faire, la liste des jeux de données (Figure 2.21-ⓐ) et la liste des groupes de jeux de données à droite (Figure 2.21-ⓑ) sont disponibles. Ces groupes permettent d'assembler les jeux de données par thème. À noter qu'un jeu de données ne fait pas nécessairement partie d'un groupe et qu'en sélectionnant un groupe, tous les jeux de données concernés sont automatiquement sélectionnés. La page d'accueil explicite aussi l'objectif et le principe de EDE (Figure 2.21-ⓒ). Un menu est accessible grâce au bouton Figure 2.21-ⓓ situé dans le coin supérieur gauche de la plateforme. À partir de ce menu, il est possible d'accéder aux visualisations (il est nécessaire d'avoir sélectionné des jeux de données et d'avoir validé ce choix au préalable), un tutoriel et aux sources de données.

Après avoir validé le choix des jeux de données à visualiser, la visualisation avec deux vues remplace la page d'accueil (cf. Figure 2.22a). Une seconde visualisation est disponible, elle a une seule vue et donc une unique carte comme le montre la Figure 2.22b. Un utilisateur peut avoir besoin de se concentrer sur une seule chose ; dans ce cas, la visualisation avec une unique vue est plus appropriée. Cela permet d'obtenir une vue d'ensemble d'une variable sans l'encombrement d'une deuxième vue. Le passage d'une visualisation avec deux vues à une visualisation avec une seule vue et vice versa est très simple. Dans le premier cas, il suffit de cliquer sur le bouton situé dans le coin supérieur gauche ou droit en fonction de la carte sélectionnée (Figure 2.22a-ⓐ ou Figure 2.22a-ⓑ). Passer d'une visualisation avec deux vues à une visualisation avec une seule vue consiste tout simplement à agrandir l'une des deux cartes. Dans le second cas, pour ajouter une deuxième carte, il suffit de cliquer sur le bouton (Figure 2.22b-ⓐ) situé dans le coin supérieur gauche ou droite selon la carte affichée.

EDE reprend les fonctionnalités de l'approche proposée consistant à avoir une visualisation avec deux vues (cf. Section 2.3.1.2). Les tooltips présentés dans la Section 2.3.1.3 sont également présents dans la plateforme. Avec ces fonctionnalités, EDE répond donc aux be-

2. Les données utilisées pour démontrer la fonctionnalité de EDE dans sa version publique (https://advanse.lirmm.net/EDE/epid-data-explorer/) proviennent de l'ECDC (https://www.ecdc.europa.eu/), Our World in Data (https://ourworldindata.org), Google (https://www.google.com/covid19/mobility/), Government Response Tracker (https://www.bsg.ox.ac.uk/research/COVID-19-government-response-tracker), Obepine (https://www.reseau-obepine.fr) et Historique meteo (https://www.historique-meteo.net) [dernière consultation de l'ensemble des sources le 04/08/2023].

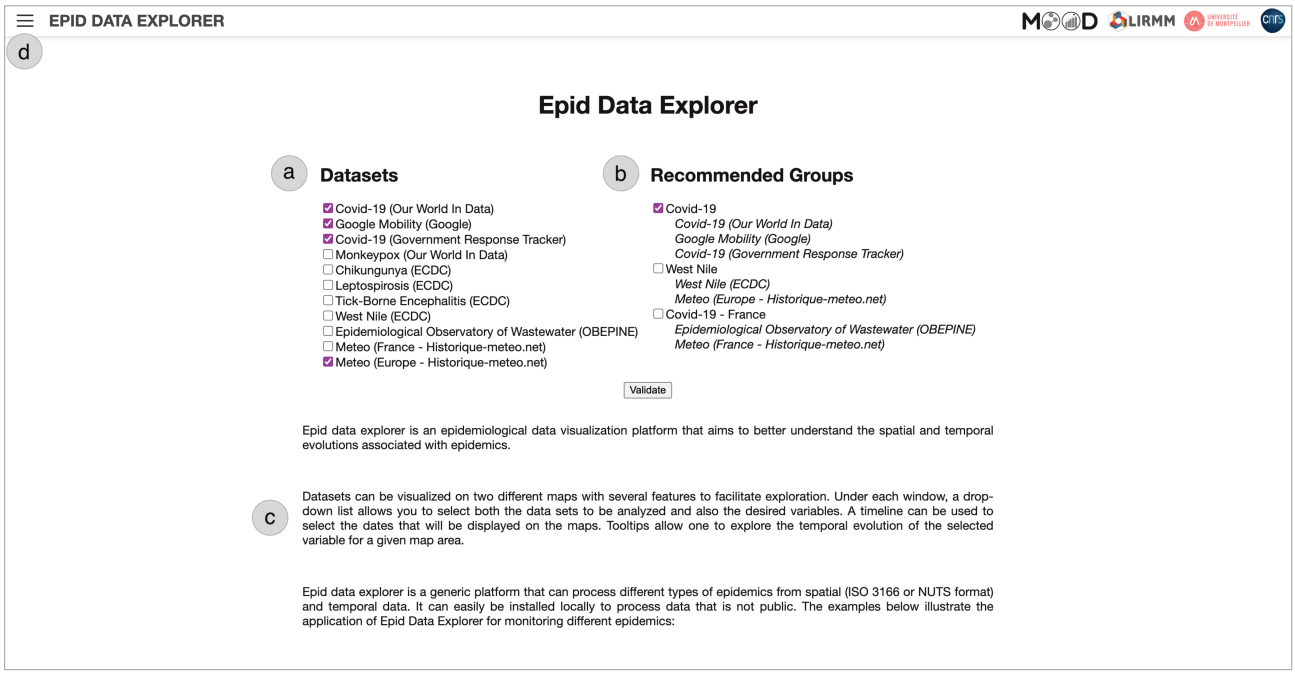
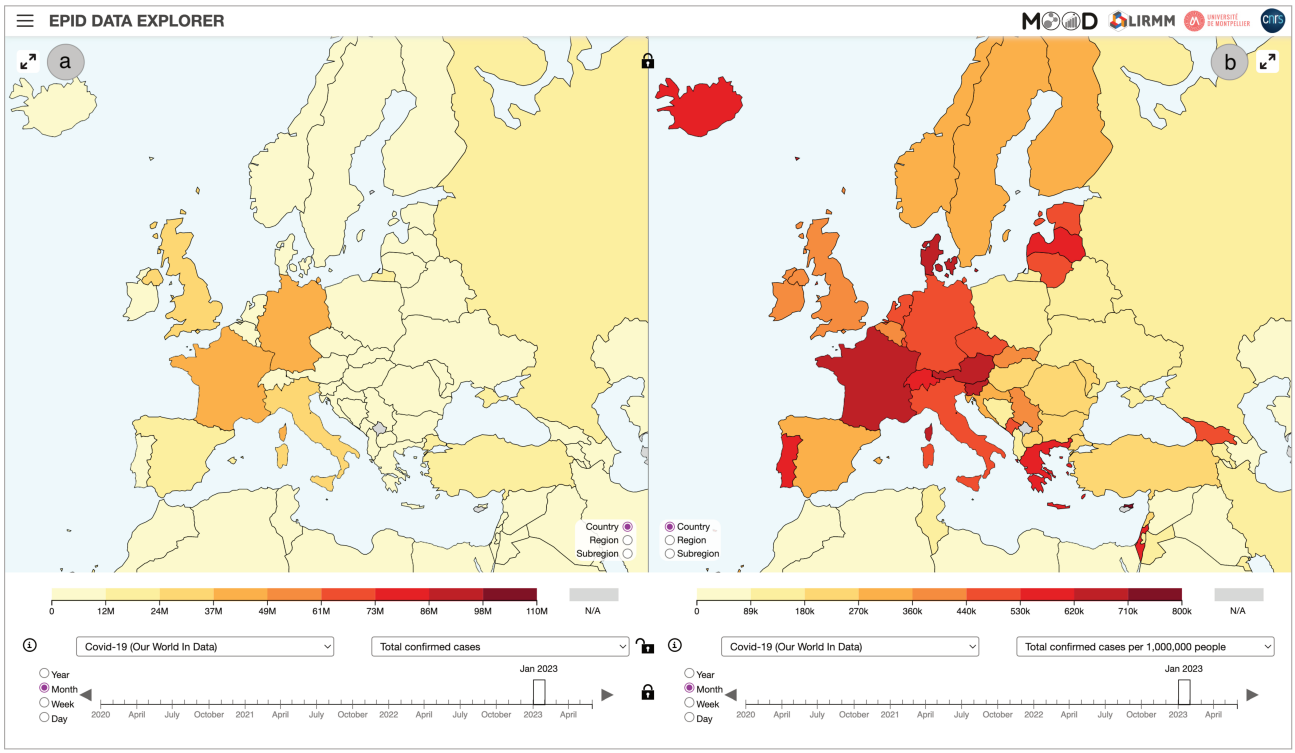


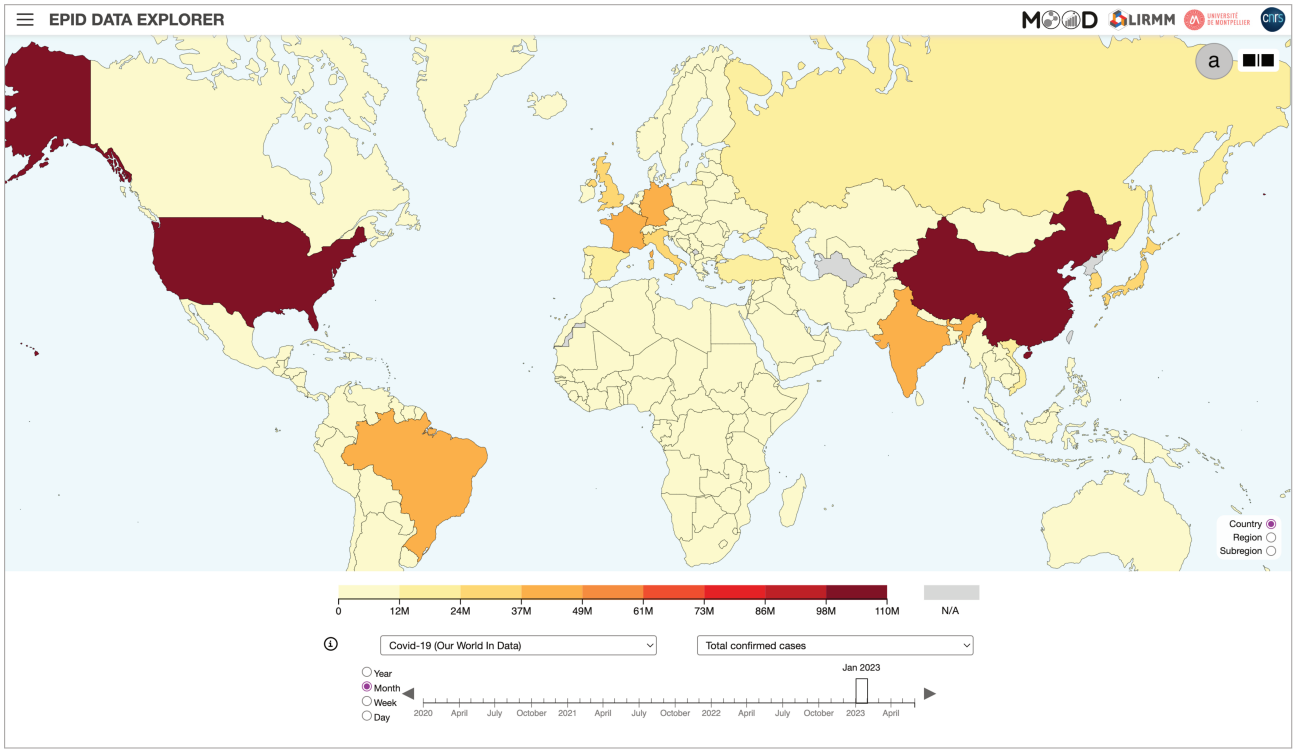
FIGURE 2.21 – Page d'accueil de la plateforme Epid Data Explorer.

soins B1 et B2. D'autres éléments ont été intégrés dans les deux visualisations proposées, ils sont illustrés dans la Figure 2.23. On retrouve une carte dans laquelle on peut naviguer (Figure 2.23-ⓐ) et dont l'échelle géographique peut être modifiée (Figure 2.23-ⓑ). Selon la nomenclature utilisée pour identifier les entités géographiques les granularités ne sont pas les mêmes. Nous prenons en compte la nomenclature ISO-3166 [3] qui englobe tous les pays du monde, nous utilisons comme niveau les pays, leurs régions et sous-régions (*Country*, *Region* et *Subregion*). La seconde nomenclature est NUTS [4], celle-ci ne concerne que l'Europe mais a plus de niveaux : *nuts0*, *nuts1*, *nuts2* et *nuts3*. Par défaut la granularité est au niveau des pays (*Country* pour ISO-3166, *nuts0* pour NUTS). Il est important de noter que le niveau de granularité géographique affiché sur la carte dépend des données. La carte peut afficher des valeurs au niveau fourni dans le jeu de données et aux niveaux supérieurs si l'agrégation a été effectuée lors de l'importation des données. En fournissant un accès à différentes granularités spatiales et à des données agrégées, EDE répond à l'exigence B3. La légende correspondant à la carte est affichée juste en dessous (Figure 2.23-ⓒ), elle s'adapte automatiquement aux données sélectionnées selon la valeur minimale et maximale mais aussi selon le type de données. Trois types de données sont pris en charge : quantitatif continu, quantitatif divergent et qualitatif ordinal. La liste déroulante pour sélectionner la variable à visualiser est toujours présente (Figure 2.23-ⓔ) mais en complément une seconde liste déroulante a été ajoutée. Celle-ci concerne la sélection du jeu de données (Figure 2.23-ⓓ). En effet, comme nous l'avons vu plus tôt lors de la présentation de la page d'accueil, EDE peut prendre en charge plusieurs jeux de données qui sont chacun composé d'une ou plusieurs variables. Juste en dessous de ces listes déroulantes est positionnée la timeline (cf. Figure 2.23-ⓕ). Avec EDE, plusieurs granularités temporelles sont accessibles. Elle peut être modifiée à partir des boutons radios présents sur la gauche de la timeline (Figure 2.23-ⓖ), les niveaux disponibles sont jour, semaine, mois et année (cf. Figure 2.24). La timeline s'adapte selon le niveau de granularité temporel sélectionné. Cela permet de répondre au besoin B3.

3. https://www.iso.org/iso-3166-country-codes.html [dernière consultation le 04/08/2023]
4. https://ec.europa.eu/eurostat/web/nuts/background [dernière consultation le 04/08/2023]



(a) Visualisation avec deux vues.



(b) Visualisation avec une seule vue.

FIGURE 2.22 – Les deux visualisations disponibles sur EDE.

L'application EDE est disponible en ligne avec des données publiques [5]. Cependant, pour répondre aux besoins B4 et B5, il est nécessaire d'installer une version locale d'EDE [6]. Suite

5. https://advanse.lirmm.net/EDE/epid-data-explorer/ [dernière consultation le 04/08/2023]
6. https://gite.lirmm.fr/advanse/EDE/epid-data-explorer [dernière consultation le 04/08/2023]

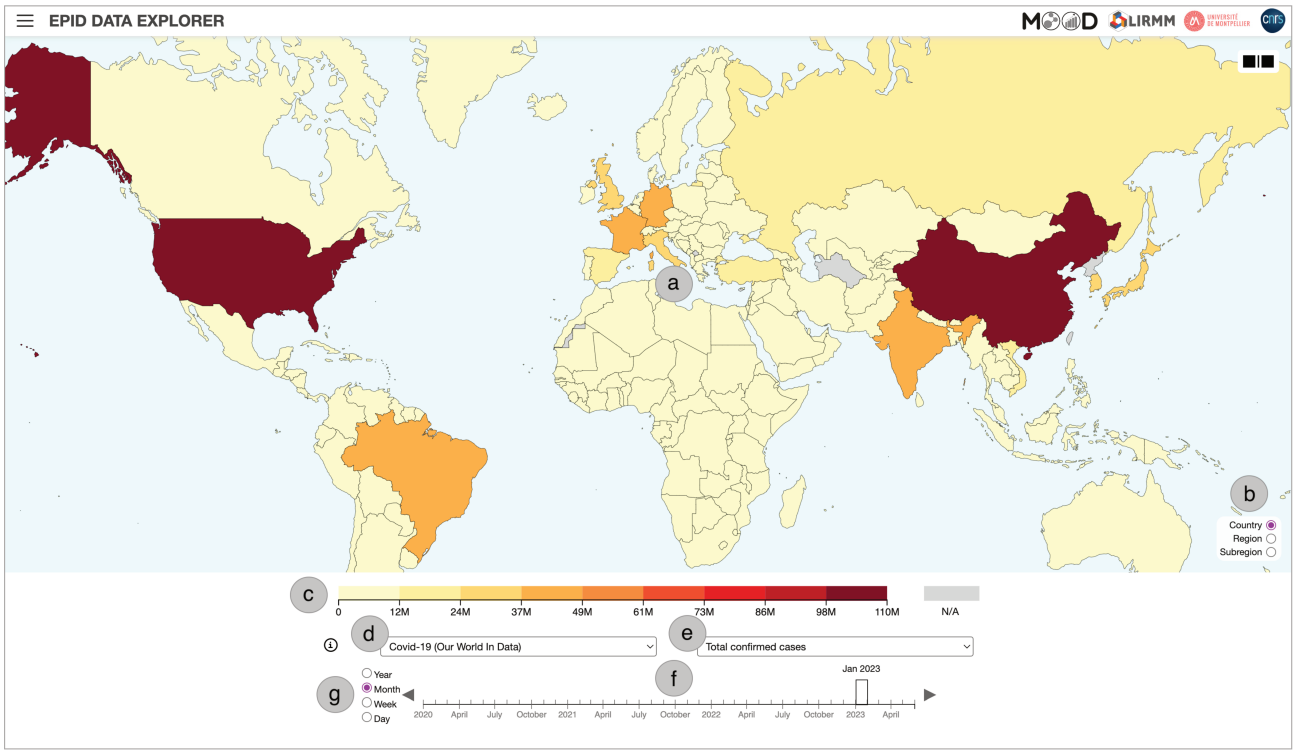
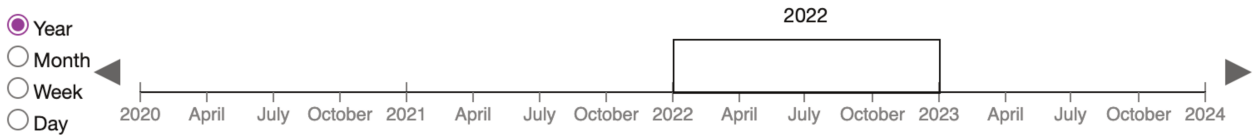
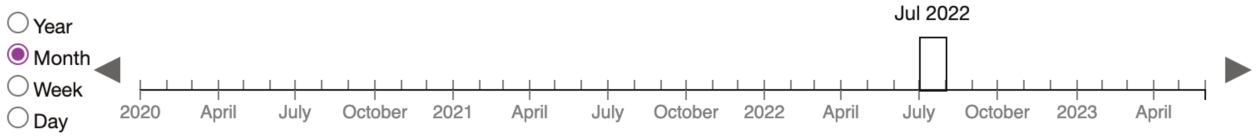


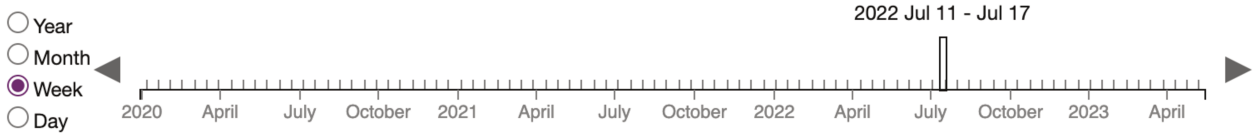
FIGURE 2.23 – Éléments d'une vue dans EDE.



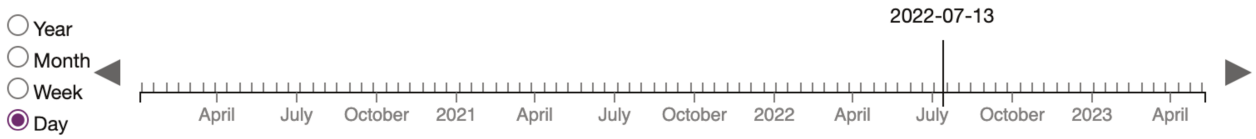
(a) Timeline par année.



(b) Timeline par mois.



(c) Timeline par semaine.



(d) Timeline par jour.

FIGURE 2.24 – Timelines disponibles en fonction de la granularité temporelle sélectionnée.

à l'installation de EDE, une nouvelle interface appelée EDE Datasource Manager est fournie (cf. Figure 2.25). Cette interface permet non seulement aux utilisateurs d'installer et d'exécu-

ter EDE localement, mais aussi d'ajouter facilement de nouveaux jeux de données. Les utilisateurs peuvent simplement spécifier l'adresse où les données sont disponibles et peuvent aussi très facilement ajouter ou supprimer des jeux de données. Lorsqu'ils accèdent à des données publiques, les utilisateurs n'ont qu'à fournir l'adresse de la source de données. Pour permettre aux utilisateurs de travailler avec leurs propres données, il est nécessaire qu'ils spécifient comment EDE Datasource Manager peut accéder à leurs données. À cette fin, une interface supplémentaire appelée EDE DataStore est proposée[7] (cf. Figure 2.25). Cette interface permet aux utilisateurs d'intégrer facilement des données locales au format CSV et d'effectuer diverses tâches de traitement des données telles que l'agrégation géolocalisée, le formatage des données et la sélection des données pertinentes. Une fois le traitement nécessaire effectué, l'interface rend les données disponibles et l'adresse fournie peut être utilisée par EDE Datasource Manager. Grâce à ces deux interfaces, exécutées localement, nous répondons aux exigences B4 et B5.

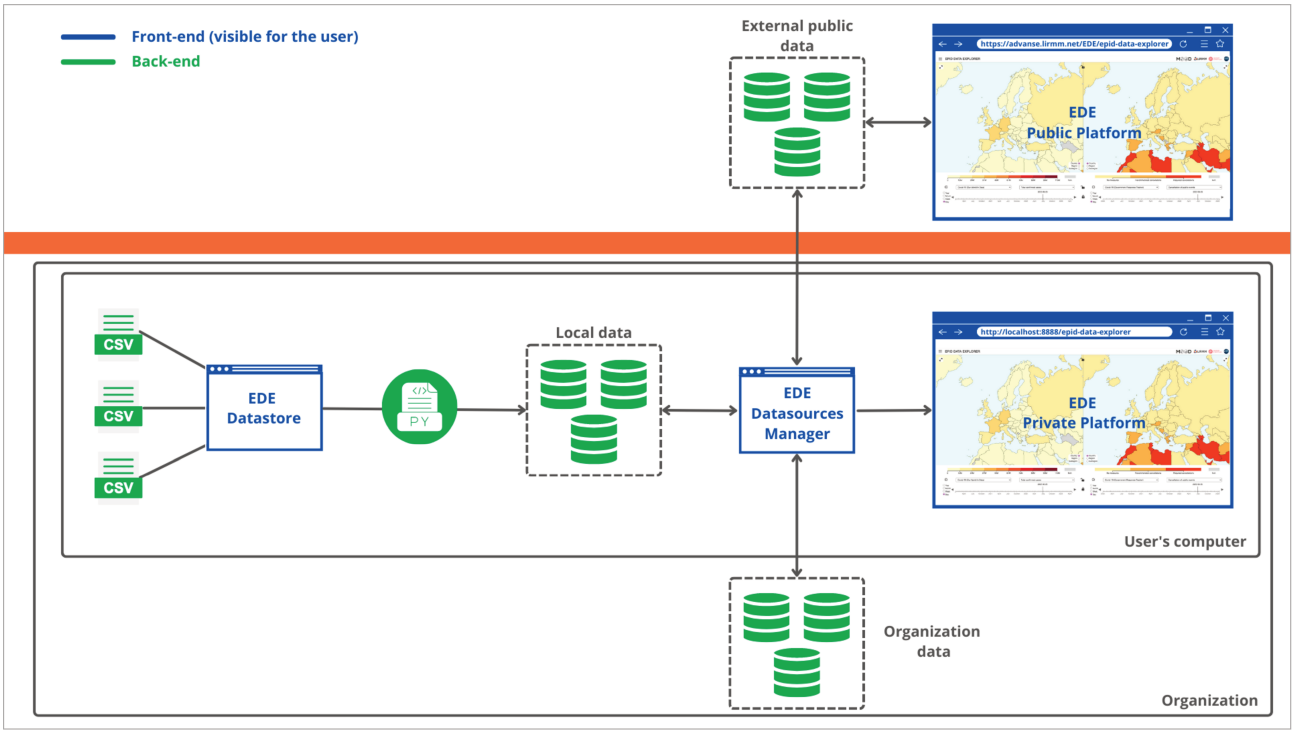


FIGURE 2.25 – Organisation des jeux de données dans EDE.

2.3.2.3 Cas d'usage

Nous proposons deux cas d'usage pour démontrer la pertinence des fonctionnalités de visualisation d'EDE.

Le premier compare une variable concernant les mesures de confinement mises en œuvre durant la pandémie de COVID-19 avec une autre relative à l'évolution des mobilités dans les transports en commun. L'objectif de ce cas d'usage est de montrer, à travers un exemple très simple, qu'il est possible d'observer des phénomènes avec les visualisations d'EDE en un coup d'œil.

Durant la pandémie de COVID-19, différents pays dans le monde ont mis en place des restrictions pour lutter contre la propagation du virus. Ces mesures ont affecté de manière significative la vie quotidienne des personnes, notamment en termes de mobilité. Pour ce cas

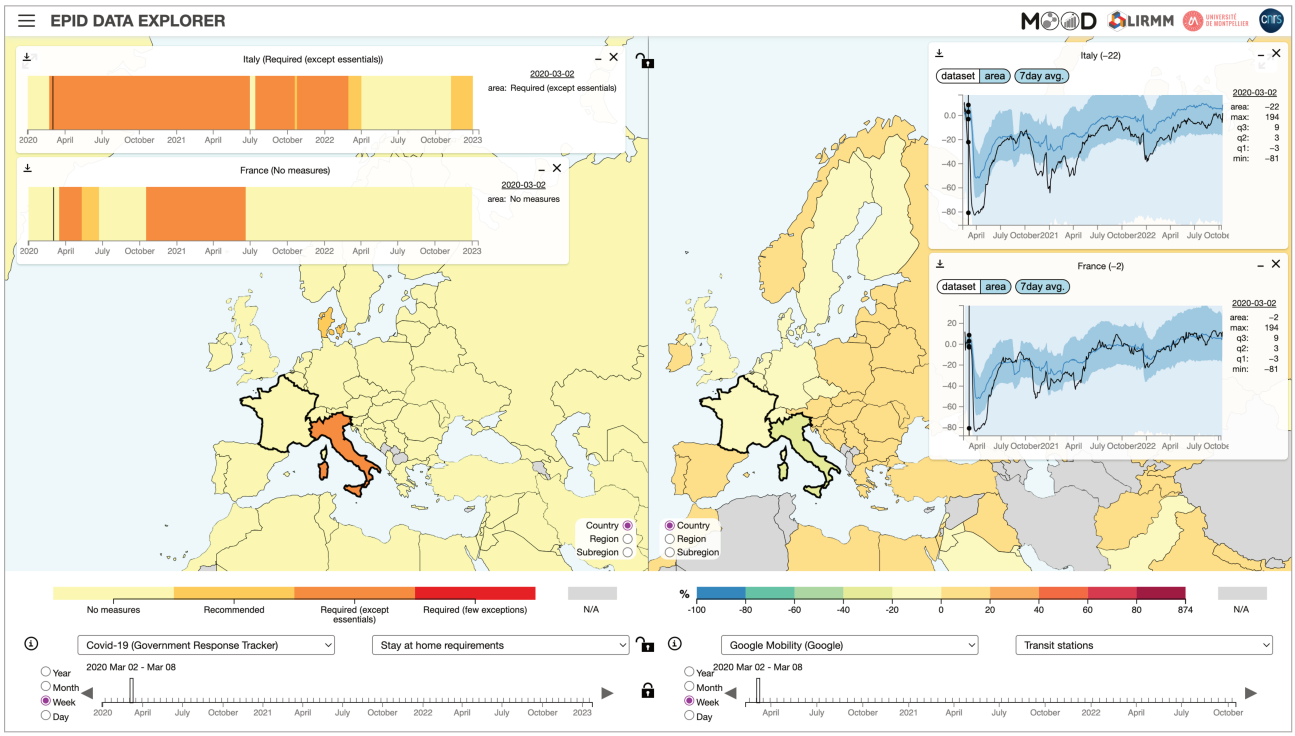
7. https://gite.lirmm.fr/advanse/EDE/dataset [dernière consultation le 04/08/2023]

d'usage, nous nous concentrerons sur l'Europe et la mise en place de ces restrictions. L'utilisation d'EDE nous permet d'observer ce phénomène très simplement, comme le montre la Figure 2.26. Deux périodes distinctes sont représentées : la semaine du 2 au 8 mars 2020 et la semaine du 23 au 29 mars 2020. Dans la Figure 2.26a, qui correspond à la première période, l'Italie est le seul pays européen à avoir mis en place des mesures de confinement (le Danemark ne faisait que des recommandations) comme l'illustre la carte de gauche. En ce qui concerne l'utilisation des transports publics (carte de droite), certains pays comme la France, l'Allemagne et le Royaume-Uni ont connu une légère baisse, tandis que d'autres ont enregistré une légère hausse (Espagne, Portugal, Irlande, Norvège, etc.). L'Italie avait déjà connu une baisse de l'utilisation des transports publics à ce moment-là, en raison des restrictions mises en œuvre. Au cours de la semaine du 23 au 29 mars, de nombreux pays ont mis en place un confinement entraînant un impact généralisé et sans précédent sur la mobilité. Ce phénomène peut être facilement observé dans la Figure 2.26b. En visualisant les deux cartes côte à côte, nous pouvons évaluer l'influence des restrictions mises en œuvre sur la mobilité dans les pays européens. En outre, nous pouvons comparer l'évolution d'une même variable dans plusieurs pays au cours de la même période. Dans la Figure 2.26, les tooltips pour l'Italie et la France sont ouverts sur toutes les cartes, ce qui permet de faire une comparaison des tendances entre ces deux pays. Sur les cartes de gauche, nous pouvons observer que l'Italie a mis en œuvre des mesures plus tôt et les a maintenues pendant une période prolongée, tandis que la France a connu deux périodes distinctes : de mars 2020 à l'été, et à partir de la fin de l'année 2020. Une tendance similaire de l'utilisation des transports publics peut être observée dans les deux pays, l'Italie connaissant des baisses plus fréquentes et plus importantes, se situe souvent en dessous de la valeur médiane.

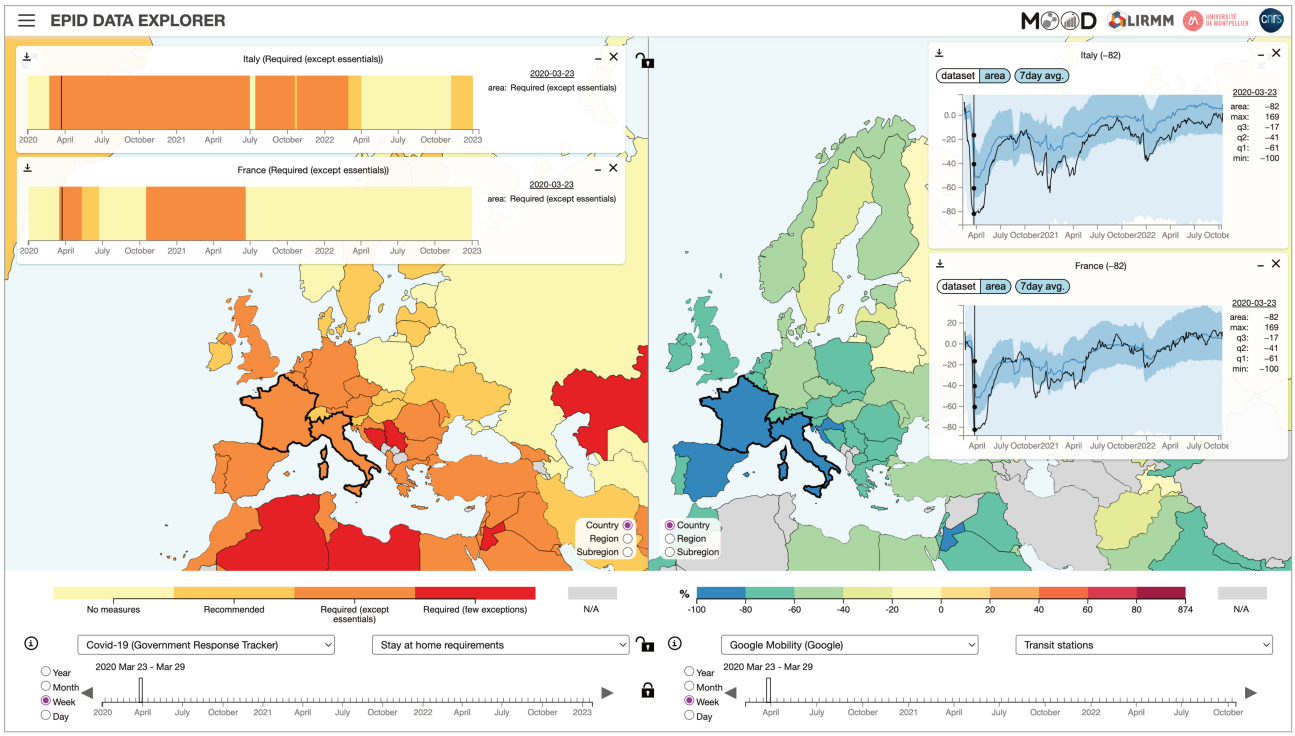
Finalement, ce premier cas d'utilisation illustre l'efficacité de la plateforme pour mettre en évidence des phénomènes et faciliter les comparaisons entre deux variables distinctes au cours de la même période, ainsi qu'entre différentes zones géographiques.

Le second cas d'usage est quant à lui basé sur l'hypothèse d'une corrélation entre la transmission de COVID-19 et les conditions météorologiques. En effet, de nombreuses études ont examiné en détail les corrélations entre les conditions météorologiques et le COVID-19. Par exemple, Ganslmeier et al. [31] ont analysé l'impact des conditions météorologiques sur la pandémie à l'aide d'un ensemble complet de données comprenant plus de 1,2 million d'observations quotidiennes sur 3700 comtés dans neuf pays pendant toutes les saisons de 2020. Dans une autre étude, Majumder et al. [61] ont réalisé une méta-analyse pour étudier les corrélations significatives entre les paramètres météorologiques (température, humidité, précipitations, rayonnement ultraviolet, vitesse du vent) et le COVID-19. Ils ont examiné 11 articles qui ont étudié en profondeur ce sujet en utilisant des jeux de données provenant de plus de 110 pays. Les résultats ont mis en évidence une corrélation notable entre la température, l'humidité et la vitesse du vent, d'une part, et le taux de mortalité et l'incidence du COVID-19, d'autre part. Dans une étude similaire, McClymont et al. [62] ont analysé 23 articles après un processus de sélection rigoureux. Ils ont évalué la conception de l'étude, la collecte et l'analyse des données, ainsi que les discussions et les limites. Les chercheurs ont constaté que la température était souvent considérée comme un facteur important dans de nombreuses études, avec une augmentation de l'incidence de COVID-19 à mesure que la température diminuait, et l'incidence la plus élevée observée dans les zones où les températures étaient plus basses. En outre, ils ont noté une association significative entre l'humidité et l'incidence du COVID-19, bien que les résultats rapportés varient d'une étude à l'autre, avec des corrélations à la fois positives et négatives.

EDE peut être utilisé pour analyser les données et identifier initialement les corrélations potentielles qui pourraient nécessiter une analyse plus approfondie. Par exemple, si l'utili-



(a) Visualisations pour la semaine du 2 au 8 mars 2020.



(b) Visualisations pour la semaine du 23 au 29 mars 2020.

FIGURE 2.26 – Visualisations des restrictions de confinement (cartes de gauche) et de l'évolution de la mobilité dans les stations de transport en commun (cartes de droite).

sateur souhaite examiner les corrélations entre l'humidité et le taux d'incidence en France, il peut récupérer des données météorologiques à partir de sources telles que le site web *Historique meteo* [8], qui fournit des informations météorologiques quotidiennes avec diverses

8. https://www.historique-meteo.net [dernière consultation le 04/08/2023]

variables. Le site *data.gouv*[9] propose une synthèse des indicateurs de suivi de l'épidémie de COVID-19 en France. Grâce à l'interface EDE Data Store, les données peuvent être intégrées sans effort. Pour ce faire, les données originales subissent une étape de prétraitement au cours de laquelle elles sont transformées à l'aide de nomenclatures telles que NUTS. Les données intégrées peuvent ensuite être incorporées de manière transparente dans la plateforme par l'intermédiaire de l'interface EDE Data Store. Les données disponibles sont au niveau départemental (NUTS3), et les utilisateurs ont la possibilité de sélectionner la fonction d'agrégation pour chaque variable afin de passer à des niveaux supérieurs tels que la région ou le pays. La Figure 2.27 illustre comment EDE permet de comparer les données météorologiques et la COVID-19 en France pendant la troisième vague de la pandémie afin d'identifier les corrélations potentielles. Elle permet d'analyser des variables telles que la température, l'humidité et la vitesse du vent en relation avec le taux d'incidence ou tout autre variable choisie dans diverses régions géographiques et au fil du temps. La carte de gauche de la Figure 2.27 indique l'humidité maximale enregistrée, tandis que la carte de droite représente le taux d'incidence.

Pour terminer, ce deuxième cas d'utilisation illustre l'utilité d'importer ses propres données pour explorer et observer les corrélations possibles. Ces observations peuvent ensuite conduire à une analyse plus approfondie et optimisée des données en sélectionnant l'approche appropriée. L'exemple présenté ici illustre les pistes de recherche qui peuvent être ouvertes en visualisant des données importées selon les besoins de l'utilisateur.

Ainsi, nous proposons une approche de visualisation se basant sur la juxtaposition pour comparer des données associées à des entités géographiques. Cette approche met en avant les principaux avantages de la juxtaposition tout en la bonifiant en y intégrant un système de synchronisation. Celui-ci fluidifie la navigation et permet d'avoir uniquement deux cartes juxtaposées sur une même fenêtre. Le lecteur ne reçoit donc pas un surplus d'informations mais n'a pas non plus la nécessité de solliciter sa mémoire en navigant sur différentes fenêtres. C'est un bon compromis et l'approche s'est montrée bien efficace et utile en l'implémentant dans la plateforme d'exploration de données épidémiologiques EDE.

2.4 Comparaison par superposition

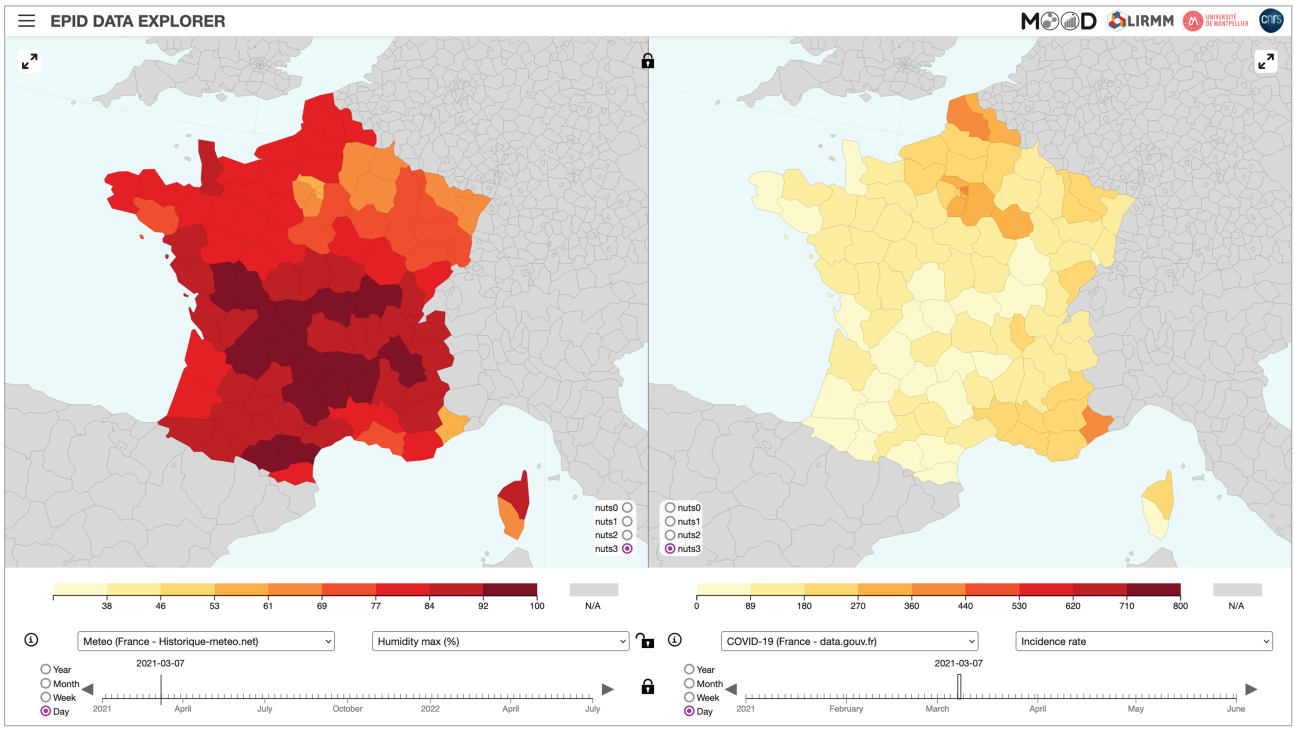
Dans cette section, nous nous centrons sur la comparaison de données par superposition. Pour rappel, dans le cadre de ce mémoire et de la visualisation de données associées à des entités géographiques, la superposition renvoie à la représentation de deux variables sur une même carte. L'objectif ici est de comparer des cartes bivariées sur des tâches spécifiques afin de déterminer quelle technique serait la plus efficace.

Pour cela, nous exposons dans un premier temps les techniques de représentation bivariées et les tâches identifiées. Ensuite, l'expérimentation est présentée et, enfin, nous décrivons les résultats obtenus pour terminer.

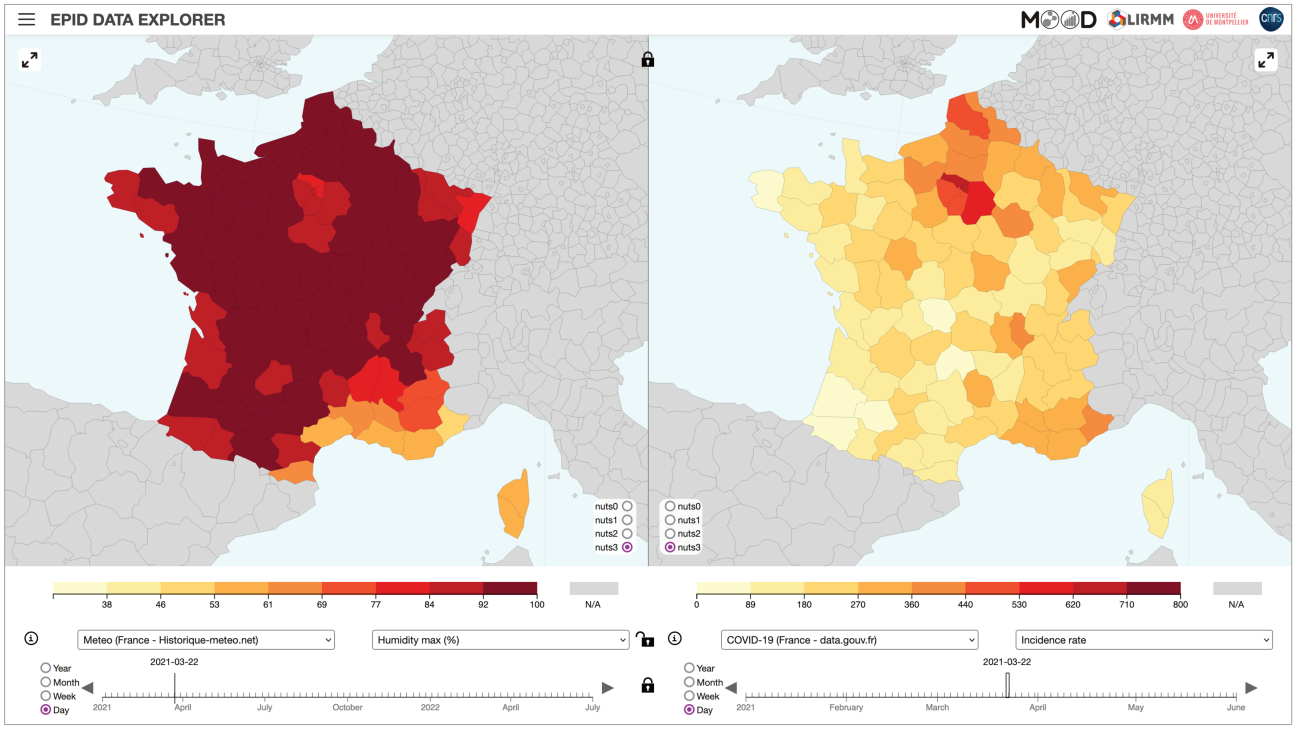
2.4.1 Comparer des représentations cartographiques bivariées

L'objectif ici est de mettre en place une étude utilisateur donnant lieu à une comparaison de technique de représentation cartographique bivariée. Cette comparaison permet de déterminer quel type de carte parmi les 5 décrits ci-dessous est le plus efficace pour effectuer

9. https://www.data.gouv.fr/fr/datasets/synthese-des-indicateurs-de-suivi-de-lepidemie-covid-19/ [dernière consultation le 04/08/2023]



(a) Visualisation pour la date du 07/03/2021



(b) Visualisation pour la date du 22/03/2021

FIGURE 2.27 – Visualisations de deux variables : une variable météorologique (humidité maximale) et une autre concernant la COVID-19 (taux d'incidence) durant la troisième vague de la pandémie en France.

une liste spécifique de tâches impliquant deux variables quantitatives géolocalisées. Nos résultats permettront de fournir des conseils en conception de cartes aux personnes peu familiarisées avec la visualisation de données, et qui souhaitent diffuser facilement des cartes à un public diversifié. Dans cette section, nous présentons d'abord les techniques de visua-

lisation impliquées, avant de passer en revue les tâches qui nous ont permis de les évaluer.

2.4.1.1 Techniques de visualisation

Les méthodes de cartographie univariées de variables quantitatives ont été présentées dans la Section 2.2.1.3 : cartes à symboles (*Sy*), cartes Choroplèthes (*Ch*) et cartogrammes (*Ca*). Ces dernières sont construites à partir d'une variable visuelle (couleur ou taille) et d'une implantation (ponctuelle ou zonale) comme illustré dans le Tableau 2.1. Nous avons basé notre choix de techniques de représentation bivariées à comparer à partir de combinaisons entre ces trois principaux types de carte :

- *SySy* est une carte de symboles dans laquelle la couleur des symboles représente la première variable et la taille des symboles représente la deuxième variable,
- *SyCa* combine une carte de symboles dans laquelle la taille du symbole représente la première variable et un cartogramme représente la deuxième variable,
- *ChSy* combine une carte choroplèthe représentant la première variable avec une carte de symboles dans laquelle la taille du symbole représente la deuxième variable,
- *ChCa* combine une carte choroplèthe représentant la première variable avec un cartogramme représentant la deuxième variable,
- *ChCh* est une carte choroplèthe bivariée basée sur le modèle *diagonal* de couleur proposé par Strode et al. [85]. Ce schéma est adapté pour représenter la relation directe entre les deux variables.

La Figure 2.28 résume l'ensemble des combinaisons et les cartes bivariées qui en résultent. Nous n'avons pas considéré la combinaison impliquant uniquement la technique de cartogramme (*CaCa*) car nous n'avons pas été en mesure de déterminer comment représenter deux variables à l'aide de la taille du polygone. Nusrat et al. [69] utilise la taille du polygone et la largeur de la bordure, qui est plus proche de notre *ChCa* que d'un hypothétique *CaCa*.

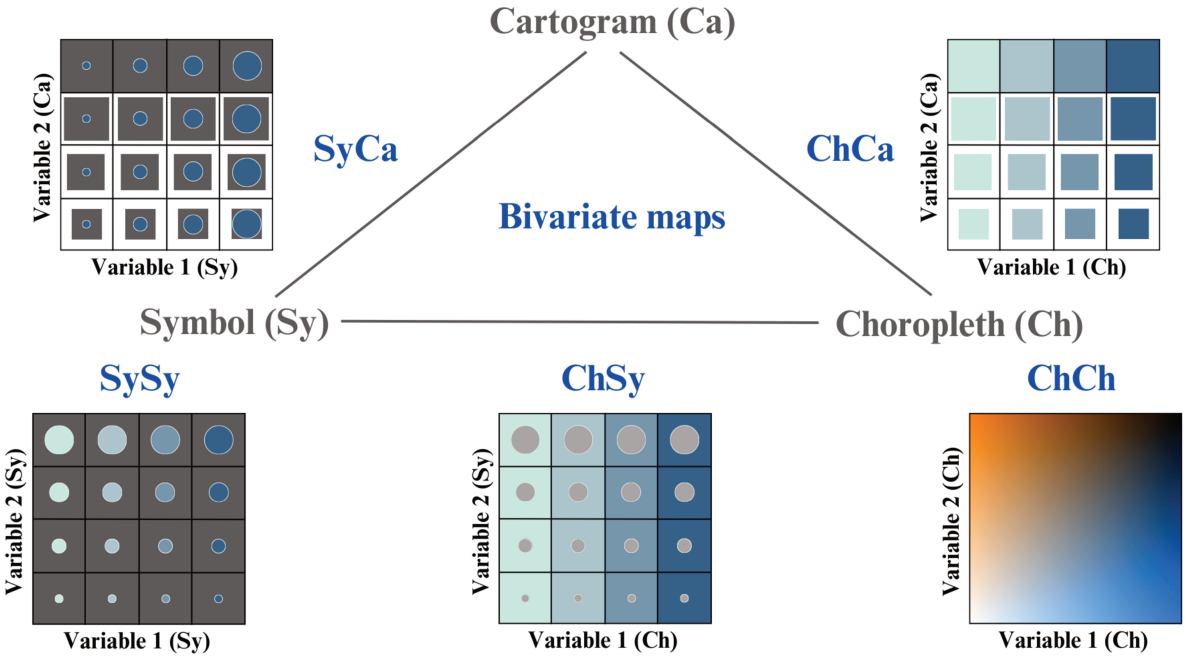


FIGURE 2.28 – Cartes bivariées comparées dans notre étude utilisateur. Elle sont issues de combinaisons entre les 3 principaux types de cartes thématiques utilisés pour représenter des valeurs quantitatives.

Comme le montre le Tableau 2.3, d'autres techniques de représentation bivariées étaient possibles à partir des différentes combinaisons possibles entre les variables visuelles et les implantations. Nous n'avons pas considéré la carte résultante de la combinaison *ponctuelle/zonale* et *couleur/couleur* car cela donnerait une carte où il deviendrait difficile de lire et interpréter les couleurs puisqu'il y en aurait dans les symboles et dans les polygones des entités géographiques. Cette visualisation serait encombrée et non pertinente. La carte avec la même implantation ponctuelle pour les deux variables et la même variable visuelle avec les couleurs (*ponctuelle/ponctuelle* et *couleur/couleur*) est très similaire à une carte choroplèthe bivariée (*ChCh*). Nous avons donc supposé que l'étudier n'aurait pas d'intérêt étant donné que l'on traite la technique *ChCh*. De plus, la visibilité des couleurs serait moindre avec les symboles. Les deux dernières combinaisons du Tableau 2.3 (*ponctuelle/ponctuelle* et *zonale/zonale* toutes deux avec *taille/taille*) ne peuvent être implémentées. Ce serait possible si au lieu d'avoir la taille comme variable visuelle on avait la largeur pour une variable et la hauteur pour l'autre par exemple. Des exemples de cartes résultantes pour chaque technique sont montrés par la Figure 2.29.

Tableau 2.3 – Combinaisons possibles entre les variables visuelles (couleur et taille) et les implantations (ponctuelle et zonale) pour générer des cartes bivariées. ✓ signifie que la carte correspondante est réalisable mais que nous ne l'avons pas incluse dans notre étude, ✗ signifie qu'elle ne l'est pas.

	Couleur/Taille	**Couleur/Couleur**	**Taille/Taille**
Ponctuelle/Zonale	*ChSy*	✓	*SyCa*
Ponctuelle/Ponctuelle	*SySy*	✓	✗
Zonale/Zonale	*ChCa*	*ChCh*	✗

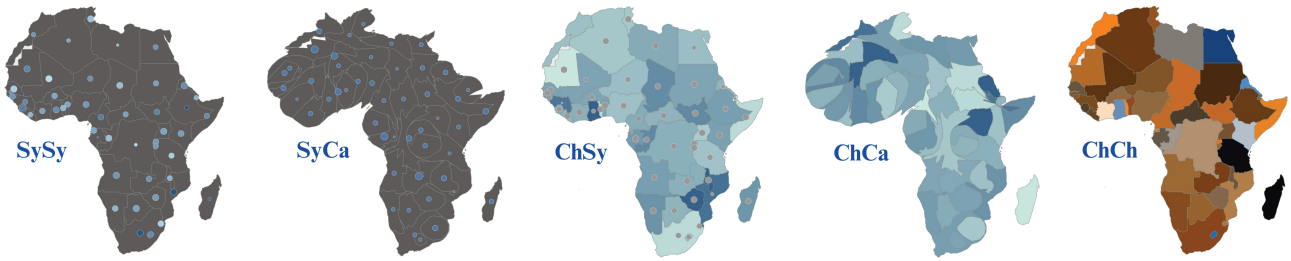


FIGURE 2.29 – Exemple de carte obtenue pour l'expérimentation de chaque technique.

L'ensemble des cartes bivariées que nous avons sélectionné nécessitent des paramètres tels que des échelles de couleur ou de taille. Comme mentionné précédemment, notre objectif est de fournir des conseils aux concepteurs de cartes ayant peu de connaissances en visualisation de données. Par conséquent, nous avons essayé de rester au plus près des paramètres par défaut d'un logiciel de visualisation bien connu et simple d'utilisation, *Tableau* [10], afin de proposer des cartes facilement reproductibles par un débutant. Pour l'échelle de taille des symboles, nous avons utilisé l'échelle par défaut de *Tableau* avec une taille minimale légèrement plus élevée pour une plus grande lisibilité. Les symboles sont placés au centroïde des polygones. L'échelle de couleur bleue séquentielle et les couleurs d'arrière-plan sont celles par défaut dans *Tableau*. Quant à l'échelle de couleurs 2D, pour *ChCh*, elle a été définie en RVB en attribuant une variable à la composante Rouge et une variable à la composante

10. https://www.tableau.com/ [dernière consultation le 04/08/2023]

Bleu. La valeur de la composante verte est la moyenne des valeurs des deux variables, ce qui rend son impact constant sur toute la palette. Enfin, les cartogrammes ont été créés à l'aide d'une implémentation existante [11] de l'algorithme décrit par Dougenik et al. [23].

2.4.1.2 Tâches

L'objectif de notre étude étant de comparer les performances des utilisateurs avec plusieurs types de cartes, nous avons défini une liste de tâches faisant intervenir deux variables quantitatives géolocalisées. La Figure 2.30 montre la liste des tâches utilisées dans notre étude utilisateur. Inspirés par l'expérimentation de Candela et al. [14], nous avons utilisé deux grands types de tâches : la détection d'une entité géographique répondant à certains critères (groupe de tâches $T1$) et l'estimation du niveau de grandeur de la valeur représentée pour une entité géographique désignée (groupe de tâches $T2$). Selon la taxonomie des tâches de recherche décrite par Munzner [64], les tâches du groupe $T1$ sont classées dans la catégorie *locate* car l'utilisateur sait ce qu'il cherche mais ne connaît pas la localisation. Ainsi, ces tâches nécessitent la capacité d'interpréter les valeurs des deux variables simultanément. Les tâches du groupe $T2$ sont classées dans la catégorie *lookup* car l'utilisateur sait ce qu'il cherche et connaît sa localisation. Les tâches de cette catégorie nécessitent de savoir différencier et dissocier les deux variables.

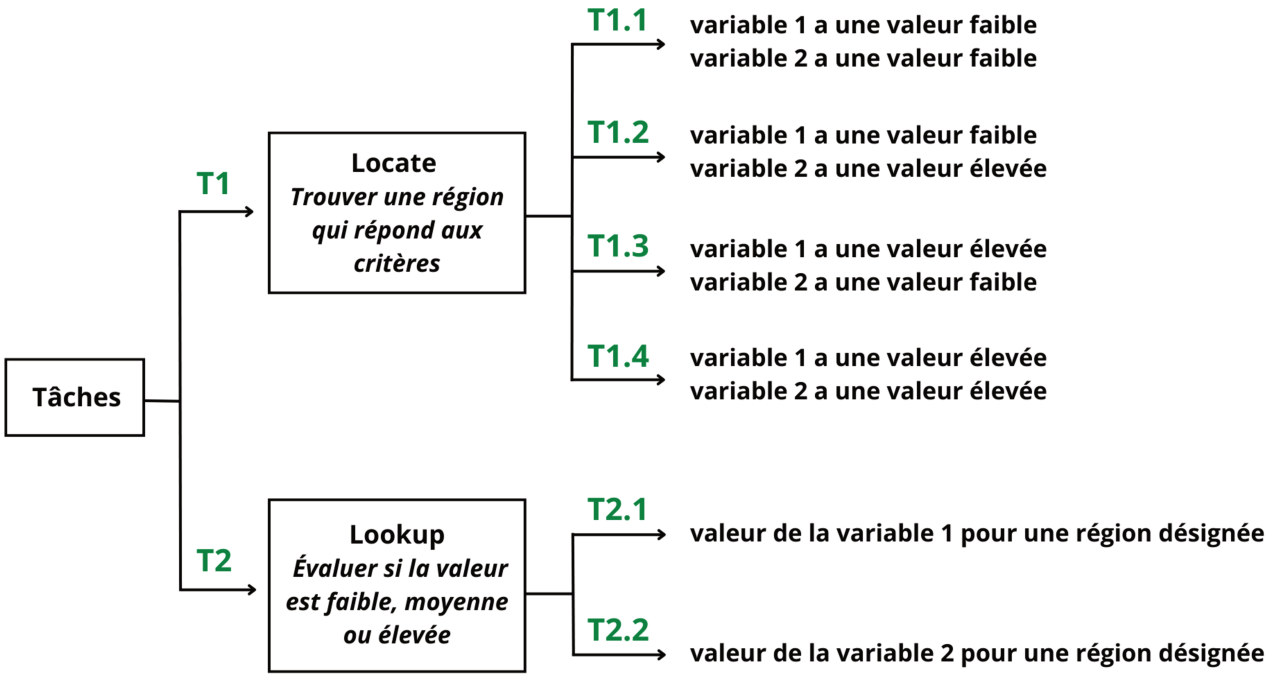


FIGURE 2.30 – Tâches utilisées pour l'expérimentation. Il existe deux types principaux : les tâches de localisation T1 (*Locate*) et les tâches de recherche T2 (*Lookup*). Ces deux types sont divisés en 4 et 2 tâches respectivement pour un total de 6 tâches.

Pour toutes ces tâches, on considère que la valeur d'une variable donnée pour une entité géographique est faible, moyenne ou élevée selon que la valeur se situe respectivement dans le premier, le deuxième ou le dernier tiers de l'intervalle $[min, max]$. Pour les tâches $T1$, les participants devaient sélectionner une entité géographique comprise dans une fourchette donnée de valeurs pour les deux variables. Nous n'avons considéré que des plages de valeurs faibles et élevées pour limiter le nombre de tâches, et donc la durée des sessions

11. https://github.com/shawnbot/topogram [dernière consultation le 04/08/2023]

d'expérimentation. Cela nous donne quatre tâches *locate* (cf. Figure 2.30). Pour les tâches du groupe $T2$, les participants devaient sélectionner si une entité géographique donnée se situait dans la fourchette de valeurs faible, moyenne ou élevée pour l'une des deux variables. Le but est d'évaluer la capacité à dissocier les deux variables sur des cartes bivariées. $T2$ se divise en deux tâches, la première impliquant la première variable et la seconde impliquant la seconde variable.

En résumé, les tâches du groupe $T1$ évaluent la lecture simultanée de deux variables sur une entité, tandis que les tâches du groupe $T2$ évaluent la lecture d'une des deux variables. Pour effectuer les tâches du groupe $T1$, l'utilisateur compare des entités géographiques entre elles. En effet, la comparaison est faite systématiquement afin de déterminer quelle entité correspond le mieux aux critères imposés par la tâche. Dans le groupe $T2$, l'objectif est d'identifier la plage de valeur d'une variable pour une entité. L'objectif ici est d'évaluer la conception des techniques, étant donné qu'une carte bivariée n'est efficace que si la variation d'une variable n'altère pas la capacité à lire l'autre.

2.4.2 Présentation de l'expérimentation

Dans cette section, nous présentons l'expérimentation. Pour cela, nous commençons par exposer les jeux de données utilisés pour générer les cartes puis le matériel mis à disposition et le profil des participants. Ensuite, nous explicitons la procédure, comment l'expérimentation a été organisée et structurée. Pour finir, nous formulons des hypothèses.

2.4.2.1 Jeux de données

Pour collecter des jeux de données dans le cadre d'une étude utilisateur, deux options sont possibles. La première consiste à générer des jeux de données fictifs. Dans ce cas, les distributions peuvent alors varier, on peut par exemple avoir une uniforme, une gaussienne et une exponentielle, ce qui tend à empêcher l'obtention de résultats spécifiques à un type de distribution. En effet, il est possible que les résultats d'une étude expérimentale diffèrent selon la distribution employée. Ainsi, il est préférable de générer de multiples jeux de données à partir de différentes distributions pour généraliser les résultats obtenus. Cependant, l'hétérogénéité des données du monde réel, en termes de distributions inhabituelles, de valeurs aberrantes, etc n'est pas prise en compte. La seconde option consiste à sélectionner des ensembles de données du monde réel présentant diverses propriétés afin de mieux saisir les structures du monde réel. Dans cette étude, nous avons choisi cette dernière option et sélectionné les jeux de données décrits dans le Tableau 2.4. Ils concernent différents sujets tels que la santé, l'agriculture, l'enseignement supérieur, le comportement social, la démographie ou bien la météo. Tous ces différents sujets permettent de diversifier les structures sous-jacentes possibles, et ils couvrent différentes régions (Afrique, Europe, États-Unis, France) afin de nuancer l'apparence des polygones et d'éviter les biais des participants qui pourraient être particulièrement familiers avec une zone géographique spécifique.

2.4.2.2 Matériel et participants

Tous les participants ont utilisé le même écran (Samsung S24A600NWU) avec les mêmes paramètres par défaut (Eye Care / Eye Saver Mode : Off et Picture / Picture Mode : Standard) et la même résolution de l'écran (2560×1440).

L'étude a porté sur un échantillon de 24 participants âgés de 19 à 59 ans (moyenne : 31,2). Les niveaux d'éducation étaient différents : 9 étaient en deuxième année d'université, 1 avait

Tableau 2.4 – Jeux de données utilisés pour générer les cartes de l'étude utilisateur. Le premier tableau présente les caractéristiques principales, les descriptions sont des hyperliens vers les ressources en ligne (dernière consultation : 03/2023). Le second tableau apporte plus de détails, une échelle logarithmique a été utilisée pour créer les visualisations afin de mieux répartir les données sur les différentes plages de valeurs (faible, moyenne et élevée) ciblées par les tâches.

	Région	Variable	Description
DS_1	Pays en Afrique	1	Espérance de vie
		2	Proportion des terres pour activités agricoles
DS_2	Pays en Europe	1	Nombre d'étudiants
		2	Dépenses de protection sociale par habitant
DS_3	États des États-Unis	1	Valeur de la production du secteur agricole
		2	Port de la ceinture de sécurité en voiture (%)
DS_4	Départements en France	1	Population
		2	Température moyenne

	Variable	Année	Valeurs	Moyenne	Médiane
DS_1	1	2015	[50.9, 76.1]	61.7	61.4
	2		[3.8, 79.4]	46.3	45.6
DS_2	1	2018	[7k, 7.6M]	815k	295k
	2		[13.9, 1.2k]	198.3	70.6
DS_3	1	2014	[49k, 52.3M]	8.1M	5.8M
	2		[69, 98]	86.4	87
DS_4	1	2021	[77k, 2.6M]	684k	541k
	2		[9.1, 17.3]	12.3	11.8

une licence, 4 avaient un master et 10 étaient doctorants ou avaient obtenu un doctorat. 10 étaient des femmes et 14 des hommes. Nous n'avons accepté que les participants n'ayant pas de troubles de la vision tels que le daltonisme et n'ayant pas reçu de formation formelle en visualisation de données. Lorsqu'on leur a demandé d'estimer leur familiarité avec la consultation des cartes, 3 ont répondu qu'ils n'avaient aucune familiarité, 8 une familiarité faible, 9 une familiarité moyenne et 4 une familiarité élevée. Lorsqu'on leur a demandé d'estimer leur familiarité avec la création de cartes, 11 participants ont estimé qu'ils n'avaient aucune familiarité, 10 une familiarité faible et 3 une familiarité moyenne. Aucun participant n'a estimé avoir un niveau élevé de familiarité avec la création de cartes.

2.4.2.3 Procédure

Un essai demande de répondre à une tâche $TASK$ bien précise à partir d'un jeu de données DS visualisé à l'aide d'une technique cartographique de visualisation bivariée $TECH$. La réponse apportée par le participant est évaluée avec une valeur de précision calculée et avec le temps qu'il a mis à répondre. Pour chaque essai, considérons un jeu de données $DS \in [DS_1, DS_2, DS_3, DS_4]$, une tâche $TASK \in [T1.1, T1.2, T1.3, T1.4, T2.1, T2.2]$ et une technique $TECH \in [SySy, SyCa, ChSy, ChCa, ChCh]$. Au total, il y a 120 configurations d'essai

différentes possibles ($TECH \times TASK \times DS$).

Pour les tâches $T1 \in [T1.1, T1.2, T1.3, T1.4]$ (*Locate*), la précision était un nombre réel de l'ensemble $\{0, 0.25, 0.5, 0.75, 1\}$, en fonction des valeurs de l'entité géographique sélectionnée par le participant et des plages de valeurs indiquées dans la tâche. La Figure 2.31 illustre comment nous calculons cette précision avec $T1.3$ comme exemple et la Figure 2.32 pour cette même tâche offre un aperçu de l'interface de l'expérimentation. Quant aux tâches $T2 \in [T2.1, T2.2]$ (*Lookup*), la précision était de 1 ou 0 : le participant a évalué la plage de la valeur correctement ou pas. La région en question était indiquée dans l'énoncé et elle était également mise en évidence sur la carte comme dans la Figure 2.33 avec le Cantal comme exemple. Cela garantissait qu'un participant qui ne savait pas où se localisait l'entité géographique en question sur la carte n'était pas pénalisé.

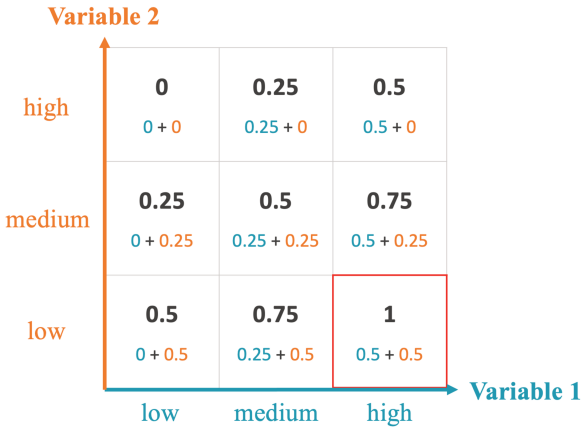


FIGURE 2.31 – Le score de précision est indiqué par le nombre noir dans chaque case. Les chiffres colorés indiquent le score obtenu pour chaque variable (0 si opposé, 0.25 si proche, 0.5 si correct). Une entité géographique tombant dans le carré rouge est une réponse tout à fait correcte : valeur élevée pour la variable 1 et valeur faible pour la variable 2.

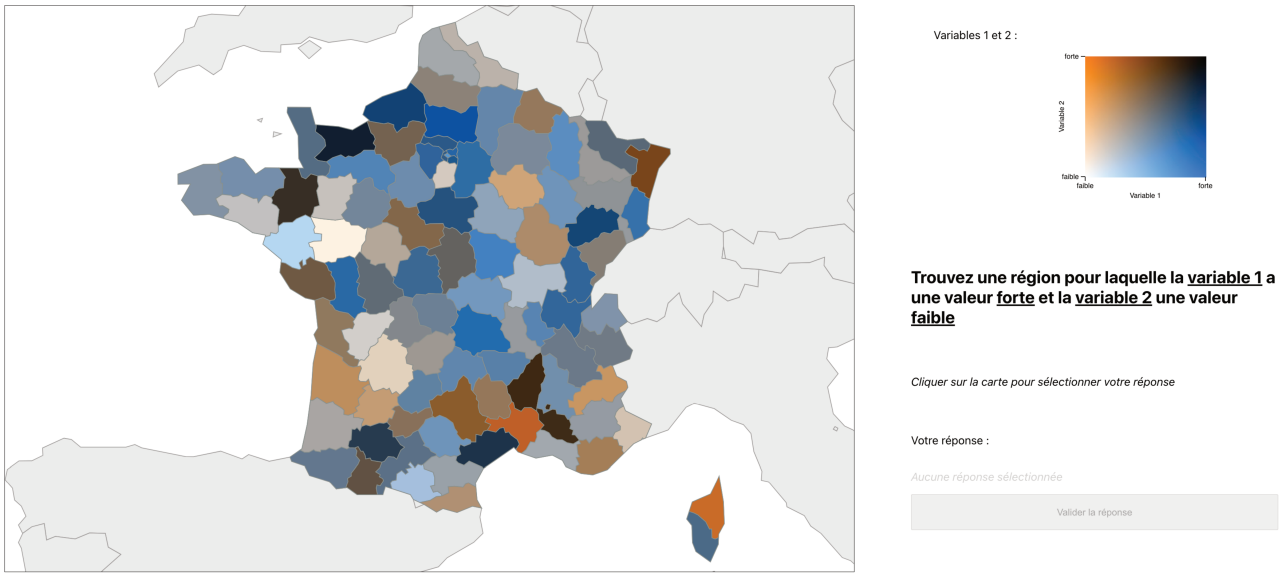


FIGURE 2.32 – Aperçu de l'interface de l'expérimentation pour le groupe $T1$ avec la tâche $T1.3$, la technique utilisée dans cet exemple est *ChCh*.

En nous basant sur une procédure inspirée de Lobo et al. [60], nous avons défini un ensemble de 24 sessions assurant un bon contrepoids. Une session est un ensemble ordonné

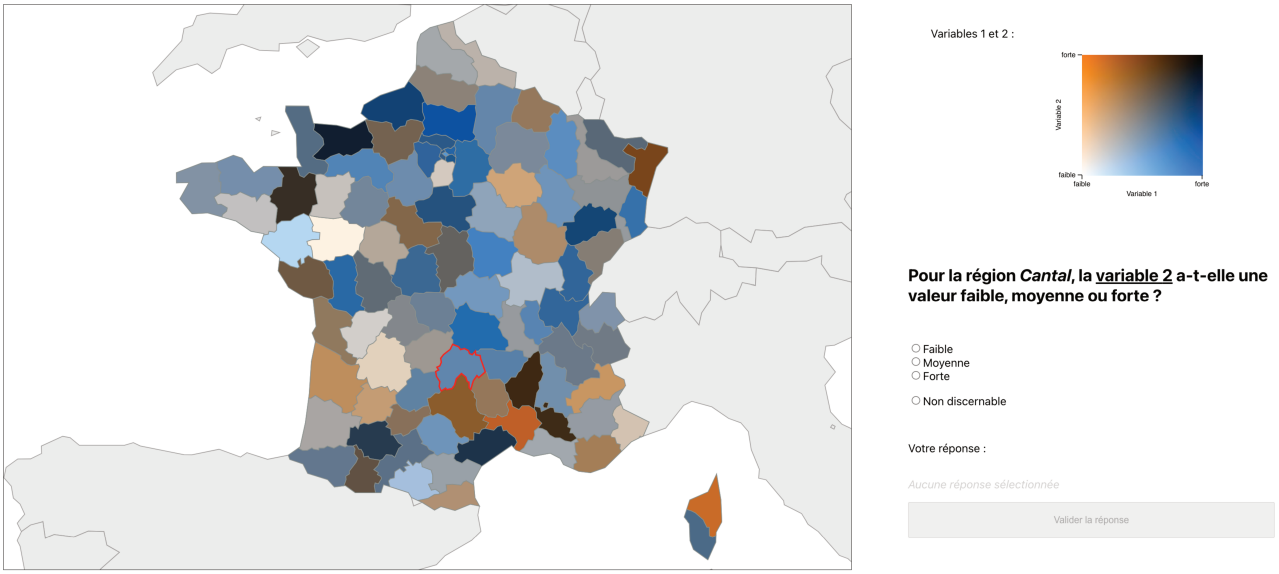


FIGURE 2.33 – Aperçu de l'interface de l'expérimentation pour le groupe $T2$ avec la tâche $T2.2$, la technique utilisée dans cet exemple est *ChCh*.

d'essais pour un participant. Comme souligné par Lobo et al. [60], un bon contrepoids est de la plus haute importance lorsqu'il s'agit de données du monde réel, car la complexité des essais ne peut pas être contrôlée.

Afin de maintenir la concentration des participants, nous avons souhaité limiter la durée d'une session à moins de 30 minutes. Une étude pilote, réalisée auprès de 4 doctorants de notre laboratoire, a montré qu'il faut environ 20 minutes pour réaliser une session de 45 essais. Nous sommes donc restés sur cette valeur. Nous avons utilisé un design intra-sujet $[5 \times 3 \times 3]$ avec les 3 variables primaires de $TECH$, $TASK$ et DS : chaque participant a rencontré chaque type de carte ($TECH$), mais n'a effectué que 3 types de $TASK$ (2 de $T1$ et 1 de $T2$) sur 3 jeux de données (DS). Cette étude pilote nous a aussi permis de constater que les participants pouvaient retenir les bonnes réponses des paires $TASK - DS$ d'une $TECH$ à l'autre. Afin d'éviter ce phénomène d'apprentissage, les associations, que nous avons appelé clé-valeur, entre les entités géographiques et les valeurs des deux variables associées, ont été mélangées. Par conséquent, le même couple de valeurs pour les deux variables n'était pas associé à la même entité géographique lorsque les participants passaient d'une $TECH$ à l'autre. Pour éviter les disparités entre les sessions, nous avons défini un ensemble clé/valeur pour chaque $TECH$ qui était le même pour toutes les sessions.

La Figure 2.34 illustre la structure en détail des sessions. Les essais des sessions ont été regroupés en 5 blocs, un par $TECH$. L'ordre des $TECH$ pour chaque session a été choisi au hasard parmi les possibilités de $5\ ! = 120$. Les participants devaient effectuer tous les essais liés à une $TECH$ avant de passer au bloc suivant. 3 $TASK$ ont été attribués à chaque session ($T1.1$, $T1.2$ et $T2.1$ pour les sessions impaires et $T1.3$, $T1.4$ et $T2.2$ pour les sessions paires) avec un ordre tiré au sort parmi les possibilités $3! = 6$. Ainsi, les participants devaient compléter 3 $TASK$ pour chaque $TECH$, l'ordre des $TASK$ étant le même pour une même session mais différent d'une session à l'autre. Les DS ont été regroupés en lots de 3 sur les 4 disponibles avec la contrainte que chaque configuration $TECH - TASK - DS$ apparaisse le même nombre de fois parmi les 24 sessions. Le premier essai de chaque paire $TECH \times TASK$ était considéré comme un entraînement et le participant pouvait poser des questions à la personne qui le supervisait. Cela nous a permis de nous assurer que le participant avait compris l'encodage visuel et la tâche à accomplir. La précision et le temps de

réponse des deuxième et troisième essais ont été enregistrés. En d'autres termes, nous avons obtenu $45 \times 24 = 1080$ de scores de précision et de temps de réponse, dont $30 \times 24 = 720$ ont été enregistrés pour effectuer l'analyse. Le contre-balancement a fait en sorte que les 120 combinaisons de $TECH - TASK - DS$ soient exécutées 9 fois par différents participants et enregistrées 6 fois.

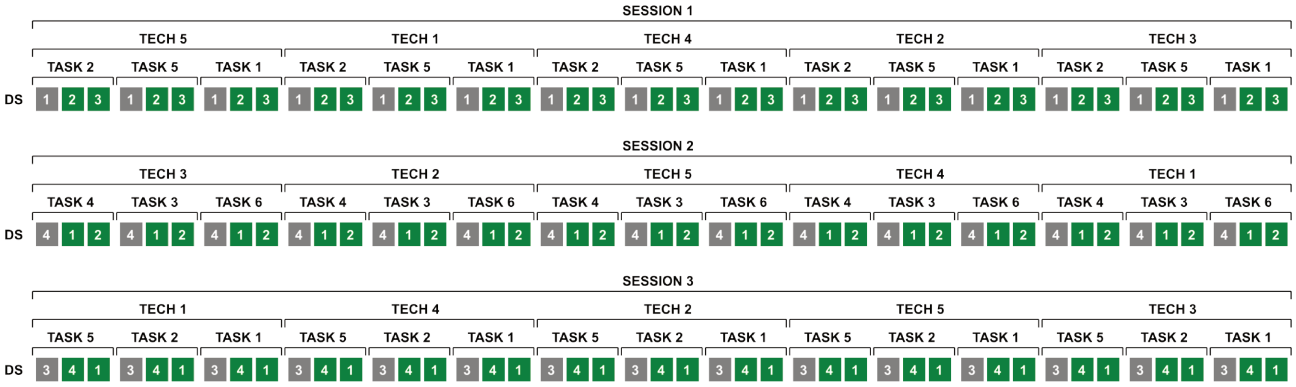


FIGURE 2.34 – Trois exemples de sessions de l'étude : à chaque fois qu'un participant changeait de *TASK*, le premier essai était considéré comme un entraînement. Seuls des essais effectif ont été utilisés pour calculer nos résultats.

Avant de débuter l'expérimentation, chaque participant a pu bénéficier d'une présentation bien précise de l'objectif de l'étude. L'ensemble des techniques ont été décrites, et pendant les essais d'entraînement, les participants étaient autorisés à poser des questions afin de s'assurer que tout était bien clair pour eux. À la fin des 45 essais, c'est-à-dire d'une session, les participants pouvaient indiquer s'ils avaient subi une baisse de concentration et, si oui, à quel moment. Ils ont également établi plusieurs classements des *TECH* en fonction de ce qu'ils ressentaient et de critères : (1) de la plus adaptée à la réalisation des tâches à la moins adaptée, (2) de la plus compréhensible à la moins compréhensible, (3) de la plus esthétique à la moins esthétique et (4) de la plus appréciée globalement à la moins appréciée. Les participants ont ensuite été invités à préciser quelques informations supplémentaires à leur propos : âge, sexe, niveau d'études actuel et familiarité avec les visualisations cartographiques (consultation et création). La durée moyenne des séances était de 21,3 minutes (min : 9,4, max : 40,2).

2.4.2.4 Hypothèses

La visualisation de deux variables sur la même carte peut grandement faciliter l'observation d'une corrélation entre deux variables. Pour ce faire, la carte doit limiter les biais d'interprétation, être compréhensible et fournir les informations de manière adaptée. Comme l'explique Nelson [65], l'attention sélective soutient que notre capacité à analyser les variables visuelles d'un symbole, c'est-à-dire la couleur et la taille dans notre cas, est affectée par d'autres variables visuelles présentes dans le même symbole. La recherche psychologique suggère que certaines combinaisons de variables visuelles peuvent améliorer ou restreindre l'attention sélective. Pour les techniques *SySy* et *ChCa* la couleur pourrait affecter la perception de la taille des symboles/polygones par exemple. Pour *ChCh*, qui a également une seule implantation (zonale), les variations de teinte sont moins perceptibles lorsque les deux valeurs sont élevées. De plus, il peut être difficile de distinguer le cas où une variable a une valeur moyenne et l'autre une valeur forte étant donné que la teinte sera sombre. Elmer [25] aborde d'ailleurs ce problème. L'étude de Sun et al. [86] a comparé les cartogrammes à

d'autres cartes thématiques telles que les cartes à symboles proportionnels (Sy) et les cartes choroplèthes (Ch). Les résultats ont montré que les autres cartes thématiques sont considérées comme plus efficaces. De plus, les cartes de symboles et les choroplèthes sont beaucoup plus couramment utilisés. Cependant, le principal avantage des cartogrammes est que, contrairement à la cartographie choroplèthe notamment, ils permettent une meilleure adéquation entre la saillance visuelle grâce à la taille et à l'importance des données. De plus, l'interprétation avec des techniques basées sur des cartogrammes peut être beaucoup plus facile si l'utilisateur connaît bien la zone géographique à analyser. En effet, comme le démontre l'étude utilisateur de Rittschof et al. [76], les résultats diffèrent selon la familiarité avec les régions. Sur la base de ces remarques, nous avons formulé les cinq hypothèses ci-dessous.

Hypothèse H1. Les techniques basées sur le cartogramme (*SyCa* et *ChCa*) devraient donner des scores de précision globalement inférieurs car elles devraient être plus difficiles à interpréter.

Hypothèse H2. Les techniques de visualisation *ChSy* et *SyCa* devraient être les plus efficaces pour les tâches $T2$. Ils ont deux implantations différentes, ce qui évite qu'une variable visuelle n'altère la perception d'une autre. Si l'hypothèse H1 est vérifiée, *ChSy* devrait être la technique la plus efficace pour les tâches $T2$.

Hypothèse H3. Les techniques utilisant des cartogrammes (*SyCa* et *ChCa*) devraient nécessiter un temps de réponse plus long que les autres.

Hypothèse H4. La technique de visualisation *ChSy* devrait être la plus populaire et considérée comme la plus facile à interpréter par les participants car c'est la plus répandue.

Hypothèse H5. Les scores de précision et les temps de réponse avec des techniques basées sur des cartogrammes (*SyCa* et *ChCa*) pourraient s'améliorer si l'utilisateur est familier avec la zone géographique représentée.

2.4.3 Résultats de l'expérimentation

Dans cette section, nous procédons d'abord à une analyse statistique des résultats obtenus par les participants en termes de précision pour ensuite faire de même au niveau des temps de réponse. A posteriori, nous examinons les réactions des participants. Enfin, nous présentons une discussion des résultats et évaluons la validité de nos hypothèses.

2.4.3.1 Précision des réponses

Les tâches du groupe $T1$ consistent à localiser une entité géographique correspondant à deux plages de valeurs données, une pour chaque variable. Par conséquent, les scores de précision sont des nombres réels de l'ensemble $\{0, 0.25, 0.5, 0.75, 1\}$. Au contraire, les tâches du groupe $T2$ impliquent l'identification d'une plage de valeurs pour une entité donnée. La réponse d'un participant est donc soit vraie, soit fausse. L'analyse statistique de ces deux types de tâches doit donc se baser sur des approches différentes compte tenu de la nature des valeurs observées.

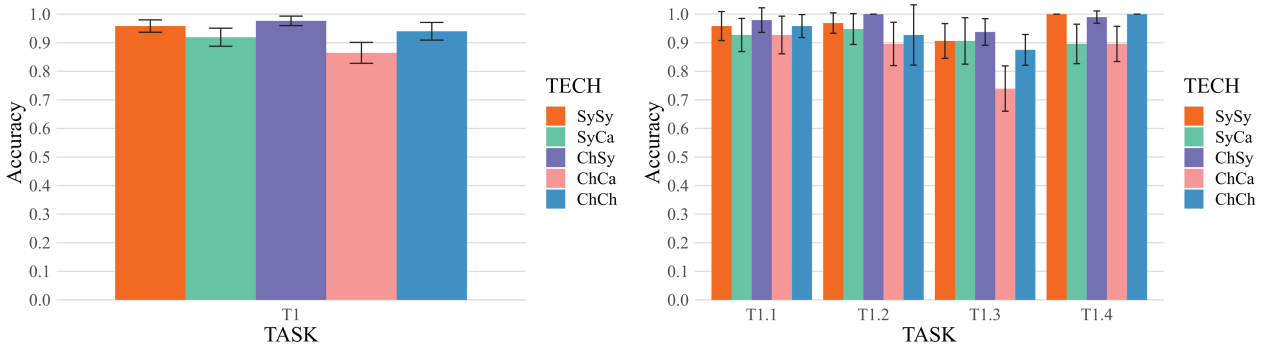
Dans les deux cas, les paramètres du modèle permettent de saisir les différences entre les techniques de visualisation. Pour toutes les estimations et les tests de ces paramètres, nous utilisons les sorties des fonctions R pour l'estimation des GLM, qui renvoient, avec l'estimation, la valeur de la statistique de Wald (valeur z) et la valeur p associée, qui teste la nullité

du paramètre. Nous utilisons le seuil de référence de 0,05 pour les valeurs p.

1) Tâches *Locate*

Comme mentionné ci-dessus, les valeurs de précision pour les tâches $T1$ représentent 5 niveaux qui ne sont pas des évaluations purement quantitatives de la réponse. Elles représentent plutôt 5 niveaux ordonnés d'une variable catégorielle. Par conséquent, nous supposons des distributions multinomiales avec une contrainte de niveaux ordonnés sur les probabilités. Pour cette contrainte, nous utilisons la fonction R *vglm* avec un lien cumulatif pour prendre en compte l'ordre. Les modèles multinomiaux sont des modèles linéaires généralisés (GLM) particuliers, qui sont similaires aux modèles ANOVA mais dont les hypothèses de distribution varient et vont au-delà de la famille gaussienne.

Pour chaque *TECH* et chaque tâche du groupe *TASK Locate* ($T1$), les moyennes des scores de précision et les intervalles de confiance sont représentés dans la Figure 2.35.



(a) Précision moyenne des tâches $T1$ par *TECH*.

(b) Précision moyenne par tâche de $T1$ et par *TECH*.

FIGURE 2.35 – Précision moyenne pour $T1$. Les barres d'erreur représentent l'intervalle de confiance à 95%.

Les résultats pour la tâche $T1.1$ sont satisfaisants quel que soit la tâche *TECH*. Nous pouvons voir sur la Figure 2.35b que tous les *TECH* ont une précision moyenne supérieure à 0.90 et un faible écart-type. *ChCh* a l'écart-type le plus faible avec une valeur d'environ 0.09, tandis que *ChCa* a l'écart-type le plus élevé avec une valeur d'environ 0.16. *ChSy* a la moyenne la plus élevée (0,98), suivie de *SySy* et *ChCh* (0,96). Les deux techniques qui incluent un cartogramme, *ChCa* et *SyCa*, ont la moyenne la plus basse (0.93). Cependant, nous n'avons pas trouvé de différence significative entre ces résultats ($p = 0.08$ entre *ChSy* et *SyCa*, $p = 0.11$ entre *ChSy* et *ChCa*, et $p > 0.30$ entre les autres paires). Lorsque les cartes basées sur les cartogrammes, à savoir *SyCa* et *ChCa*, sont comparées ensemble à toutes les autres, la différence n'est pas non plus significative ($p = 0.09$).

Pour la tâche $T1.2$, *ChSy* n'a eu que des réponses correctes et *SySy* a également obtenu d'excellents résultats, avec une précision moyenne de 0,97. Ils sont suivis par *SyCa* (0.95), *ChCh* (0.93), et enfin *ChCa* (0.89). Comparées aux 100% de réponses correctes de *ChSy*, toutes les autres techniques sont nettement moins précises. Nous pouvons également noter que l'écart-type des scores de précision de *SySy* est plus faible (0.08) que les autres (> 0.13), à l'exception de *ChSy* bien entendu. La différence entre les cartes basées sur les cartogrammes *SyCa* et *ChCa* comparées ensemble aux autres est significative ($p = 0.02$).

ChSy donne les meilleurs résultats pour la tâche $T1.3$, avec une précision moyenne d'environ 0.94, suivi par *SySy* et *SyCa* (0.91), mais l'écart-type est plus élevé pour ce dernier,

0.19 contre 0.14. *ChCh* (0,87) vient ensuite et *ChCa* est le moins performant, avec une précision moyenne de 0,74. Cette différence est significative ($p < 0,02$ entre ChCa et tout autre $TECH$), mais les différences observées entre les autres paires ne le sont pas ($p > 0,12$ pour toutes les paires). Les cartes basées sur les cartogrammes *SyCa* et *ChCa* comparées conjointement aux autres présentent également une différence significative ($p = 0.03$).

Pour la tâche $T1.4$, deux $TECH$ n'ont fourni que des réponses correctes : *ChCh* et *SySy*. *ChSy* a également obtenu une excellente précision moyenne de 0,99. *SyCa* et *ChCa* ont obtenu des scores significativement plus bas (0,89, $p = 0,02$ entre ces derniers et *ChSy*) et ont des écarts types plus élevés (environ 0,15 alors que les autres ont 0 ou 0,05). Comparées aux 100% de réponses correctes de *ChCh* et *SySy*, toutes les autres techniques sont significativement moins précises. Statistiquement, les cartes basées sur les cartogrammes *SyCa* et *ChCa* sont très similaires l'une à l'autre ($p = 0,86$) et différentes des autres ($p < 0,01$).

Enfin, comme nous avons émis l'hypothèse que les techniques basées sur les cartogrammes sont plus performantes si les zones géographiques sont plus familières, nous avons effectué une analyse statistique pour tester cette hypothèse. Les participants vivaient tous en France, nous pouvions donc supposer que la précision serait meilleure pour DS_2 et DS_4 (respectivement les pays européens et les départements français). Nous avons trouvé une différence significative entre DS_2 et DS_1 (pays africains, $p = 0.02$) mais pas entre DS_2 et DS_3 (états américains, $p = 0.74$). Nous n'avons pas trouvé de différences significatives entre DS_4 et les autres ensembles de données. Il convient de noter qu'il existe également une différence significative entre DS_1 et DS_3 ($p = 0.04$).

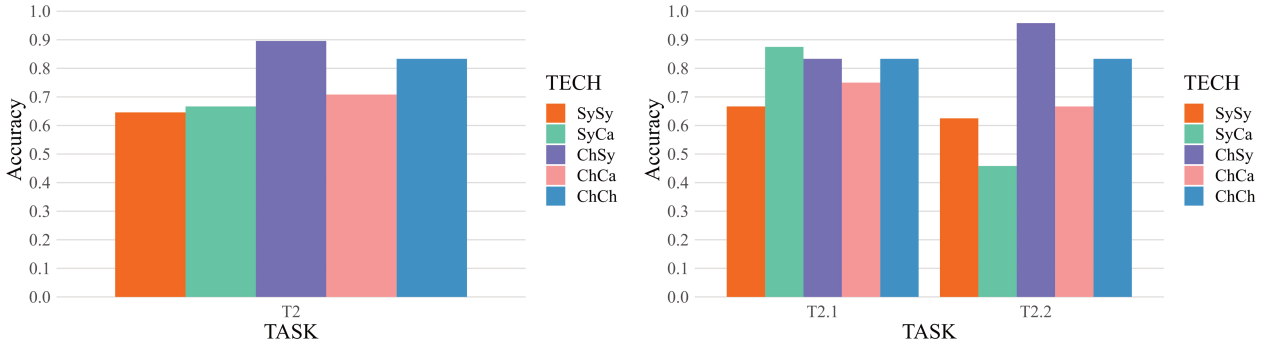
Globalement, comme on peut le voir dans la Figure 2.35a, pour les tâches $T1$, *ChSy* semble être la technique la plus efficace en termes de précision des réponses. En effet, elle présente des différences significatives par rapport à d'autres techniques : $p < 0.01$ avec *ChCa* et *SyCa*, $p = 0.04$ avec *ChCh*. *SySy* semble également être un bon concurrent, même s'il n'y a pas de différence significative par rapport aux autres ($p > 0.08$), à l'exception de *ChCa* ($p < 0.01$). *ChCh* donne des résultats hétérogènes : il est parmi les meilleurs pour $T1.1$ et $T1.4$, qui pour rappel consistent respectivement à localiser une entité ayant comme valeurs pour les deux variables faible/faible et élevé/élevé, et parmi les moins bons pour $T1.2$ et $T1.3$ (respectivement faible/élevé et élevé/faible). Il semble donc mieux adapté pour les valeurs extrême, lorsque les participants recherchent des régions sombres ou claires que lorsqu'ils recherchent des régions colorées. Enfin, les deux techniques qui utilisent la méthode des cartogrammes (*SyCa* et *ChCa*) ont les précisions de réponses les plus faibles pour toutes les tâches de $T1$ ($p < 0.01$ entre eux et les autres). Plus précisément, *ChCa* obtient des résultats significativement moins bons que tous les autres $TECH$ ($p = 0,01$ avec *SyCa* et $p < 0,01$ avec les autres), tandis que *SyCa* n'est significativement moins bon que de *ChSy* ($p < 0,02$). La différence entre *ChCa* et *SyCa* est principalement due aux différentes moyennes obtenues dans la tâche $T1.3$.

2) Tâches *Lookup*

Les réponses aux tâches du groupe $T2$ sont dichotomiques, soit vraies, soit fausses. Nous supposons donc une distribution de Bernoulli, et utilisons donc la fonction glm de R pour estimer les modèles, avec la famille binomiale et le lien logit canonique.

Les moyennes des scores de précision pour chaque $TECH$ et tâche du groupe $TASK$ "Lookup" $T2$ sont présentées dans la Figure 2.36.

Pour la tâche $T2.1$, nous avons demandé aux participants d'évaluer la valeur de la première variable, qui est représentée par la couleur du symbole pour *SySy*, la taille du symbole pour *SyCa*, et la couleur du polygone pour *ChSy*, *ChCa* et *ChCh*. *SySy* a obtenu le plus pe-



(a) Précision moyenne des tâches $T2$ par $TECH$.

(b) Précision moyenne par tâche de $T2$ et par $TECH$.

FIGURE 2.36 – Précision moyenne pour $T2$.

tit nombre de réponses correctes avec une moyenne de $0,67$, suivi de $ChCa$ $(0,75)$, $ChSy$ et $ChCh$ $(0,83)$, et $SyCa$ $(0,87)$, comme le montre la Figure 2.36b. Ces taux de réponses correctes ne sont pas significativement différents ($p \geq 0.1$). Statistiquement, les cartes basées sur les cartogrammes $SyCa$ et $ChCa$ ne sont pas significativement différentes des autres ($p = 0.65$).

Quant à la tâche $T2.2$, nous avons demandé aux participants d'évaluer la valeur de la seconde variable, représentée par la taille du symbole pour $SySy$ et $ChSy$, la taille du polygone pour $SyCa$ et $ChCa$, et la couleur du polygone pour $ChCh$. Les scores moyens sont beaucoup plus disparates. Les résultats sont assez médiocres pour $SyCa$: plus de 50% des réponses sont incorrectes (la moyenne est de 0.46). Au contraire, $ChSy$ obtient d'excellents résultats, avec une moyenne de $0,96$. De plus, cette différence est très significative ($p < 0.01$). Les résultats pour $SyCa$ présentent également une différence significative par rapport à $ChCh$ ($p < 0.01$), qui se place en deuxième position avec une moyenne de 0.83. On peut également noter que $ChCa$ et $SySy$ donnent des résultats significativement moins bons $(0,67$ et $0,63)$ que $ChSy$ ($p < 0,05$). Il existe également une différence significative entre les cartes basées sur les cartogrammes et les autres avec $p < 0,01$.

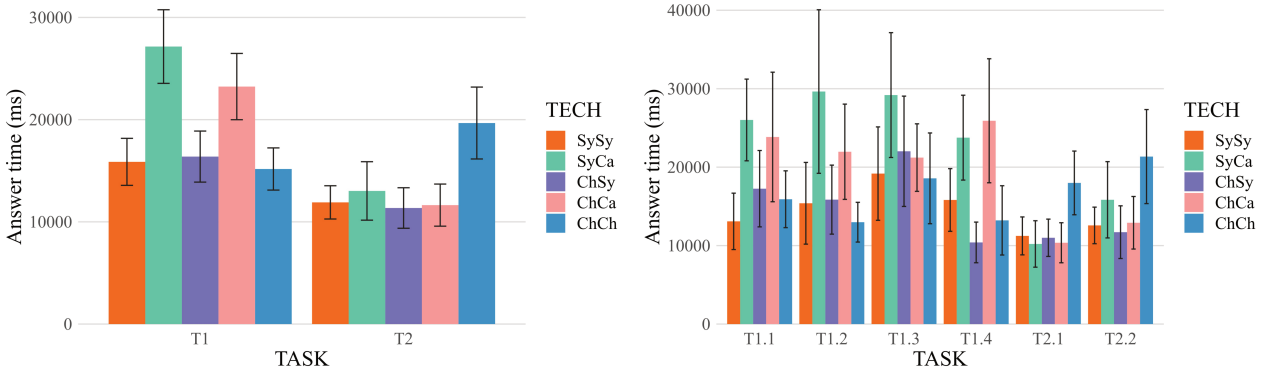
Étant donnée que l'objectif des tâches $T2$ est d'évaluer les combinaisons des variables visuelles, nous pouvons nous attarder sur les performances relatives à ces combinaisons. Les participants devaient estimer la taille des symboles dans $T2.1$ pour $SyCa$ et dans $T2.2$ pour $SySy$ et $ChSy$. Les scores obtenus sont très variables : $ChSy$ arrive en tête (0.96), puis $SyCa$ (0.87, $p < 0.01$ avec $ChSy$), et enfin $SySy$ (0.63, $p = 0.02$ avec $ChSy$ et $p = 0.04$ avec $SyCa$). Ainsi, l'estimation de la taille du symbole semble être la plus précise lorsque la deuxième variable visuelle est la couleur du polygone, suivie de la taille du polygone et enfin de la couleur du symbole. En ce qui concerne la couleur du polygone, nous n'observons pas de telles variations de précision : $ChSy$ et $ChCh$ ont un score de 0.83 et $ChCa$ un score de 0.75 pour $T2.1$ (non significatif, $p = 0.48$). Ainsi, dans ce cas, la deuxième variable visuelle ne semble pas avoir un effet important sur la précision de l'estimation de la couleur des polygones. Enfin, nous observons une variation lorsque les participants estiment la taille du polygone (cartes basées sur des cartogrammes) : $SyCa$ a un score de 0.46 et $ChCa$ un score de 0.67 pour $T2.2$ (non significatif cependant, $p = 0.15$). Il semble donc que la taille du polygone soit mieux estimée lorsque la deuxième variable visuelle est la couleur du polygone plutôt que la taille du symbole, mais cette conclusion n'est pas corroborée par l'analyse statistique.

Globalement, comme l'illustre la Figure 2.36a, $ChSy$ a la précision moyenne la plus élevée pour les tâches $T2$ (0.89). $ChCh$ (0.83) est en seconde position, suivi de $ChCa$ (0.71), $SyCa$ (0.67), et $SySy$ (0.66). Les différences entre $ChSy$ et $SySy$ ($p < 0,01$), $SyCa$ ($p < 0,01$) et $ChCa$

($p = 0,03$) sont significatives. Il est à noter que la différence entre le meilleur concurrent de *ChSy*, étant *ChCh*, et la technique la moins précise, *SySy*, est également significative ($p = 0.04$). Les autres paires ne présentent pas de différences significatives.

2.4.3.2 Temps de réponse

Pour chaque essai, à l'exception des entraînements, nous avons enregistré le temps qui a été nécessaire au participant pour répondre à la question. Cette mesure, purement quantitative n'est pas normalement distribuée. Nous supposons la distribution exponentielle, ou la distribution gamma plus largement. Nous avons donc utilisé la fonction *glm* de R avec la famille gamma et le lien identité. La Figure 2.37 montre les résultats par *TASK* et *TECH*.



(a) Temps de réponse moyen pour *T*1 et *T*2 par *TECH*.

(b) Temps de réponse moyen par *TASK* et par *TECH*.

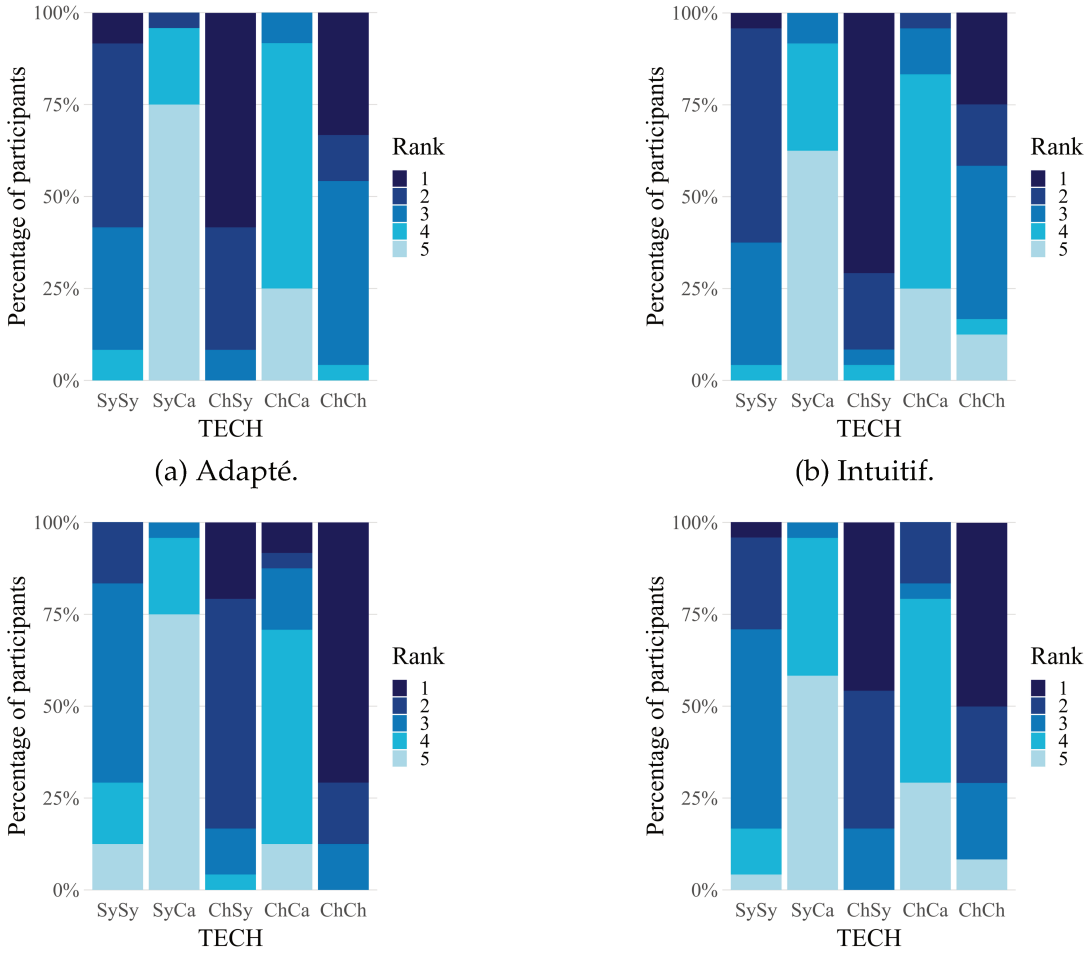
FIGURE 2.37 – Temps de réponse moyens. Les barres d'erreur représentent l'intervalle de confiance à 95%.

Globalement (cf. Figure 2.37a), pour les tâches *T*1, les cartes basées sur des cartogrammes, *SyCa* et *ChCa*, ont nécessité beaucoup plus de temps (environ 27 et 23 secondes en moyenne) que les autres (entre 15 et 17 secondes). Ces différences sont significatives ($p < 0,01$), tandis que celles entre les autres ne le sont pas ($p \geq 0,45$). Pour les tâches *T*2, les temps de réponse étaient plus longs pour *ChCh* (aux alentours de 19 secondes) que pour les autres (entre 11 et 13 secondes). Cette différence est significative ($p < 0,01$), alors que les autres ne le sont pas ($p \geq 0,28$). La Figure 2.37b montre le temps de réponse moyen pour chaque tâche et technique de visualisation. Étant donné que le temps de réponse moyen par tâche individuelle est presque cohérent avec celui qui vient d'être présenté pour l'ensemble des groupes *T*1 et *T*2, nous nous concentrerons uniquement sur un résultat surprenant : nous pouvons observer que la réalisation de la tâche *T*1.3 avec *ChCa* ne prend pas plus de temps qu'avec les autres techniques non basées sur des cartogrammes ($p \geq 0,50$). Par conséquent, trouver une entité géographique avec une valeur élevée représentée par la couleur d'un polygone et une valeur faible représentée par la taille du polygone est plus rapide que l'inverse (*T*1.2) et plus ou moins la même chose qu'avec les autres techniques.

2.4.3.3 Retours des participants

À la fin d'une session, les participants ont été invités à classer les cinq techniques de visualisation en leur attribuant une note de 1 (la meilleure) à 5 (la pire) selon différents critères. Les classements étaient les suivants : de la plus adaptée à la moins adaptée à la réalisation des tâches, de la plus intuitive à la moins intuitive, de la plus esthétique à la

moins esthétique, et enfin leurs propres préférences (cf. Figure 2.38). Nous avons également donné la possibilité d'ajouter des commentaires libres.



(a) Adapté.

(b) Intuitif.

(c) Esthétique.

(d) Préférences personnelles.

FIGURE 2.38 – Avis des participants : classement des types de cartes selon différents critères.

En général, nous pouvons noter que les techniques basées sur les cartogrammes ont obtenu les scores les plus bas : pour l'ensemble des critères, elles ont obtenu entre 71% et 96% des votes pour l'avant dernière ou dernière position. À l'opposé, *ChSy* a obtenu les scores globaux les plus élevés, avec au moins 83% en première et seconde position pour chaque critère. Les résultats pour *SySy* et *ChCh* sont plus hétérogènes. Alors que *SySy* est considéré comme plus adapté et plus intuitif (respectivement 58% et 62% en première ou seconde position) que *ChCh* (respectivement 46% et 42% en première ou seconde position) *ChCh* surpasse cependant *SySy* en termes de préférence (71% contre 29% en première ou seconde position). C'est également le cas pour l'esthétique, critère pour lequel *ChCh* devance bien largement *SySy* et plus de peu *ChSy* (87% contre 17% et 83% pour la première ou seconde position dans le classement).

Les participants pouvaient aussi ajouter des commentaires facultatifs à la fin de la session s'ils le souhaitaient. Sur les 24 participants, 14 ont écrit des commentaires, soit environ 58%. La totalité des commentaires, sans aucun exception, étaient liés à la difficulté de lire les cartogrammes, plus particulièrement sans connaître la zone géographique.

2.4.3.4 Discussion sur les résultats

Notre expérience confirme H1 (cf. Tableau 2.5). Pour les tâches $T1$, la précision était significativement plus faible avec les techniques basées sur les cartogrammes, *SyCa* et *ChCa*. Entre les deux, *ChCa* a obtenu les moins bons résultats durant notre expérimentation. Pour $T2.1$, les participants devaient estimer la première variable, c'est-à-dire la variable représentée par la taille du symbole pour *SyCa* et par la couleur du polygone pour *ChCa*. Pour cette tâche, nous avons observé une précision moyenne plus faible pour les techniques basées sur les cartogrammes, mais la différence n'était pas significative. Pour $T2.2$, lorsque les participants devaient estimer la valeur de la deuxième variable, représentée à l'aide du cartogramme, les résultats ont été significativement plus mauvais que les autres techniques. Dans l'ensemble, les commentaires des participants suggèrent également que *SyCa* et *ChCa* sont plus difficiles à interpréter : ils ont d'ailleurs classé ces techniques au dernier rang en termes d'adéquation et d'intuitivité.

Puisque notre expérimentation a confirmé H1, H2 suggère que *ChSy* est la technique la plus efficace pour les tâches $T2$ car c'est la seule technique avec deux implantations différentes et sans cartogramme. En ce qui concerne $T2.2$, les résultats le confirment. Pour $T2.1$, il n'y a pas de différence significative entre *ChSy* et ses concurrents, mais sa précision moyenne est la deuxième plus élevée ; et surtout les résultats en eux-mêmes ont été très largement satisfaisants. Si l'on considère les deux tâches $T2$, *ChSy* est la plus efficace et les différences sont significatives par rapport à toutes les autres techniques, à l'exception de *ChCh*. Par conséquent, H2 n'est que partiellement confirmée, mais nous pouvons raisonnablement penser qu'elle peut être confirmée.

Comme le montre le Tableau 2.5, H3 a été confirmée pour les tâches $T1.1$, $T1.2$ et $T1.4$. Pour $T1.3$, *SyCa* a nécessité plus de temps que les autres techniques, mais pas *ChCa*. Cependant, pour l'ensemble des tâches $T1$, H3 a été confirmée étant donné que la différence est significative. Ainsi, les cartes basées sur les cartogrammes ont des temps de réponse beaucoup plus longs pour les tâches $T1$. Pour les tâches $T2$, toutes les techniques, y compris les cartes basées sur des cartogrammes, ont obtenu des résultats plus ou moins similaires, à l'exception de *ChCh*. Nous supposons que cela peut s'expliquer par le fait qu'il est difficile pour l'œil humain de distinguer certaines nuances de teinte, en particulier lorsqu'une variable est moyenne et l'autre élevée. La teinte est alors trop foncée et il est très difficile, voire impossible, de déterminer quelle variable est moyenne et quelle variable est élevée. Ainsi, H3 est rejetée pour les tâches $T2$.

ChSy a été la technique la mieux classée par les participants. Elle a également été considérée comme la plus appropriée et la plus intuitive par la majorité des participants et a été bien classée en termes d'esthétique (surpassée uniquement par *ChCh*) et de préférence. Pour toutes ces raisons, nous pouvons confirmer H4.

Les réactions des participants, rapportées ci-dessus, semblent confirmer H5. Plus de la moitié d'entre eux (58%) ont laissé des commentaires, et tous sans exception ont mentionné la difficulté d'interpréter les cartogrammes lorsque la zone géographique n'était pas bien connue. Cependant, nous n'avons pas trouvé de preuves statistiques lorsque nous avons comparé les résultats obtenus sur l'Europe et la France aux autres (tous les participants vivaient en France). Peut-être faudrait-il accorder plus d'importance aux commentaires des participants car notre analyse statistique était basée sur l'hypothèse que les participants connaissaient mieux les pays européens et les départements français que les pays africains et les États américains ; ce qui n'est peut-être pas vrai. La France métropolitaine est effectivement composée de 96 départements et il est probable que la plupart des personnes vivant en France ne connaissent la forme que de quelques-uns d'entre eux. On peut donc raisonnablement penser que H5 est confirmée. Cela soulève la question de savoir si les techniques

Tableau 2.5 – Aperçu des hypothèses validées par tâche : ✓ signifie que nous avons observé une différence significative en faveur de l'hypothèse, ✗ signifie que les résultats ne soutiennent pas l'hypothèse.

	*T*1.1	*T*1.2	*T*1.3	*T*1.4	*T*2.1	*T*2.2
H1 (précision) cartogrammes moins bons	✗	✓	✓	✓	✗	✓
H2 (précision) *ChSy* meilleur pour *T*2					✗	✓
H3 (temps) cartogrammes moins bons	✓	✓	✗	✓	✗	✗

basées sur les cartogrammes pourraient être plus efficaces dans des situations particulières. Pour rappel, les participants n'étaient pas des experts en cartographie. Avec des utilisateurs connaissant très bien la zone cartographiée et habitués à ce type de représentation, *SyCa* et *ChCa* pourraient peut-être révéler tout leur potentiel. Cependant, d'après les résultats de notre expérimentation, ces deux techniques ne sont pas adaptées au grand public, non experts en cartographie.

Notre étude présente des similitudes avec celle menée par Elmer [25], dont l'objectif était d'examiner l'impact des caractéristiques perceptuelles des cartes bivariées sur leur interprétation. Cependant, les deux études diffèrent sur plusieurs points :

- L'étude d'Elmer compare 8 techniques de visualisation, dont celles que nous nommons *ChSy*, *ChCa* et *ChCh*, basées sur différentes approches (variation de la hauteur/largeur du symbole, variation de l'orientation du symbole, variation de l'opacité, texture des polygones, et diagrammes à barres en tant que symbole). Notre objectif étant de nous concentrer sur des cartes communes facilement reproductibles avec un outil simple tel que *Tableau*.
- Elmer s'est appuyé sur un seul ensemble de données fictives, présenté aux participants, ils avaient donc déjà un à priori. Nous de notre côté, nous n'avons donné aucune information à propos des données aux participants. La carte de base était composée des comtés de l'Ohio, aux États-Unis, avec des modifications au niveau de la rotation, de la taille et de la topologie. Dans notre étude, nous avons préféré utiliser des ensembles de données réelles provenant de différentes régions du monde (nous expliquons ce choix dans la Section 2.4.2.1).
- La procédure est différente : dans l'étude d'Elmer, chaque type de carte était associé à un bloc de 8 questions, et l'ordre des types de cartes présentés ainsi que l'ordre des questions dans chaque bloc étaient aléatoires. Notre procédure est différente, notamment parce qu'elle inclut les mêmes questions pour différents ensembles de données (cf. Section 2.4.2.3).
- La majorité des participants dans l'étude d'Elmer étaient des spécialistes, alors que nous nous sommes concentrés sur les participants qui n'avaient pas d'expérience en matière de visualisation de données et de cartographie (cf. Section 2.4.2.2).

Notre objectif n'était pas de reproduire la même étude mais plutôt de la compléter, notamment en nous concentrant sur des cartes facilement reproductibles pour un public non expert. Cependant, il est intéressant de comparer les principaux résultats. Comme mentionné

plus haut, Elmer a également comparé *ChSy* et *ChCh*. Il y avait aussi une technique basée sur des cartogrammes de type *ChCa*, mais elle n'était pas contiguë, nous ne la prenons donc pas en compte. Comme dans notre étude, Elmer n'a pas noté de différence significative entre *ChSy* et *ChCh* en termes de précision, mais la précision moyenne de *ChCh* était légèrement supérieure (0,99 contre 0,97). En ce qui concerne le temps de réponse, *ChSy* figurait parmi les techniques les plus rapides dans l'étude d'Elmer, mais ce n'était pas le cas de *ChCh*. Notre étude a trouvé des résultats équivalents pour $T2$ (avec une différence significative) mais pas pour $T1$. Enfin, les commentaires des participants étaient très similaires dans les deux études. Dans l'étude d'Elmer, *ChSy* et *ChCh* ont été jugés de manière très positive ; ils ont été considérés comme adaptés à l'exécution des tâches et les informations contenues ont été considérées comme bien comprises. Dans notre expérimentation, *ChSy* et *ChCh* ont également été les techniques les plus appréciées et ont eu le plus de retours positifs.

En résumé, malgré les conditions différentes, les résultats de nos expérimentations sont cohérents avec ceux d'Elmer. Cela permet de renforcer l'idée que *ChSy* et *ChCh* sont des techniques de représentations bivariées bien efficaces.

Nous avons identifié plusieurs limites au niveau de notre expérimentation. Il existe de nombreuses combinaisons pour créer des cartes à deux variables représentant des variables quantitatives. Nous nous sommes basés sur les types de cartes les plus courants, mais il est tout à fait possible que nous ayons omis des alternatives potentiellement plus efficaces. De plus, il est possible que les cartes à deux variables que nous avons sélectionnées pour cette étude puissent donner des résultats différents selon les paramètres sélectionnés. Par exemple, la taille des symboles dans *SySy*, *SyCa* et *ChSy* est basée sur l'échelle par défaut de *Tableau* avec une taille minimale légèrement plus grande. Des tailles de symbole minimales et/ou maximales différentes produiraient des cartes diverses et potentiellement d'autres résultats. Pour *ChSy* et *SyCa*, nous avons choisi d'utiliser la taille du symbole, mais nous aurions également pu choisir la couleur comme variable visuelle. Nous avons fait ce choix parce que la taille est connue pour être une variable visuelle plus précise que la couleur comme l'indique Munzner [64]. Pour *SyCa* et *ChCa*, nous avons opté pour des cartogrammes contigus, mais il existe de nombreux autres types de cartogrammes (cf. l'état de l'art proposé par Nusrat et al. [71]) qui pourraient éventuellement être plus efficaces pour les cartes bivariées. Il convient également de noter que nos résultats ne concernent que les échelles de couleurs que nous avons choisies. D'autres échelles de couleurs pourraient donner des résultats différents. Cependant, si l'on prend en compte les travaux de Liu et Heer [59], les échelles de couleurs *blues* (*SySy*, *ChCa*, *ChSy*) et *blueorange* (*ChCh*) ont donné des taux d'erreur très proches. Sur la base de ces travaux, nous devrions donc avoir limité les disparités entre les techniques étudiées.

Pour résumer, les résultats de notre étude sont basés sur nos cartes, réalisées avec les paramètres décrits ci-dessus, et la généralisation de ces résultats à d'autres cartes et/ou paramètres est délicate. Nous n'avons pas pu tester un plus grand nombre de cartes dans le cadre de cette expérimentation en raison de contraintes liées au processus (durée de la session, nombre de participants, etc.). De futurs travaux sur cet aspect seraient bénéfiques et permettraient de compléter cette étude.

2.5 Discussion

Pour développer la compréhension générale de la comparaison, Gleicher et al. [37] a proposé une taxonomie des conceptions visuelles pour la comparaison. De nombreux systèmes

et approches ont été répertoriés mais au final, toutes les conceptions sont assemblées à partir des éléments constitutifs de la juxtaposition, superposition et des encodages explicites. La taxonomie a donc été définie à partir de ces 3 catégories :

- La juxtaposition consiste à présenter plusieurs objets séparément, elle fait appel à la mémoire de l'utilisateur.
- La superposition permet de visualiser des objets dans le même espace, utilise le système visuel.
- L'encodage explicite quant à lui, fournit un codage visuel des relations entre les objets, nécessite un calcul pour déterminer les relations.

Dans ce chapitre, nous avons traité la juxtaposition et la superposition. Nous avons défini la superposition, dans le cadre de la visualisation de données associées à des entités géographiques, comme étant une visualisation dans laquelle un même fond de carte est utilisé pour deux ou plusieurs variables. Toutes les techniques de visualisation par superposition se déclinent en réalité en deux catégories : la superposition et les encodages explicites. Si on se tient aux définitions et à la taxonomie de Gleicher et al., certaines cartes bivariées que nous avons étudié sont des encodages explicites et non pas des superpositions.

La Figure 2.39 illustre ces trois catégories à travers des exemples simples. Dans ce chapitre nous nous sommes penchés sur Figure 2.39-ⓐ dans la Section 2.3, sur Figure 2.39-ⓑ et ⓓ dans la Section 2.4. Certaines des techniques étudiées dans la Section 2.4 sont des encodages explicites : *ChCh*, *SySy* et *ChCa*. Il s'agit des cartes bivariées utilisant la même implantation pour les deux variables. Cependant, que ce soit une superposition ou bien un codage explicite, on a un même carte de fond. Dans ces deux cas, chaque entité géographique est représentée une unique fois.

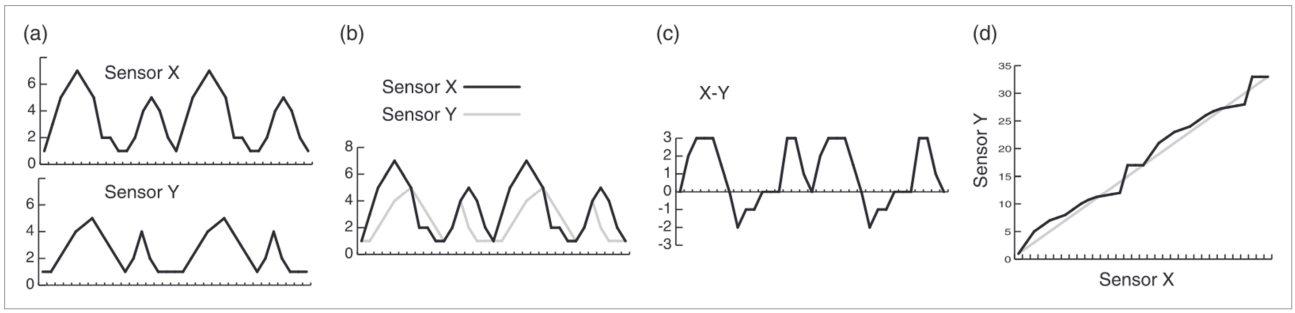


FIGURE 2.39 – Les catégories de comparaisons définies par Gleicher et al. [37] : (a) juxtaposition, (b) superposition, (c, d) codage explicite des relations.

La dimension temporelle a été traitée uniquement dans le cadre de la juxtaposition, combiner notre approche de juxtaposition avec des cartes bivariées ou même multivariées est une perspective afin de prendre en compte la dimension temporelle et ainsi observer des évolutions tout en comparant des variables.

Nous avons évoqué d'autres types de cartes dans la Section 2.2.1.3 telles que les cartes dasymétriques qui sont sous-exploitées. Nous ne les avons pas étudié car elles sortent du sujet principal de ce chapitre qui est la visualisation de données associées à des entités géographiques. Les entités géographiques ne sont effectivement pas prises en compte dans ce type de carte.

2.6 Conclusion

Nous avons abordé dans ce chapitre les approches de comparaison de données associées à des entités géographiques à travers la visualisation. Les techniques de visualisation était des cartes et les approches de comparaison consistaient à soit juxtaposer les cartes ou bien les superposer, et dans ce cas, avoir un même fond de carte pour plusieurs variables.

Nous avons tenté d'exploiter au maximum les avantages de la comparaison par juxtaposition en proposant une approche interactive basée sur un système de synchronisation entre les dimensions principales des visualisations : la variable, la date et la zone géographique. Cette approche a été mise en application grâce à la plateforme de visualisation de données épidémiologiques Epid Data Explorer.

Concernant la comparaison par superposition, nous avons étudié 5 types de cartes bivariées pour représenter des données quantitatives associées à des entités géographiques. Pour cela, nous avons élaboré une expérimentation basée sur 6 tâches divisées en deux catégories, une première évaluant l'efficacité de la technique pour représenter simultanément deux variables et une seconde pour vérifier que la lecture d'une variable n'est pas entravée par la seconde variable. L'analyse des résultats montre que la carte bivariée combinant une carte choroplèthe pour la première variable et des symboles proportionnels pour la seconde est la plus performante en termes de précision, de temps de réponse et de préférences des utilisateurs. La carte choroplèthe bivariée s'est elle aussi montrée très convaincante, même si les temps de réponse étaient plus élevés pour les tâches de la seconde catégorie. La carte bivariée combinant la taille et la couleur des symboles a été efficace pour les tâches de première catégorie, mais beaucoup moins pour les autres tâches. Comme l'explique la théorie de la sélection attentive, cette technique ne permet pas de distinguer les deux variables étant donné que l'implantation est la même pour les deux. Enfin, les cartes bivariées représentant l'une de leurs variables par un cartogramme ont présenté les moins bonnes performances.

CHAPITRE

3

DONNÉES ASSOCIÉES À DES RELATIONS ENTRE DES ENTITÉS GÉOGRAPHIQUES

Sommaire

3.1 Introduction 62
3.2 État de l'art 63
3.2.1 Visualisation 64
3.2.1.1 Géographie abstraite 64
3.2.1.2 Géographie déformée 65
3.2.1.3 Géographie explicite 66
3.2.2 Cartes de flux 67
3.2.3 Distorsion de carte 72
3.3 Algorithme de distorsion FMD (Flow Map Distortion) 74
3.3.1 Première distortion 75
3.3.2 Courbes 79
3.3.3 Deuxième distorsion 79
3.4 Création de carte de flux un-à-plusieurs 81
3.4.1 Création du graphe initial 81
3.4.2 Définition de l'arbre 83
3.4.3 Simplification de l'arbre 83
3.4.4 Dessin des flux 86
3.5 Discussion 87
3.5.1 Jeux de données 87
3.5.2 Temps de calcul 87
3.5.3 Résultats 88
3.6 Conclusion 93

3.1 Introduction

Le chapitre précédent était consacré à la visualisation et à la comparaison de données quantitatives liées à des entités géographiques indépendantes les unes des autres. Dans ce nouveau chapitre, notre attention se porte sur la représentation de données quantitatives qui décrivent les relations entre ces entités géographiques. Ici, l'objectif n'est plus de comparer les valeurs des différentes entités, mais plutôt d'analyser les relations entre elles.

Divers types de relations peuvent exister entre les entités géographiques. Nous nous concentrons principalement sur les déplacements comme, par exemple, les migrations, les échanges commerciaux, ou encore les transports. La représentation de ces mouvements a été un défi pendant de nombreuses décennies. Ces dernières années, avec l'explosion des volumes de données et la mondialisation croissante, la mobilité et les échanges n'ont cessé de croître. Ainsi, la visualisation de données quantitatives concernant les relations entre les entités géographiques demeure un enjeu à l'ordre du jour.

Une approche traditionnelle, remontant aux années 1800, consiste à cartographier les relations sous forme de flux, i.e. à superposer des liens représentant les flux sur une carte de base constituée d'entités géographiques. L'épaisseur des liens reflète l'ampleur des flux. Toutefois, cette méthode présente des inconvénients liés à l'encombrement dû aux chevauchements entre les flux et les entités géographiques. Par conséquent, des informations géographiques peuvent être perdues, certains pays se trouvant masqués. Une alternative pour visualiser les flux consiste à tracer des lignes de même épaisseur et à utiliser une échelle de couleurs séquentielle pour représenter les quantités associées. Cependant, cette technique de visualisation rend complexe la perception des ordres de grandeur et la comparaison des valeurs sans biais pour l'œil humain. L'utilisation de l'épaisseur comme variable visuelle pour représenter des valeurs quantitatives est en effet considérée comme plus efficace que l'utilisation des couleurs, comme le souligne notamment Munzner [64].

Les cartes de flux de Charles-Joseph Minard (1781-1870), un ingénieur civil français [75], ont particulièrement retenu notre attention. La Figure 3.1 présente un exemple de ses cartes. Sa méthode de représentation est très intuitive : chaque flux est représenté par une ligne d'épaisseur proportionnelle à sa valeur, et aucune de ces lignes ne se superpose aux entités géographiques grâce à une déformation du fond de carte. Étant donné que l'objectif de ces cartes n'est en aucun cas de mesurer les surfaces des entités géographiques, la distorsion est une bonne solution pour éviter les problèmes d'encombrement tout en faisant varier les épaisseurs des flux. De plus, les entités géographiques demeurent aisément identifiables et reconnaissables.

C'est à partir de cette constatation que nous avons défini notre objectif : concevoir des cartes de flux inspirées de celles de Minard. Plus précisément, nous visons à créer des cartes de flux où les épaisseurs des flux varient et où le fond de carte s'ajuste en fonction de ces flux. Notre méthode nous permet de conserver tous les avantages des cartes élaborées à la main par Minard tout en automatisant leur création. Nos travaux de recherche se concentrent exclusivement sur les cartes de flux de type "un-à-plusieurs", c'est-à-dire celles qui représentent les flux partant d'une origine unique vers plusieurs destinations.

La contribution se présente sous trois axes majeurs :

1. Nous proposons une nouvelle approche de la visualisation des cartes de flux un-à-plusieurs basée sur la distorsion du fond de carte ;
2. Nous proposons une méthode de distorsion de fond de carte afin d'éliminer les chevauchements entre les entités géographiques et les flux ;
3. Nous proposons une nouvelle méthode pour générer des cartes de flux un-à-plusieurs qui sont adaptées à notre méthode de distorsion.

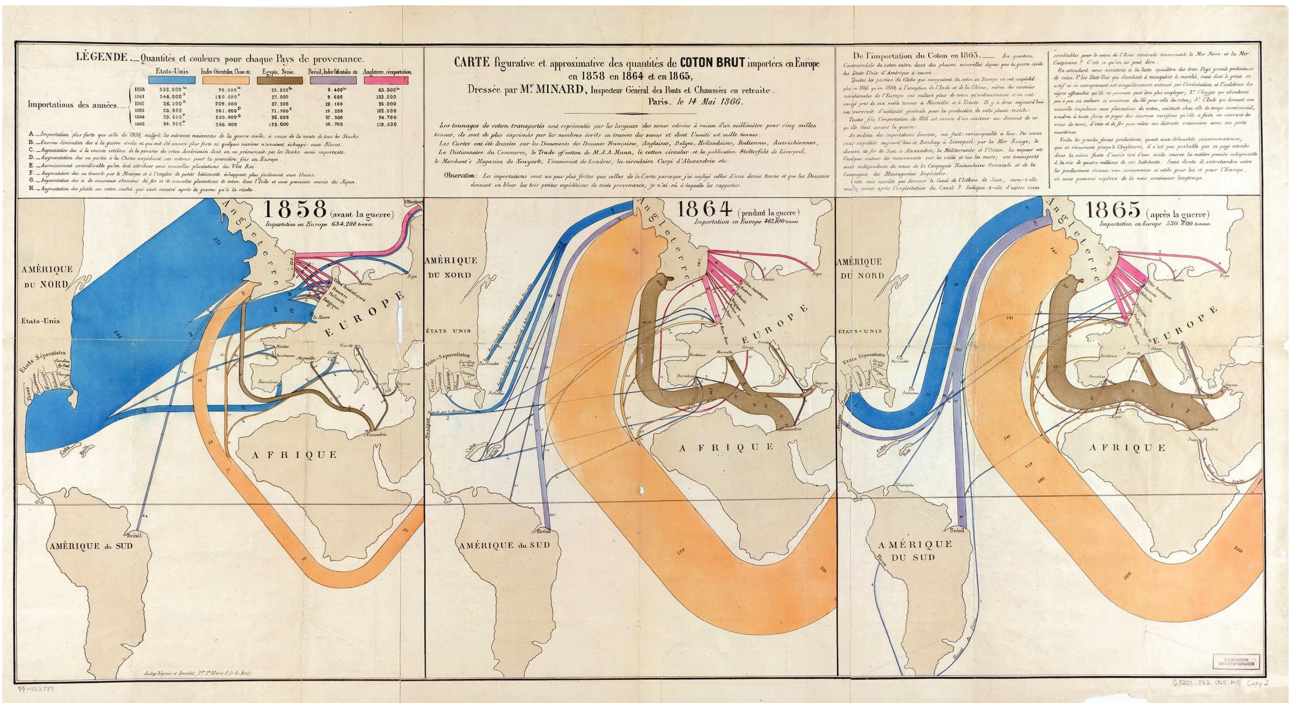


FIGURE 3.1 – Cartes de flux de Charles-Joseph Minard sur les quantités de coton brut importées en Europe, une couleur est attribuée à chaque pays exportateur [75].

Ces travaux ont été publiés dans une revue internationale des sciences de l'information géographique [95] :
Viau, L., Sallaberry, A., Rodriguez, R., Girres, J.-F., & Poncelet, P. (2023). Polygon vector map distortion for increasing the readability of one-to-many flow maps. *International Journal of Geographical Information Science*, 37(6), 1288-1314. http://doi.org/10.1080/13658816.2023.2190374

Le chapitre est structuré comme suit : nous commençons par un état de l'art sur la visualisation de données quantitatives associées à des relations entre les entités géographiques dans la Section 3.2. Ensuite, la Section 3.3 décrit notre méthode de distorsion de fond de carte. Notre méthode de création de carte de flux un-à-plusieurs, adaptée et complémentaire à notre méthode de distorsion, est présentée dans la Section 3.4. Nous proposons par la suite une discussion dans la Section 3.5 avant de conclure ce chapitre dans la Section 3.6.

3.2 État de l'art

La méthode que nous proposons s'inscrit à la croisée des domaines de la visualisation des flux structurés par des graphes géographiques et de la distorsion de cartes. À notre connaissance, de nombreuses méthodes ont été proposées dans ces deux domaines, cependant, depuis l'époque de Minard, elles n'ont pas été combinées au sein d'une approche similaire à la nôtre. Dans un premier temps, nous présentons les techniques existantes pour visualiser des données quantitatives associées aux relations entre des entités géographiques. Ensuite, nous passons en revue les méthodes spécifiques à la création de cartes de flux. Enfin, nous présentons des exemples d'applications de distorsion de cartes qui nous ont inspirés dans l'élaboration de notre méthode de distorsion.

3.2.1 Visualisation

La visualisation de données quantitatives associées aux relations entre des entités géographiques, couramment désignée sous le terme de visualisation des données OD, pour Origine-Destination, s'inscrit dans le domaine de la visualisation de données géospatiales. Un cadre de conception pour la visualisation des données OD a été proposé par Tennekes et Chen [89]. Ils précisent que la représentation des entités géographiques ainsi que les données relatives à leurs relations peuvent être représentées telles quelles, ou bien être agrégées via une abstraction des données. Les informations géographiques, qui constituent le contexte des données, peuvent être représentées de trois manières distinctes : abstraite, déformée, ou explicite, comme le détaille l'état de l'art sur la visualisation des graphes géospatiaux de Schöttler et al. [82].

Dans les prochaines sections, nous présentons les différentes méthodes de visualisation en fonction de ces trois types de représentations géographiques, en distinguant à chaque fois les méthodes effectuant une abstraction des données de celles qui ne le font pas. Par ailleurs, nous avons relevé, en nous basant sur la littérature existante, qu'il est possible de distinguer deux formes principales de visualisation : les diagrammes nœuds-liens et les matrices.

3.2.1.1 Géographie abstraite

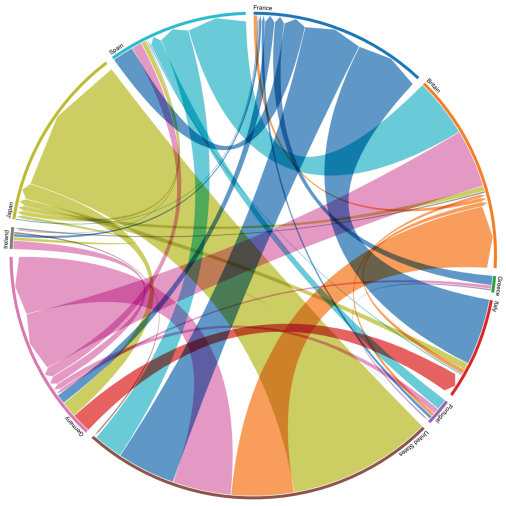
Certaines techniques de visualisation optent pour une représentation géographique abstraite. Dans ce contexte, la position d'un élément n'est pas déterminée précisément en fonction de sa localisation géographique. C'est le cas des diagrammes de membrures (de l'anglais *Chord diagram* en suivant la traduction proposée dans la documentation d'ArcGIS[1]), utilisés à la fois par Abel et Sander [39] et Hennemann [44]. Cette méthode consiste à disposer les entités géographiques le long de la circonférence d'un cercle et à les connecter les unes aux autres. Chaque entité est associée à une couleur et les liens ont des épaisseurs qui varient en fonction des valeurs associées. La Figure 3.2a illustre un exemple de ce type de visualisation. La représentation des nœuds et des liens est explicite.

Une autre technique qui représente l'information géographique de manière abstraite a été proposée par Xiao et Chun [99]. Appelée kriskogramme elle est utilisée pour visualiser des flux migratoires. Cette visualisation est composée d'une ligne de localisation représentée sous forme d'un segment. Sur cette ligne les entités géographiques sont projetées en fonction de leur position géographique de l'est vers l'ouest. Chaque flux entre deux entités, l'origine et la destination, est représentée sous forme d'un demi-cercle. La représentation des nœuds et des liens est, comme pour les diagrammes de membrures, explicite.

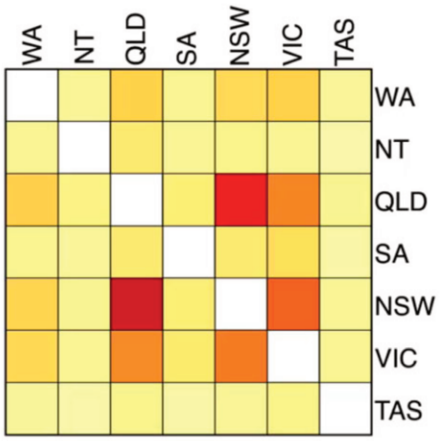
Toutes les méthodes présentées jusqu'à maintenant se présentent sous la forme d'un digramme nœuds-liens mais il existe aussi des solutions se basant sur des matrices d'adjacence (Bach et al. [5]). La Figure 3.2b illustre un exemple de matrice OD permettant de visualiser les quantités de flux entre des entités géographiques.

Les matrices constituent une bonne alternative aux diagrammes nœuds-liens pour résoudre le problème de densité et d'encombrement du graphe d'après Ghoniem [36]. Cependant, son évaluation concerne les graphes généraux et non pas les graphes géospatiaux qui ont leurs propres particularités.

1. https://doc.arcgis.com/fr/insights/latest/create/chord-diagram.htm [dernière consultation le 12/09/2023]



(a) Diagramme de membrures : visualise les dettes entre des pays durant la crise de l'euro en 2011. Chaque couleur correspond à un pays "origine". https://observablehq.com/@d3/directed-chord-diagram



(b) Matrice OD : représente les migrations internes de la population en Australie entre ses 8 États fédéraux (Schöttler et al. [82]).

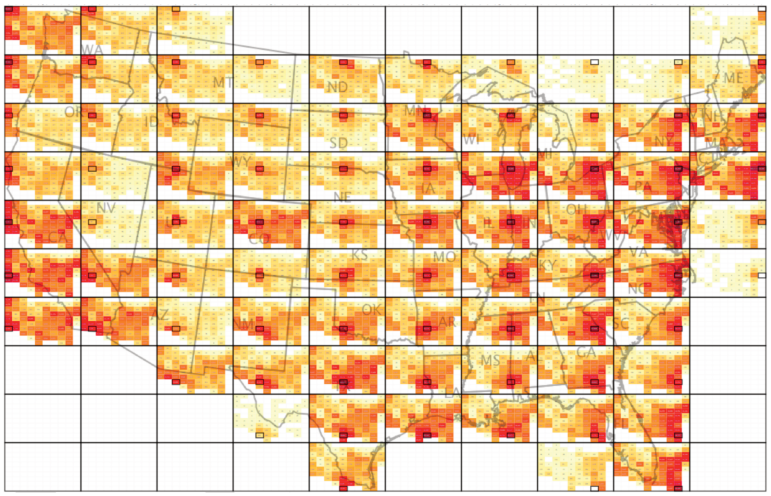
FIGURE 3.2 – Exemples de visualisations ayant une représentation de la géographie abstraite.

3.2.1.2 Géographie déformée

Les techniques se servant d'une représentation de la géographie déformée déplacent les entités géographiques et/ou altèrent leur forme. Dans la plupart des cas, cette déformation est employée dans le but de clarifier la visualisation et de mettre en évidence les interactions entre les entités géographiques.



(a) Exemple de distorsion continue : les distances correspondent à la durée d'un voyage en train (Bouts et al. [8]).



(b) Exemple de carte OD : c'est un ensemble de 721 432 vecteurs de migration de comté à comté des États-Unis qui est représenté. Une grande cellule correspond à une entité de départ, elle est composée d'une carte de densité de chaque destination utilisant la même grille (Wood et al. [98]).

FIGURE 3.3 – Exemples de visualisations ayant une représentation de la géographie déformée.

Certaines techniques [8, 4] représentent des mesures, telles que des temps de trajets par exemple, en s'appuyant sur les distances entre les entités géographiques (cf. Figure 3.3a). Pour cela, une distorsion continue est effectuée. Dans ce type de visualisation, la distance entre deux entités ne correspond pas à la distance géographique et aucune abstraction des données n'est requise pour clarifier la représentation. Par exemple, la carte OD proposée par Wood et al. [98] utilise une grille spatiale régulière superposée à une carte et regroupe les entités dans chaque cellule de la grille. En d'autres termes, les entités géographiques sont transformées en une grille de tuiles identiques. Cette transformation constitue une distorsion discontinue. Le volume des flux entre chaque paire de cellules, c'est-à-dire la valeur quantitative entre deux entités, est ensuite codé à l'intérieur de la cellule à l'aide d'une carte imbriquée et de couleurs. La Figure 3.3b illustre un exemple de cette technique de visualisation.

3.2.1.3 Géographie explicite

Les représentations géographiques qui cartographient les informations telles quelles, sans aucune distorsion, sont explicites. Tous les éléments géographiques sont représentés à l'aide d'une projection cartographique.

Souvent, les visualisations de données quantitatives associées à des relations entre des entités géographiques qui font appel à une représentation de la géographie explicite, consistent à superposer un fond de carte à un diagramme nœuds-liens (cf. l'exemple de la Figure 3.4). C'est le type de représentation le plus courant, mais également celui pour lequel les problèmes d'encombrement sont les plus fréquents et importants. Ces visualisations sont appelées cartes de flux.

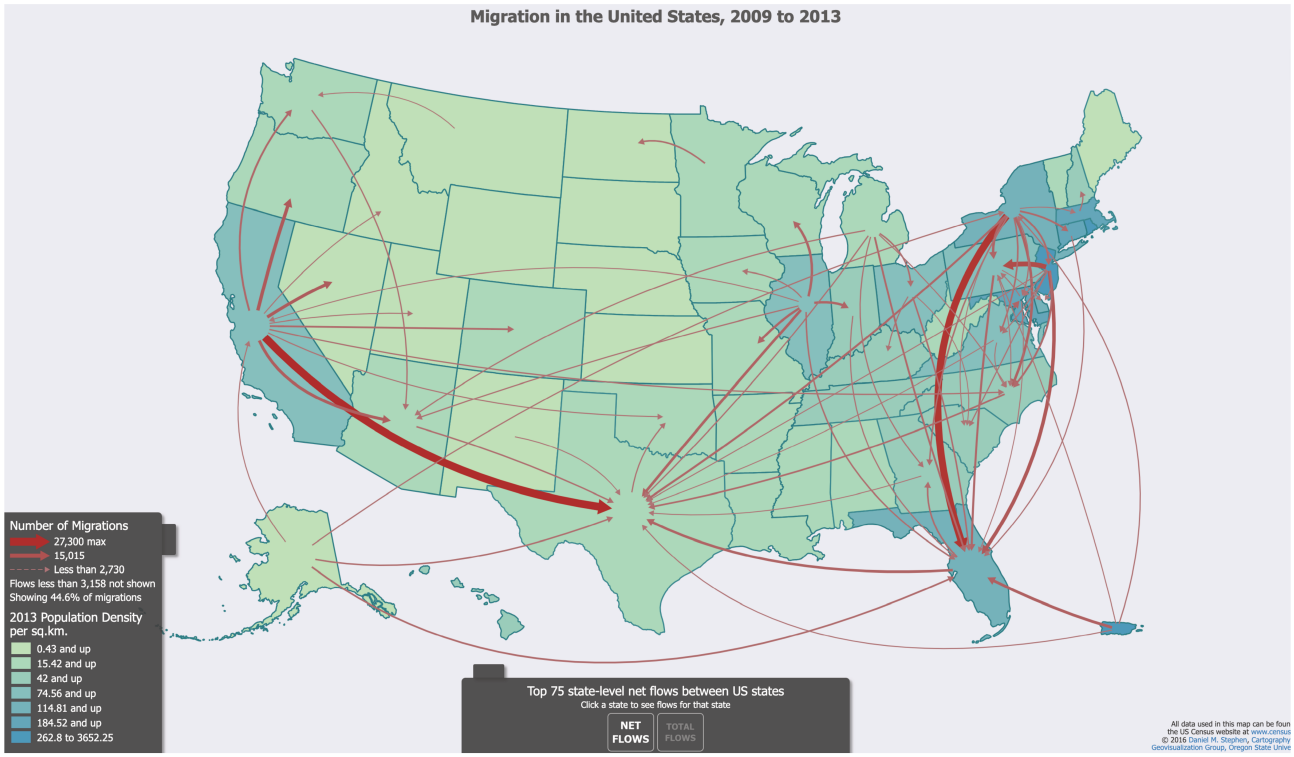


FIGURE 3.4 – Représentation de données entre des entités géographiques, avec une représentation de la géographie explicite, la plus courante : on superpose un fond de carte à un diagramme nœuds-liens (http://usmigrationflowmapper.com/).

Dans leur étude utilisateur, Yang et al. [101] comparent les cartes OD et les cartes de flux

avec leur nouvelle technique MapTrix (cf. Figure 3.5). Cette visualisation consiste à représenter de manière explicite les informations géographiques à l'aide de deux cartes identiques : l'une dédiée aux flux sortants et l'autre aux flux entrants. Entre ces deux cartes, une matrice d'adjacence est présente et tous ces éléments sont interconnectés par des segments. Les valeurs sont représentées dans la matrice au moyen d'une échelle de couleurs. Des interactions permettent, par exemple, de mettre en évidence les éléments associés à une entité d'origine.

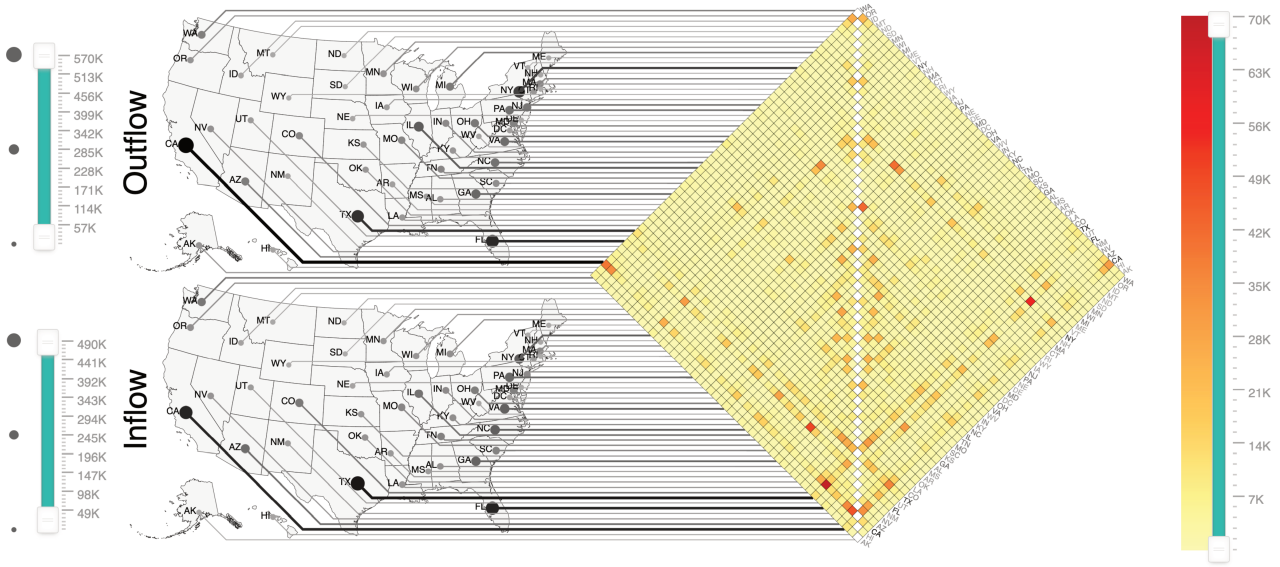


FIGURE 3.5 – Aperçu de MapTrix (https://vis.yalongyang.com/maptrix/index.html).

Les représentations où un diagramme nœuds-liens est superposé à un fond de carte restent tout de même les plus courantes, les cartes de flux se distinguent en effet des autres méthodes.

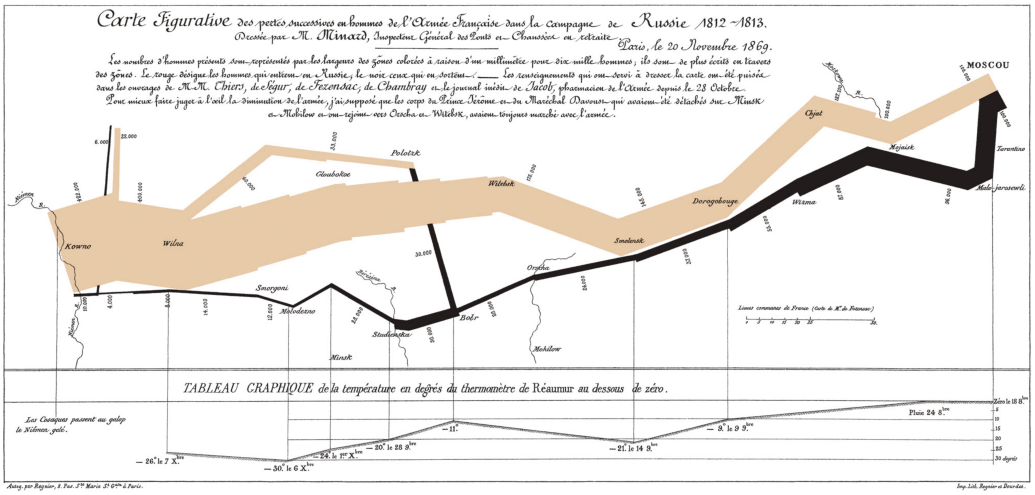
3.2.2 Cartes de flux

À notre connaissance, c'est l'ingénieur Henry Drury Harness qui est à l'origine de la première carte de flux (cf. Figure 3.6a). Cette carte a été publiée dans le cadre d'un rapport portant sur le potentiel de la construction de chemins de fer en Irlande et illustre la quantité de trafic de marchandises par route et par canal. Au fil des années, cette technique de visualisation des flux, basée sur la représentation des volumes par l'épaisseur des flux, a été explorée, notamment par Charles-Joseph Minard. L'une de ses visualisations les plus célèbres, élaborée en 1869, relate la campagne de Russie de 1812 (cf. Figure 3.6b). Minard a conçu bien d'autres visualisations de flux telles que les cartes de la Figure 3.1 présentées précédemment.

Les cartes de flux de Tobler à la fin des années 1980 [91] ont marqué le début de la création de telles cartes sur ordinateur. À cette époque, la qualité des visualisations produites était encore loin de celle des cartes élaborées à la main. Jenny et al. [47] ont défini des principes de conception pour les cartes de flux. Ces principes sont le fruit des conclusions d'une analyse quantitative ainsi que des résultats d'études utilisateur sur le dessin de graphes. Ils englobent des aspects tels que la réduction du chevauchement entre les flux, l'ajustement de la largeur des flux pour représenter la quantité, et bien d'autres encore. Sur la base de ces principes de conception, Jenny et al. [48] ont mis en place une méthode qui repose sur l'utilisation de modèles de force pour produire des cartes de flux OD où les flux sont représentés par des courbes de Bézier quadratiques. Les liaisons sont explicitement représentées et ne



(a) Carte de Henry Drury Harness représentant les quantités relatives de trafic dans différentes directions en Irlande en 1837. Cette carte est issue de l'Atlas accompagnant le 2e rapport des commissaires aux chemins de fer en Irlande en 1838 (`https://digital.ucd.ie/view/ivrla:45724`).

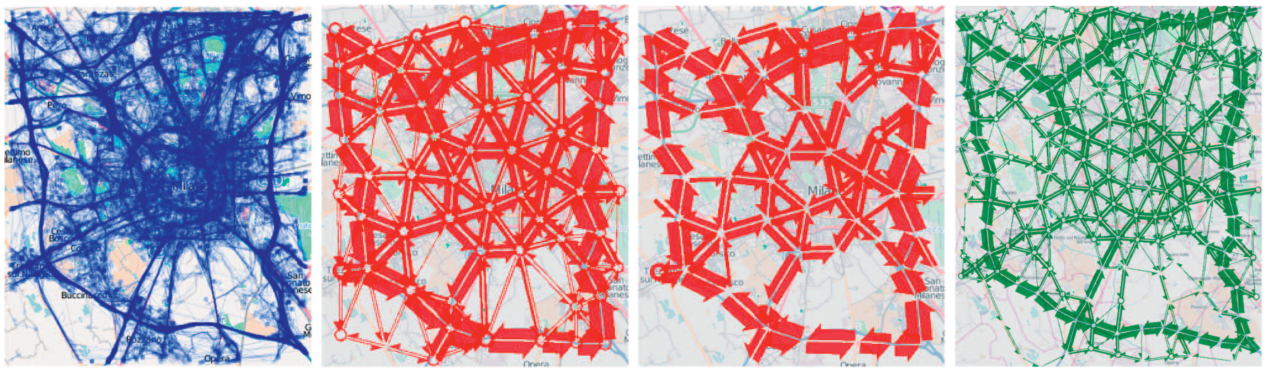


(b) Carte de Charles Joseph Minard représentant le nombre d'hommes dans l'armée française durant la campagne de Russie en 1812 [75].

FIGURE 3.6 – Les premières visualisations de flux élaborées à la main.

comportent pas de nœuds intermédiaires. Les cartes de Tobler et celles de Jenny et al. [91, 48] offrent une représentation aussi explicite que possible des liens. Toutes les informations sont présentes, mais avec l'augmentation du nombre de flux, ces cartes peuvent rapidement devenir encombrées. En effet, chaque flux est représenté de manière indépendante des autres. De plus, selon [6], la représentation des échanges par des lignes droites est insatisfaisante. Une telle représentation est trop approximative et ne correspond pas au mode de transport sous-jacent, comme le transport maritime, par exemple.

Les données à visualiser peuvent constituer un ensemble de trajectoires entre différentes entités géographiques. Cependant, lorsque la quantité de trajectoires devient importante, les représenter directement sur une carte devient rapidement laborieux. Il est possible de recourir à une méthode d'abstraction des données pour réduire les chevauchements entre les trajectoires, ce qui rendrait la visualisation moins encombrée, comme le fait Adrienko [1]. Cette méthode repose sur l'agrégation des données de mouvements, transformant ainsi des ensembles de trajectoires en flux agrégés entre des entités géographiques. Les épaisseurs de ces flux agrégés varient en fonction du nombre de trajectoires concernées (cf. Figure 3.7).



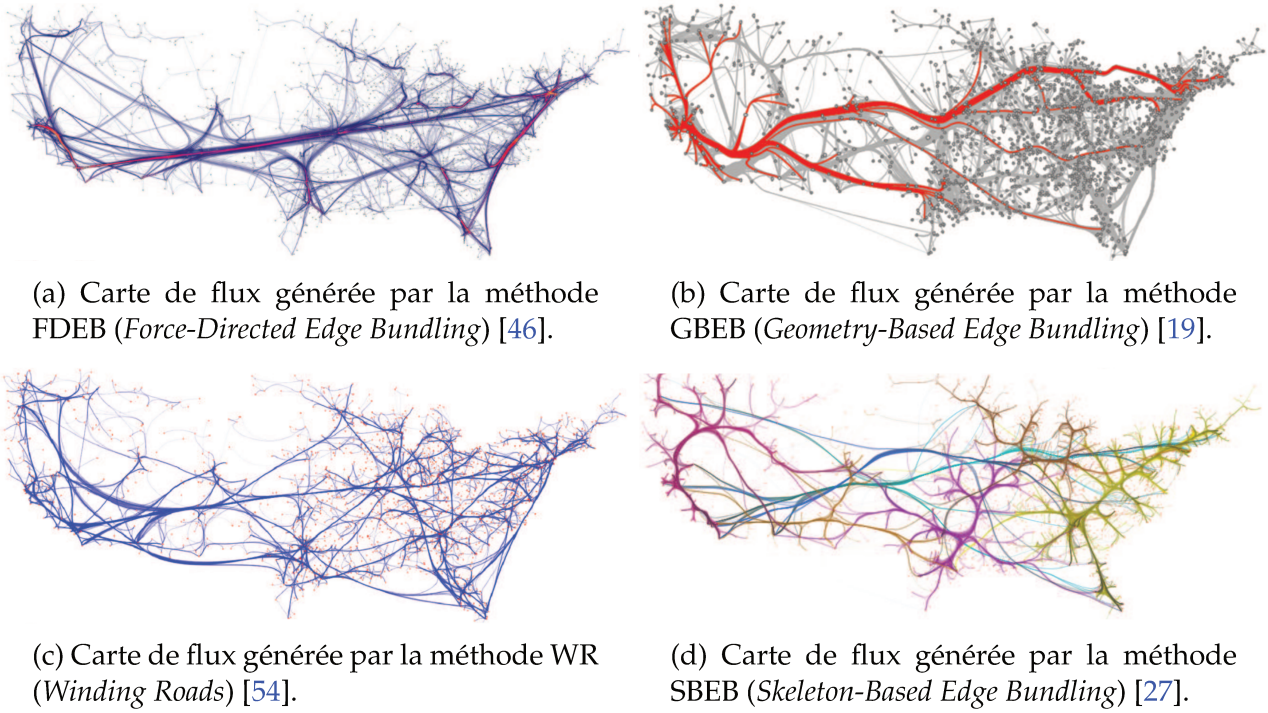
(a) (b) (c) (d)

FIGURE 3.7 – Cartes de flux obtenues par Adrienko [1] : (a) 6 187 trajectoires de voitures dessinées de manière explicite. (b) Les trajectoires sont agrégées, tous les flux sont visibles, le nombre maximal de trajectoires pour un flux est de 253. (c) Les flux sont agrégés, la valeur maximale est 253 et la minimale 50. (d) Ensemble des trajectoires agrégées avec un degré d'abstraction moins important, l'épaisseur maximale correspond cette fois-ci à 255 trajectoires.

La méthode proposée par Guo et Zhu [38] se concentre sur la compréhension des tendances complexes des flux, en particulier avec de grandes quantités de données. Pour ce faire, ils ont développé une méthode d'estimation de la densité des flux et une nouvelle technique de généralisation des cartes de flux. Des motifs de flux sont extraits et représentés sous forme de flux sur une carte. Une autre méthode similaire, également de Guo et Zhu [105], est hiérarchique et généralise les flux à différents niveaux tout en prenant en charge la multi-résolution cartographique des flux. Les flux sont regroupés en fonction de leur similitude, chaque groupe étant représenté par un flux représentatif. Il faut noter qu'après Tobler [91], la plupart des méthodes automatiques proposées pour résoudre le problème de l'encombrement des cartes consistaient à agréger les liens d'un graphe.

Afin de réduire le nombre de croisements d'arêtes, Phan et al. [74] ont proposé une méthode basée sur du clustering hiérarchique. Elle utilise le regroupement hiérarchique pour déterminer la disposition de la carte. Le regroupement crée un arbre reliant une racine à un ensemble de destinations. Cette approche vise à minimiser les croisements d'arêtes et modifie les positions des nœuds tout en préservant leurs positions relatives. La méthode GBEB

(*Geometry-Based Edge Bundling*) de Cui et al. [19] utilise un maillage de contrôle permettant de lancer le processus de collecte des arêtes. Ce dernier consiste à forcer toutes les arêtes à passer par des points de contrôle définis sur le maillage. Cependant, les résultats obtenus peuvent présenter des variations de courbure importantes (cf. Figure 3.8b). Une autre méthode de regroupement des arêtes pour les graphes généraux, nommée FDEB (*Force-Directed Edge Bundling*) et proposée par Holten et Van Wijk [46], permet d'éviter ce problème. Leur méthode basée sur les forces modélise les arêtes comme des ressorts flexibles qui s'attirent les uns les autres (cf. Figure 3.8a). Cependant, cette méthode est plus lente que la précédente. Debiasi et al. [20] ont amélioré la proposition de Holten et Van Wijk en développant une nouvelle méthode de génération automatique de cartes de flux qui décrit le mouvement et la force d'attraction ou de répulsion entre les points en utilisant les lois de la physique. Gansner et al. [32] ont également abordé le problème des courbures créées par le maillage de contrôle. Ils ont proposé une méthode de regroupement agglomérante basée sur le principe de minimisation de l'encre nécessaire pour représenter les arêtes, tout en imposant des contraintes supplémentaires sur la courbure des courbes résultantes. La méthode WR, proposée par Lambert et al. [54], regroupe également les arêtes pour réduire efficacement l'encombrement. Cette méthode se base sur une grille construite à partir du graphe original. À partir de cette grille les plus courts chemins sont calculés et les chemins les plus fréquents sont utilisés pour regrouper les arêtes. Une autre méthode, SBEB (*skeleton-based edge bundling*), proposée par Ersoy et al. [27], est applicable aux graphes généraux et aux graphes géographiques. Elle se base sur les axes médians, attirant progressivement les arêtes vers les lignes centrales pour finalement les agréger. Ersoy et al. compare sa méthode à GBEB [19], FDEB [46], et WR [54] (cf. Figure 3.8). SBEB s'est avéré être considérablement bien plus rapide que FDEB et légèrement plus rapide que WR.



(a) Carte de flux générée par la méthode FDEB (*Force-Directed Edge Bundling*) [46].

(b) Carte de flux générée par la méthode GBEB (*Geometry-Based Edge Bundling*) [19].

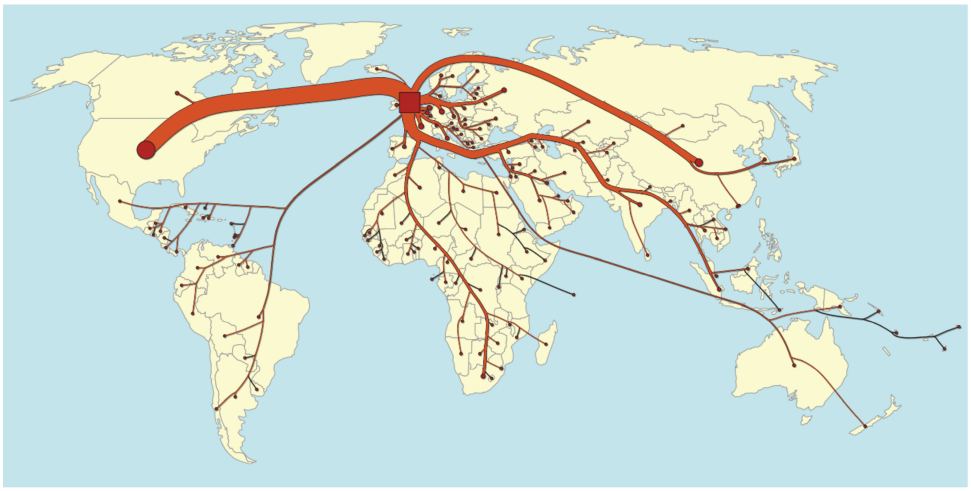
(c) Carte de flux générée par la méthode WR (*Winding Roads*) [54].

(d) Carte de flux générée par la méthode SBEB (*Skeleton-Based Edge Bundling*) [27].

FIGURE 3.8 – Comparaison de cartes de flux d'Ersoy et al. [27]. Les cartes ont été crées à partir d'un graphe géospatial qui agrègent les arêtes (*Edge Bundling*). Toutes les cartes représentent les mêmes données sur les migrations aux États-Unis.

D'autres méthodes visant à générer des cartes de flux un-à-plusieurs, sont basées sur un

arbre de Steiner, comme celle de Buchin et al. [12]. Cette méthode utilise un type particulier d'arbre de Steiner impliquant des spirales logarithmiques. Elle améliore le traçage des flux en évitant les croisements (cf. Figure 3.9a). Une autre approche pour créer une carte de flux un-à-plusieurs basée sur un arbre de Steiner a été proposée par Sun [87]. Elle consiste à trianguler l'espace pour créer un réseau, puis à utiliser cette triangulation pour définir un arbre de Steiner. Ensuite, une série d'opérations de cartographie automatique est réalisée pour améliorer les lignes représentant les flux et leur attribuer des couleurs en fonction de leur valeur (cf. Figure 3.9b).



(a) Carte de flux générée par la méthode de Buchin et al. [12].



(b) Carte de flux générée par la méthode de Sun [87].

FIGURE 3.9 – Exemples de carte de flux un-à-plusieurs se basant sur un arbre de Steiner.

Les cartes de flux peuvent également être visualisées en 3D [20, 53] et certaines d'entre elles sont même intégrées dans des environnements de réalité virtuelle [102, 104]. Dans ces représentations, les liens entre les entités géographiques ont été agrégés. Elles affichent directement le globe terrestre, ce qui en fait la représentation géographique la plus explicite possible. En effet, toutes les informations géographiques, y compris les positions exactes, les distances et les surfaces, sont conservées dans ces visualisations.

Le Tableau 3.1 synthétise l'ensemble des méthodes présentées pour visualiser des données quantitatives entre des entités géographiques. La méthode basée sur les distances entre les entités pour représenter les valeurs associées à leur relation [4, 8] n'y est pas mentionnée car il s'agit d'un cas bien particulier. Il ne s'agit ni d'une matrice ni d'un diagramme nœuds-liens. Cependant, cette méthode est pertinente car elle permet de visualiser une me-

sure à travers la distance entre les entités géographiques, comme les temps de trajets illustrés dans l'exemple de la Figure 3.3a. La carte de flux de Phan et al. [74] n'est pas non plus incluse dans le tableau car sa représentation géographique n'est ni explicite en raison du repositionnement des nœuds, ni déformée puisque les informations géographiques ne pas représentées. En effet, contrairement aux autres méthodes de edge bundling, on ne peut pas superposer le diagramme nœuds-liens à une carte. Des étiquettes sont ajoutées aux nœuds, c'est-à-dire aux entités géographiques, afin de les identifier. Cette méthode diffère de ce que nous souhaitons réaliser car elle ne représente pas les informations géographiques de manière explicite.

À partir du Tableau 3.1 nous pouvons constater que la plupart des visualisations visent à conserver les informations géographiques sans les déformer, en les superposant à un diagramme nœuds-liens, ce qui les classe donc comme des cartes de flux. Il y a moins de méthodes basées sur l'utilisation de matrices. Nous n'avons pas trouvé de méthode similaire à celle que nous proposons, c'est-à-dire un diagramme nœuds-liens superposé à une carte déformée.

Tableau 3.1 – Récapitulatif des techniques existantes pour visualiser des valeurs quantitatives associées à des relations entre des entités géographiques.

Représentation géographique	Matrice	Diagramme nœuds-liens
Abstrait	Matrice d'adjacence [5]	Diagrammes de membrures [39, 44] Kriskogramme [99]
Explicite	MapTrix [101]	Edge Bundling [19, 20, 27, 32, 46, 54] Abstraction des données [1, 38, 105] Arbre de Steiner [12, 87] Représentation des liens explicite [48, 91] 3D [20, 53, 102, 104] MapTrix [101]
Déformé	Carte OD [98]	?

3.2.3 Distorsion de carte

La distorsion de carte à laquelle nous faisons référence se rapproche le plus des cartogrammes et du fish-eye. Les cartogrammes ont déjà été présentés dans le chapitre précédent, dans la Section 2.2 , nous nous concentrerons donc uniquement sur les techniques de fish-eye dans cette section.

En photographie, un objectif fish-eye est un objectif à très grand angle qui présente en détail les objets proches tout en montrant les objets plus éloignés avec moins de détails. Cette notion a été adaptée aux graphes par Sarkar et Brown [81].

En cartographie, un fish-eye signifie que la projection de la carte est à échelle variable. Les vues fish-eye sont particulièrement utiles pour se concentrer sur une zone particulière contenant de nombreuses informations tout en conservant une vision des éléments plus éloignés. La zone d'intérêt, appelée focus, est agrandie mais les environs, le contexte, restent

visibles. Cette approche est connue sous le nom de fish-eye *Focus+Context*. La difficulté réside dans la capacité à donner accès à toutes les informations détaillées de la zone d'intérêt tout en préservant le contexte. La vue fisheye *Focus+Context* a été proposée comme solution pour la navigation personnelle en temps réel par Harrie et al. [41]. L'utilisateur a besoin à la fois d'une vue d'ensemble et de détails supplémentaires autour de sa position. Cette méthode utilise une carte évolutive qui s'ajuste en temps réel pour que la position de l'utilisateur soit toujours mise en avant à grande échelle. Cependant, les résultats se détériorent avec l'augmentation de l'intensité du grossissement. Carpendale et al. [15] ont présenté une série d'améliorations pour un fort grossissement en introduisant de nouvelles techniques de distorsion. Le principal inconvénient des vues fish-eye *Focus+Context* réside dans la déformation de l'ensemble de la carte et dans le coût de calcul élevé. De plus, la densité d'informations devient excessive dans les coins de la carte.

Takahashi [88] a amélioré ce type de vue fish-eye en introduisant le concept de *Focus+Glue+Context*. Une zone de "collage" (*glue*) est incluse entre les zones de mise au point (focus) et de contexte et les distorsions dans ces zones de mise au point et de contexte sont complètement éliminées. La zone de mise au point est affichée à une grande échelle pour présenter les détails de la carte tandis que la zone de contexte est à une petite échelle. La zone de collage relie la zone d'intérêt au contexte. La Figure 3.10 présente la vue fish-eye *Focus+Glue+Context*, on peut constater que, contrairement à la vue constituée d'un simple grossissement, la zone de mise au point est bien mise en avant sans pour autant restreindre la visibilité du contexte. Yamamoto et al. [100] ont proposé une méthode de génération de cartes de haute qualité basée sur ce concept, avec des vues *Focus+Glue+Context* qui peuvent être utilisées dans les services de cartographie en ligne en raison de leur rapidité de génération. De même, Haunert et Sering [43] tentent d'améliorer les vues fish-eye en cartographie en soutenant que la distorsion devrait de préférence avoir lieu dans les zones où la densité d'information est plus faible. Aucune fonction de cartographie prédéfinie n'est donc appliquée. Cette méthode repose sur une approche d'optimisation visant à minimiser les distorsions tout en respectant un ensemble de contraintes et d'objectifs.

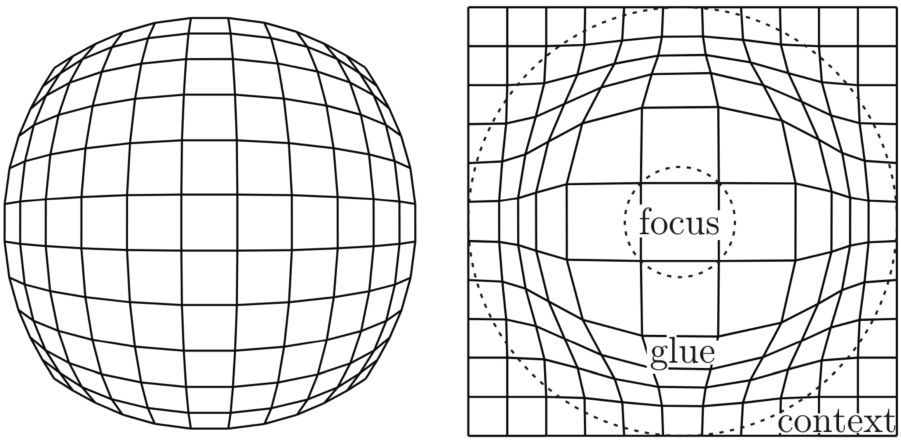


FIGURE 3.10 – Comparaison de deux vues [43]. À gauche, un simple grossissement est fait tandis qu'à droite il s'agit du concept *Focus+Glue+Context*.

L'objectif de notre méthode de distorsion des cartes de flux est très similaire à celui du fish-eye : effectuer un ensemble de grossissements dans des zones particulières tout en gardant l'ensemble de la carte visible.

3.3 Algorithme de distorsion FMD (Flow Map Distortion)

Comme nous l'avons vu, la visualisation des flux en utilisant la largeur pour représenter une quantité génère des chevauchements entre les flux et les zones géographiques. Dans cette section, nous proposons une méthode pour supprimer ces chevauchements d'une carte de flux un- à-plusieurs existante sans modifier la largeur des flux. Par la suite, nous illustrons notre approche à l'aide de flux sur des zones maritimes. La Figure 3.11 illustre les étapes de cette méthode : une première déformation, puis le tracé de courbes, et enfin une seconde déformation qui supprime complètement les chevauchements.

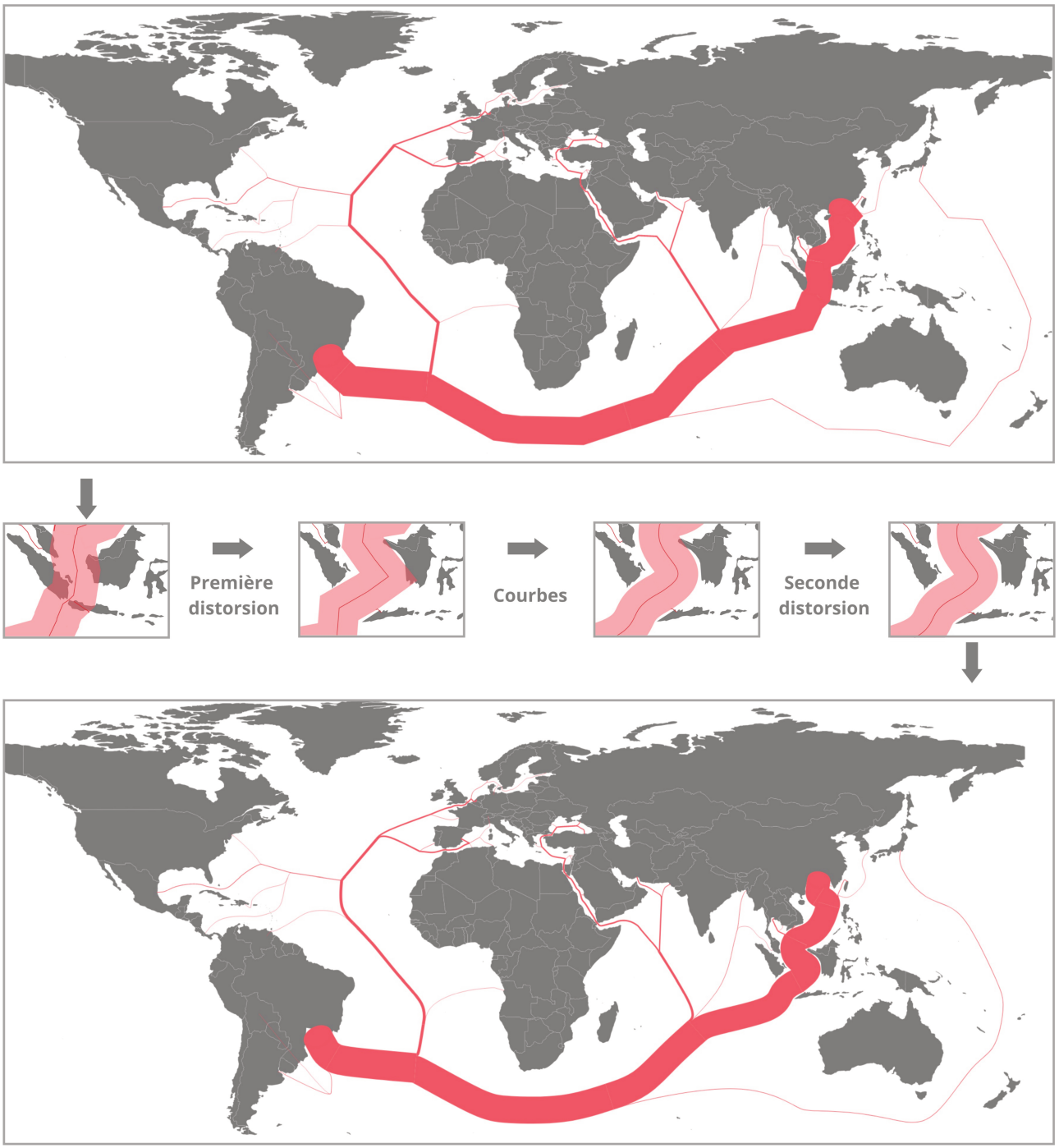


FIGURE 3.11 – Aperçu des principales étapes de la méthode FMD.

3.3.1 Première distortion

La méthode de distorsion mise en place consiste à faire un ensemble de grossissements dans les zones maritimes pour lesquelles l'espace n'est pas suffisant pour faire passer des flux d'une certaine épaisseur. En créant l'espace suffisant, il n'y a alors plus de chevauchement. Pour faire un grossissement, nous nous sommes basés sur la formule proposée par Kadmon et Shlomi [50] pour ses projections polyfocales. Celles-ci consistent à redéfinir la projection cartographique afin de créer un ensemble de grossissements comme illustré dans la Figure 3.12.

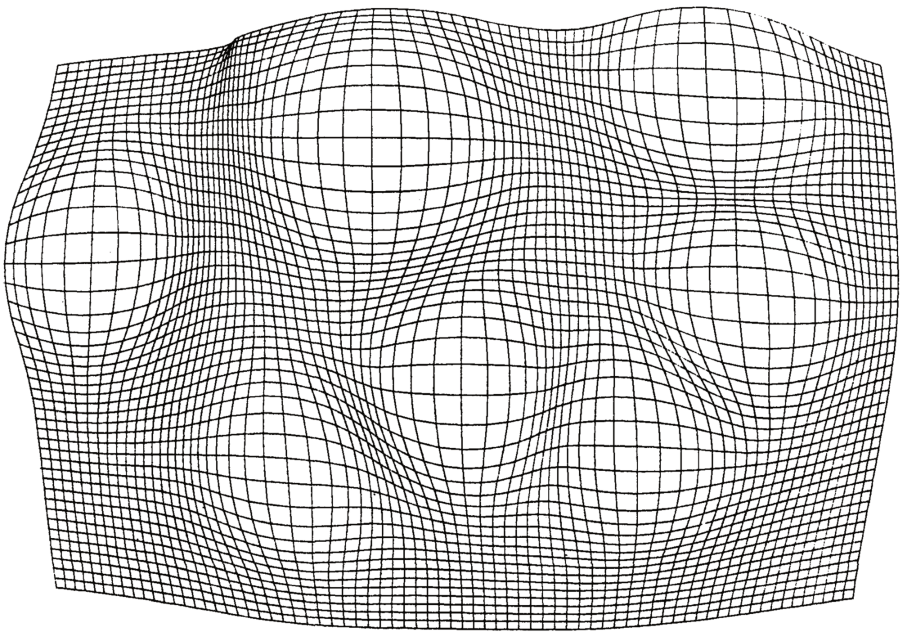


FIGURE 3.12 – Projection polyfocale de Kadmon [50].

Considérons le point focal, le point central du grossissement, de coordonnées (x_f, y_f) et un autre point p de coordonnées (x, y) à déplacer. D est la distance entre p et le point focal. Les nouvelles coordonnées de p sont notées (x', y'), et définies comme suit :

$$\begin{aligned} x' &= x + \frac{A(x - x_f)}{1 + C \cdot D^2} \\ y' &= y + \frac{A(y - y_f)}{1 + C \cdot D^2} \end{aligned} \tag{3.1}$$

La valeur de A est le niveau de grossissement et C le taux de variation de A. Plus la valeur de C est élevée, plus le grossissement sera localisé et donc à proximité du point focal. À l'inverse, moins la valeur de C est élevée, plus le grossissement se généralise dans l'entièreté de la carte. Dans notre cas, le grossissement doit être localisé donc on fixe $C = 1$.

Étant donnée que nous nous concentrons sur les cartes de flux un-à-plusieurs, nous pouvons utiliser un arbre pondéré où la racine correspond à l'entité géographique de départ et les feuilles à l'ensemble des destinations. Les nœuds internes sont les points de contrôles de la carte d'origine. Le poids des arêtes permet de définir l'épaisseur des flux (comme à gauche dans la Figure 3.13). Pour effectuer les grossissements, on divise dans un premier temps l'arbre en un ensemble de chemins, que l'on nomme séquence, tels que chaque nœud constituant la séquence soit de degré 2 (à droite dans la Figure 3.13). On attribue alors à chaque séquence un poids correspondant au poids d'une de leurs arêtes (les poids de toutes

les arêtes d'une même séquence sont égaux, voir dans la Figure 3.13). Considérons un ensemble S de séquences de l'arbre tel que $S = \{S_1, S_2, ..., S_s\}$. $\forall S_i$ avec $1 \leq i \leq s$, w_i est le poids de la séquence.

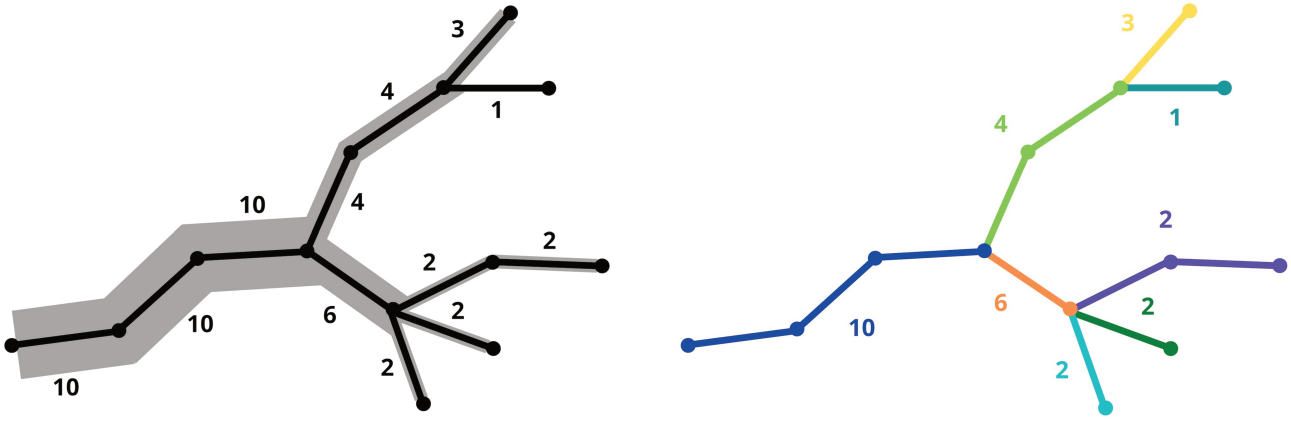


FIGURE 3.13 – Arbre (à gauche) et ses séquences (à droite). Chaque couleur correspond à une séquence (8 au total) et le nombre au poids de la séquence.

La méthode consiste, durant la première distorsion, à traiter l'ensemble de séquences S dans l'ordre décroissant du poids des arêtes. Pour chacune de ces séquences, nous disposons d'un ensemble de nœuds. S'il existe des chevauchements entre la séquence et une entité géographique, le nœud le plus proche d'un sommet d'un polygone d'une entité géographique est recherché. Ce nœud devient alors un point focal et nous appliquons la formule (cf. Algorithme 1) à tous les sommets des polygones correspondant aux entités géographiques et à tous les nœuds de la séquence. Ce processus est répété jusqu'à ce qu'il n'y ait plus de chevauchement (cf. Algorithme 2).

Algorithme 1 Créer un nouveau grossissement

Entrée : P un ensemble de points
Sortie : P' un ensemble de points après grossissement

1: **fonction** DÉFORMER(P, $focal$, A, $C = 1$)
2: **pour chaque** $p \in P$ **faire**
3: $D \leftarrow \sqrt{(x_p - x_{focal})^2 + (y_p - y_{focal})^2}$
4: $x'_p \leftarrow x_p + \frac{A \times (x_p - x_{focal})}{1 + C \times D^2}$
5: $y'_k \leftarrow y_p + \frac{A \times (y_p - y_{focal})}{1 + C \times D^2}$
6: **fin pour**
7: **retourner** P'
8: **fin fonction**

Considérons D' la distance entre le point focal et le sommet d'un polygone le plus proche après la transformation. R est l'épaisseur de la séquence divisé par 2. Pour supprimer les chevauchements, on doit trouver la valeur de A_{min}, la valeur minimale de A pour avoir $D' \geq R$.

Algorithme 2 Première distorsion

Entrée : S l'ensemble des séquences, K l'ensemble des sommets formant les polygones des entités géographiques, $\epsilon \in]0,1[$ un facteur d'amortissement

Sortie : S l'ensemble des séquences avec de nouvelles coordonnées, K l'ensemble des sommets formant les polygones des entités géographiques avec de nouvelles coordonnées

pour chaque $S_i \in S$ **faire**
  $chevauchement \leftarrow True$
  $r_i \leftarrow \frac{w_i}{2}$
  tant que $chevauchement$ **faire**
    $distMin \leftarrow \{\min distance(u,k) | \forall u \in S_i, \forall k \in K\}$
    si $distMin > r_i$ **alors**
      $chevauchement \leftarrow False$
    fin si
    $focal \leftarrow \{u \in S_i | \min distance(u,k)\ \forall u \in S_i, \forall k \in K\}$
    $A_{min} \leftarrow$ solution de l'équation (3.6)
    $A \leftarrow A_{min} \times \epsilon$
    $S \leftarrow$ DÉFORMER$(S, focal, A)$
    $K \leftarrow$ DÉFORMER$(K, focal, A)$
  fin tant que
fin pour

$$
\begin{aligned}
D' \geq R \Leftrightarrow & \sqrt{\left[\left(x + \frac{A_{min}(x - x_f)}{1 + C \cdot D^2}\right) - x_f\right]^2 + \left[\left(y + \frac{A_{min}(y - y_f)}{1 + C \cdot D^2}\right) - y_f\right]^2} \geq R \\
\Leftrightarrow & \left[\left(x + \frac{A_{min}(x - x_f)}{1 + C \cdot D^2}\right) - x_f\right]^2 + \left[\left(y + \frac{A_{min}(y - y_f)}{1 + C \cdot D^2}\right) - y_f\right]^2 \geq R^2
\end{aligned}
\tag{3.2}
$$

En isolant A_{min} et en développant la formule on obtient :

$$
\left[\left(x + \frac{A_{min}(x - x_f)}{1 + C \cdot D^2}\right) - x_f\right]^2 = \left(\frac{x - x_f}{1 + C \cdot D^2}\right)^2 A_{min}^2 + \left(\frac{2(x - x_f)^2}{1 + C \cdot D}\right) A_{min} + (x - x_f)^2 \tag{3.3}
$$

$$
\left[\left(y + \frac{A_{min}(y - y_f)}{1 + C \cdot D^2}\right) - y_f\right]^2 = \left(\frac{y - y_f}{1 + C \cdot D^2}\right)^2 A_{min}^2 + \left(\frac{2(y - y_f)^2}{1 + C \cdot D}\right) A_{min} + (y - y_f)^2 \tag{3.4}
$$

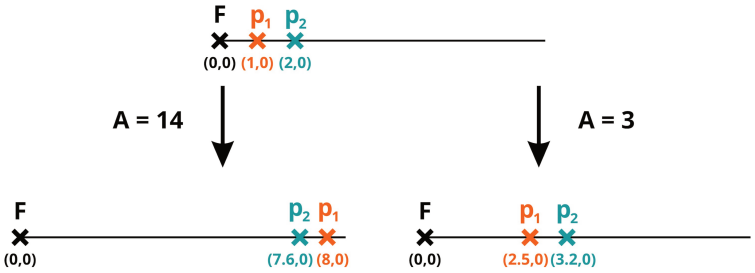
En additionnant (3.3) et (3.4) on obtient alors un polynôme du second degré de la forme $\alpha A_{min}^2 + \beta A_{min} + \gamma$ avec :

$$
\begin{aligned}
\alpha &= \left(\frac{x - x_f}{1 + C \cdot D^2}\right)^2 + \left(\frac{y - y_f}{1 + C \cdot D^2}\right)^2 \\
\beta &= 2\frac{(x - x_f)^2 + (y - y_f)^2}{1 + C \cdot D} \\
\gamma &= (x - x_f)^2 + (y - y_f)^2
\end{aligned}
\tag{3.5}
$$

Il ne reste alors plus qu'à résoudre l'équation ci-dessous dont le résultat correspond à la valeur minimale de A permettant de supprimer tout chevauchement :

$$\alpha A_{min}^2 + \beta A_{min} + \gamma - R^2 = 0 \tag{3.6}$$

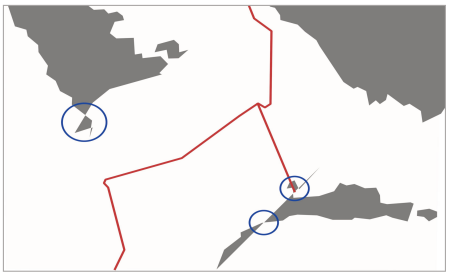
Initialement, des séquences très épaisses peuvent passer par des zones étroites comme celle de la Figure 3.14b. Supprimer les chevauchements en fixant $A = A_{min}$ en une unique itération peut créer des croisements au sein même des polygones (cf. Figure 3.14a et Figure 3.14c). C'est pour cette raison que l'on atténue la valeur A à l'aide d'un facteur d'amortissement $\epsilon \in]0,1[$ en définissant A comme suit : $A = A_{min} \times \epsilon$. La déformation est modérée et on conserve des polygones dits simples, sans croisements [2]. Les grossissements se font donc de manière récursive et progressive. (Figure 3.14d).



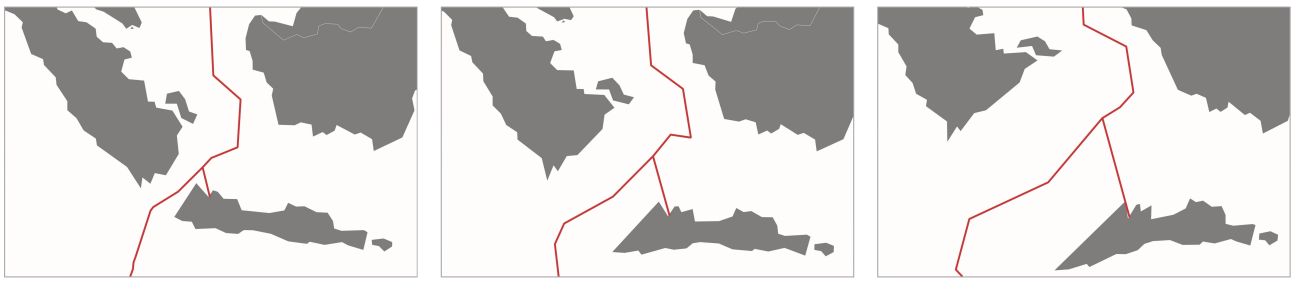
(a) Exemple pour repositionner les points p_1 et p_2 avec $A = 14$ et $A = 3$. F est le point focal. Dans le premier cas, la valeur de A est trop élevée : p_1 et p_2 ne conservent pas leur position relative.



(b) Entités géographiques initiales, sans distrosion.



(c) Exemple de distorsion en une itération avec $A = A_{min}$: des croisements (entourés en bleu) sont crées.



(d) Exemple de distorsion récursive : pour chaque grossissement, on a $A = A_{min} \times 0.2$. Aucun croisement n'est crée. Dans cet exemple, toutes les itérations ne sont pas illustrées.

FIGURE 3.14 – Comparaison des résultats obtenus après une distorsion en une itération avec $A = A_{min}$ et une distorsion récursive avec $A = A_{min} \times 0.2$.

2. Nous avons fixé $\epsilon = 0.2$. Empiriquement, cette valeur donnait les meilleurs résultats en ne créant aucun croisement sans pour autant trop ralentir le processus.

Les nœuds ne sont pas traités dans un ordre séquentiel et un même nœud peut être le point focal d'un grossissement plusieurs fois. Cette première étape de distorsion ne se termine que lorsque les chevauchements ont été complètement supprimés, mais il ne faut pas oublier que le niveau de grossissement est réduit. Nous avons constaté empiriquement que cela empêchait la création de croisements dans les polygones des entités géographiques. En outre, l'ordre séquentiel entraînait toujours le déplacement des entités géographiques dans la même direction que la trajectoire du flux, ce qui créait plus de distorsion que nécessaire. Un autre avantage du traitement du nœud le plus proche d'un sommet à chaque itération est qu'un nœud n'est jamais traité lorsqu'une distorsion n'est pas nécessaire.

Considérons l le nombre total de sommets formant les polygones des entités géographiques, m la taille de l'arbre, et n le nombre de séquences. La complexité en temps pour vérifier s'il y a des chevauchements est $O(c.l.m)$, avec c le nombre d'itérations. Pour définir le point focal, la complexité en temps est aussi $O(c.l.m)$. Pour la distorsion c'est $O(c.m.n + c.l.n)$. Comme le nombre de séquences est inférieur à la taille de l'arbre, $n.l$ est inférieur à $m.n$ et la complexité en temps de la première distorsion est alors $O(c(l.m + m.n))$. S'il y a beaucoup de séquences dans l'arbre T, autrement dit si n tend vers m, alors la taille des séquences diminue et tend vers 1. La complexité se rapproche alors de $O(c(l.m + m^2))$.

À la fin de cette première étape de distorsion, les chevauchements ont été supprimés tout en conservant des caractéristiques géographiques reconnaissables et cohérentes. Nous passons ensuite au dessin des courbes afin de fluidifier l'aspect des flux.

3.3.2 Courbes

Durant cette seconde étape, nous transformons chaque séquence de l'arbre en une courbe. Cette étape est réalisée juste après la première distorsion car la création des courbes implique des modifications de l'arbre. Et modifier l'arbre sans avoir créé de l'espace au préalable pourrait provoquer des croisements entre l'arbre et les entités géographiques.

Pour cette étape, nous avons testé et comparé différentes courbes. Notre méthode a obtenu les meilleurs résultats avec la Spline de Catmull-Rom et la Spline de base (B-spline).

La Spline de Catmull-Rom utilise des points de contrôle et un paramètre α. Son comportement dépend fortement du choix de la valeur de α. Si $\alpha = 0$, des auto-intersections peuvent se produire, et si $\alpha = 1$, la courbe s'éloigne trop du polygone de contrôle [103]. Dans notre cas, $\alpha = 0,5$ semble être le plus approprié. Le principal avantage de cette courbe par rapport à la B-spline est qu'elle passe par les points du polygone de contrôle, c'est-à-dire la trajectoire du flux. Cela réduit les risques de créer de nouveaux chevauchements puisque la courbe est tracée très près de la trajectoire originale. Cependant, la B-spline a l'avantage de lisser la trajectoire davantage que la Catmull-Rom.

La B-spline est une spline cubique de base redressée avec des points de contrôle. Holten [45] a proposé de l'utiliser dans le regroupement hiérarchique des arêtes. La spline est redressée en fonction du paramètre β de la courbe. $\beta = 1$ est le plus adapté à nos besoins (cf. Figure 3.15). Nous avons donc opté pour ce type de courbe, qui lisse les chemins sans les superposer aux entités géographiques.

3.3.3 Deuxième distorsion

Durant les tests, nous avons remarqué que le tracé des courbes pouvait créer de légers chevauchements dans certains cas. Une deuxième distorsion est donc nécessaire afin de s'assurer que tous les chevauchements sont supprimés. Pour cette seconde distorsion, on utilise

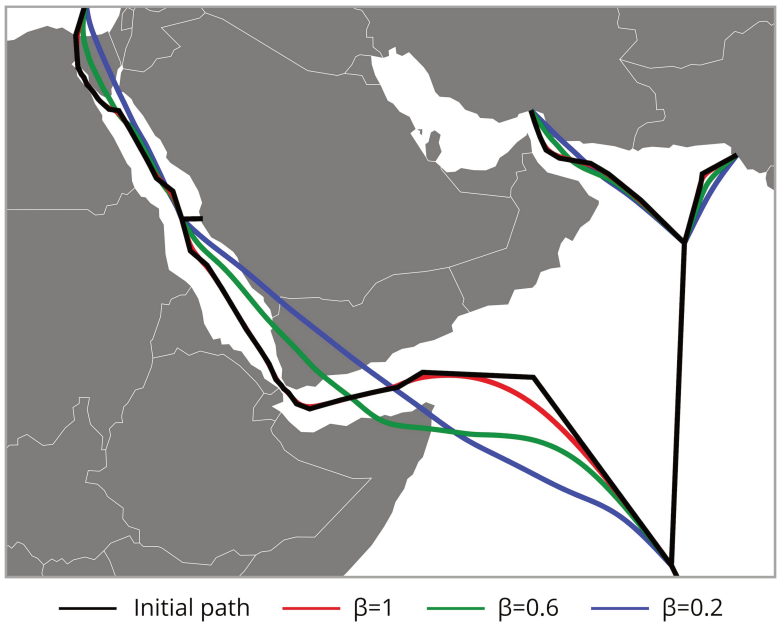


FIGURE 3.15 – Comparaison des différentes valeurs de β pour générer des courbes B-spline.

de nouveau la formule (3.1) et on résout l'équation (3.6) qui ont été présentées dans la Section 3.3.1 (cf. Algorithme 3).

Algorithme 3 Deuxième distorsion

Entrée : P l'ensemble des courbes, K l'ensemble des sommets formant les polygones des entités géographiques

Sortie : P l'ensemble des courbes déplacées, K l'ensemble des sommets formant les polygones des entités géographiques avec de nouvelles coordonnées

pour chaque $P_i \in P$ **faire**
  pour chaque $p \in P_i$ **faire**
    $dist \leftarrow \min distance(p,k), \forall k \in K$
    $r_i \leftarrow \frac{w_i}{2}$
    si $dist < r_i$ **alors**
      $focus \leftarrow p$
      $A_{min} \leftarrow$ solution de l'équation (3.6)
      $A \leftarrow A_{min}$
      $P \leftarrow$ DÉFORMER$(P, focus, A)$
      $K \leftarrow$ DÉFORMER$(K, focus, A)$
    fin si
  fin pour
fin pour

L'algorithme est presque le même que pour la première distorsion. Ils se différencient au niveaux des points focaux : ils ne sont pas définis à partir des nœuds des séquences, mais à partir des points le long des courbes de chaque séquence. En effet, l'ordre est cette fois-ci séquentiel. À ce stade, peu de changements sont encore nécessaires, c'est pour cette raison qu'une seule itération sur un point est très largement suffisant pour éliminer complètement toute superposition.

Considérons l le nombre total de sommets formant les polygones des entités géographiques, m le nombre de points dans les courbes. La complexité en temps pour vérifier s'il y

a des chevauchements est $O(l.m)$. Pour la distorsion c'est $O(l.m + m^2)$, et donc la complexité temporelle de la seconde distorsion est également $O(l.m + m^2)$.

À l'issue de ces trois étapes, nous avons obtenu des cartes de flux sans chevauchement. La seule condition à respecter initialement est qu'aucune arête de l'arbre ne croise une entité géographique. Différentes méthodes peuvent être utilisées pour créer une carte de flux. Dans la section suivante, nous en proposons une qui garantit aucun croisement entre les entités et l'arbre et qui limite les surfaces de chevauchement. Ainsi, le nombre de grossissements à effectuer est réduit lors de la déformation des entités géographiques.

3.4 Création de carte de flux un-à-plusieurs

Dans cette section, nous proposons une nouvelle méthode pour créer des cartes de flux un-à-plusieurs. Créer les flux à partir de cette méthode permet d'optimiser l'algorithme de distorsion FMD bien qu'il soit possible de l'appliquer à partir de cartes de flux issues de méthodes différentes. Les flux sont dessinés le long des axes médians des entités géographiques, ce qui empêche de créer des croisements entre eux et réduit les chevauchements avec les entités géographiques. Ainsi, le nombre d'itérations dans le processus FMD est minimisé tout comme la déformation des entités géographiques qui en résulte.

Nous allons maintenant présenter les étapes de notre méthode pour créer une carte de flux un-à-plusieurs : (1) la création d'un graphe connexe à partir d'un squelette (cf. Figure 3.16a et Figure 3.16b), (2) la définition d'un arbre représentant les flux (cf. Figure 3.16c), (3) une simplification de cet arbre, et (4) le dessin des flux (cf. Figure 3.16d).

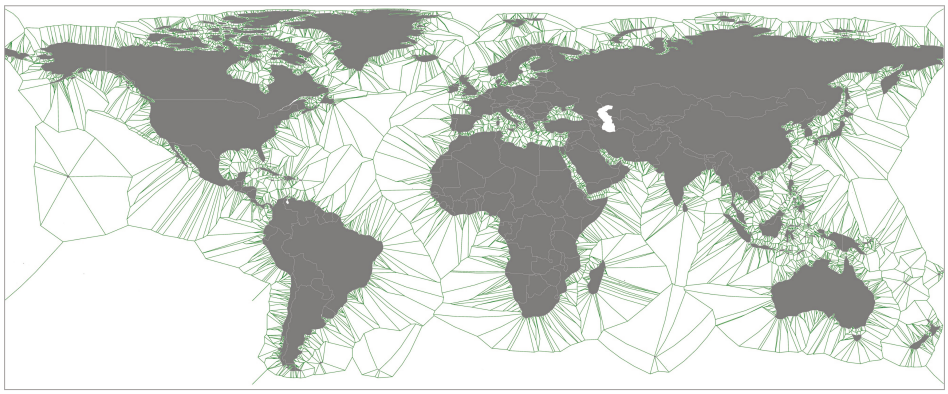
3.4.1 Création du graphe initial

La première étape de notre méthode consiste à créer un graphe, à partir de la carte de base, dont les arêtes se situent le plus loin possible des entités géographiques. Ce graphe doit contenir des chemins reliant l'ensemble des entités géographiques, il est donc connexe. Ce graphe sera ensuite utilisé pour définir le tracé des flux (cf. Section 3.4.2). Notre méthode est basée sur un squelette droit (cf. Figure 3.16a). Introduits par Aichholzer et al. puis généralisés [3, 2], les squelettes droits (*straight skeletons*) peuvent être considérés comme des graphes connexes dont les nœuds incluent les sommets des polygones (dans notre cas, les sommets formant les côtes des entités géographiques) et dont les arêtes se situent à proximité des axes médians entre les polygones des entités géographiques. Contrairement aux axes médians, qui peuvent contenir des courbes paraboliques, les squelettes droits sont composés exclusivement de segments droits. Felkel et Obdrzalek[29] a proposé une méthode appliquée aux polygones simples et Eppstein avec Erickson [26] une méthode qui construit un squelette droit amélioré.

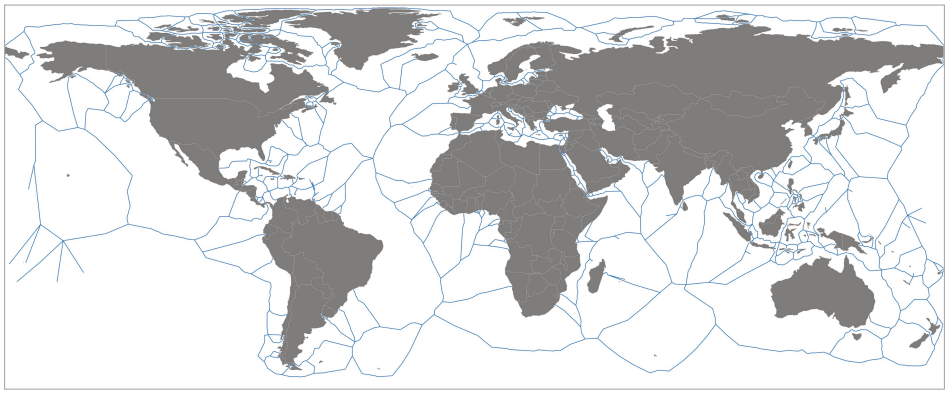
Dans notre mise en œuvre, nous avons utilisé une bibliothèque,[3] qui nous donne un squelette comme illustré dans la Figure 3.16a. Nous avons construit un squelette droit à partir d'un polygone avec des "trous" qui correspondent aux faces internes des polygones formant les entités géographiques (c'est-à-dire la terre, les continents).

Comme les pays côtiers comportent plusieurs sommets, il existe plusieurs chemins menant à chaque pays du squelette. Pour simplifier ce dernier, nous sélectionnons un sommet par pays côtier (dans notre implémentation, le sommet le plus proche du plus grand port

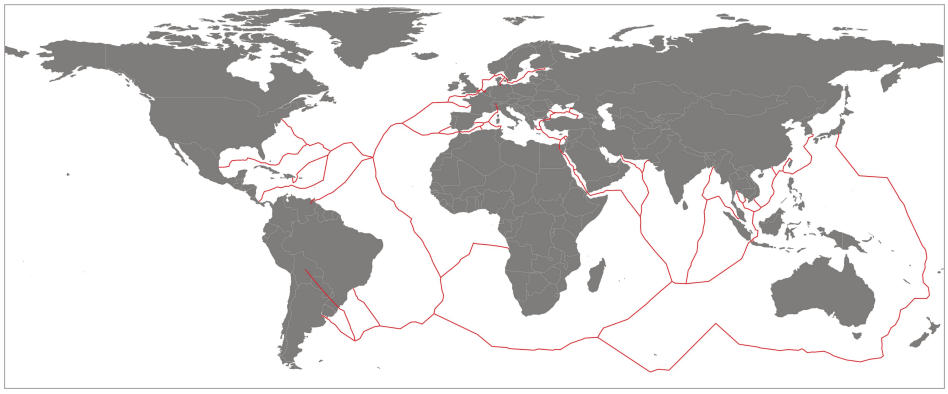
3. The Computational Geometry Algorithms Library (CGAL) : `https://www.cgal.org` [dernière consultation le 04/08/2023]



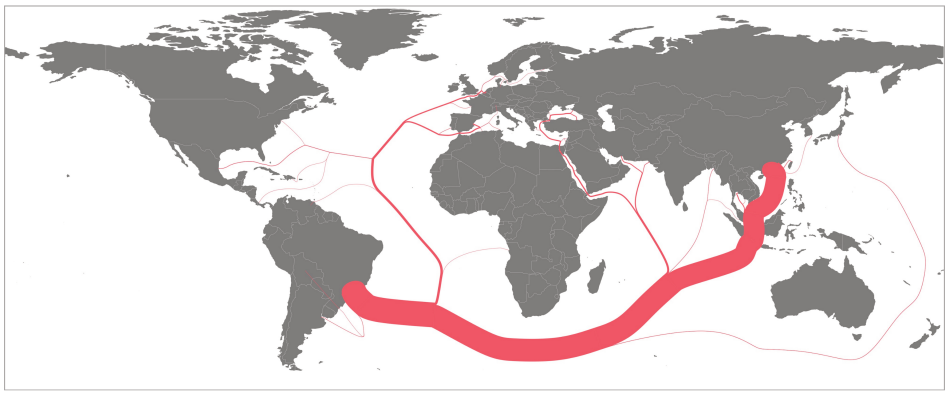
(a) Squelette



(b) Graphe



(c) Arbre



(d) Carte de flux

FIGURE 3.16 – Étapes pour créer une carte de flux un-à-plusieurs

du pays) et supprimons itérativement les sommets de degré 1 non sélectionnés (et les arêtes correspondantes). Nous obtenons ainsi un graphe G, illustré par la Figure 3.16b.

3.4.2 Définition de l'arbre

La deuxième étape de notre méthode consiste à faire correspondre les flux au graphe G afin de créer les chemins de flux. Lors de l'étape précédente, on a besoin uniquement des polygones du fond de carte. À partie de cette étape, le jeu de données que l'on souhaite visualiser est pris en compte. Les données doivent se composer d'une origine (le pays de départ), d'un ensemble de cibles (les destinations) ainsi que d'une valeur quantitative pour chaque paire origine/destination. À partir du graphe G, crée lors de l'étape précédente, nous calculons le chemin le plus court entre l'origine et chaque destination. Nous avons implémenté l'algorithme de Dijktra pour cela. Nous avons testé trois possibilités pour définir le poids w des arêtes :

- $w = 1$, il n'y a pas de pondération.
- $w = d$, où d est la distance entre les deux nœuds. Les longues arêtes sont ainsi pénalisées.
- $w = d + (p \times (180 - a))$, plus la valeur a de l'angle (en degrés) avec l'arête précédente est petite, plus le poids w est élevé. Les petits angles sont donc pénalisés.

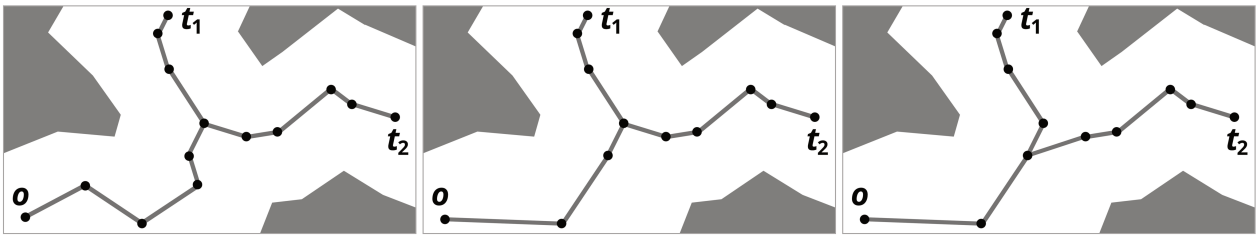
Les résultats présentés sont basés sur la dernière définition, avec $p = 5$. Nous discuterons plus en détails du choix de ce paramètre dans la Section 3.5.

Nous créons ensuite un graphe $T = (V, E)$ à partir de l'ensemble des plus courts chemins (cf. Figure 3.16c). Le graphe T est connecté, acyclique et orienté vers les destinations ; il s'agit donc d'un arbre orienté dans lequel la racine représente l'origine des flux et les feuilles représentent les destinations, ou point d'arrivée, des flux.

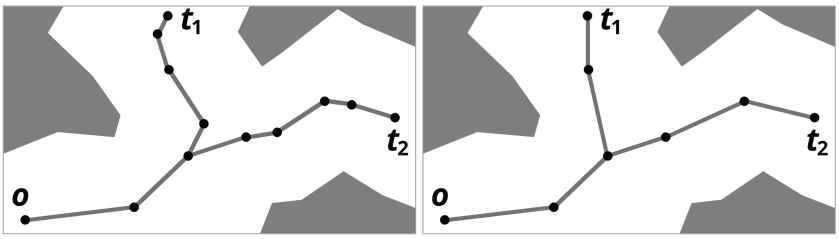
3.4.3 Simplification de l'arbre

Lorsque nous visualisons l'arbre mis en place durant l'étape précédente, nous pouvons observer de nombreuses courbes inutiles le long des chemins (cf. Figure 3.17a). Cette étape consiste à améliorer l'apparence des flux en lissant le tracé des chemins. Il existe une très grande variété de solutions pour la simplification. L'essentiel est de ne pas créer de croisements entre les chemins et les entités géographiques. Nous proposons d'effectuer une série d'améliorations successives présentées dans la Figure 3.17. Nous décrirons en détail chacune des étapes permettant d'obtenir une meilleure apparence des flux (voir résultat final sur la Figure 3.17e).

Tout d'abord, nous supprimons les nœuds qui forment des détours, pour passer de la Figure 3.17a à la Figure 3.17b. Considérons quatre nœuds successifs v_0, v_1, v_2, v_3 et trois arêtes (v_0, v_1), (v_1, v_2) et (v_2, v_3). Tous ces éléments sont illustrés dans la Figure 3.18a. Nous commençons par calculer les coordonnées polaires des nœuds v_1, v_2 et v_3 dans l'intervalle $[0, 2\pi[$. Le centre du cercle pour définir les coordonnées polaires correspond au nœud précédent. Par exemple, pour v_2 nous utiliserons v_1. Les coordonnées polaires sont notées (r_1, θ_1), (r_2, θ_2) et (r_3, θ_3). Nous considérons 4 intervalles pour les valeurs de θ : $[0, \frac{\pi}{2}[$, $[\frac{\pi}{2}, \pi[$, $[\pi, \frac{3\pi}{2}[$ et $[\frac{3\pi}{2}, 2\pi[$. Par exemple, si θ_1 et θ_3 sont dans le même intervalle et θ_2 dans un autre, nous supprimons le nœud v_1 (ce qui nous donne l'arête (v_0, v_2)). C'est le cas dans la Figure 3.18b : θ_1 et θ_3 sont dans l'intervalle $[0, \frac{3\pi}{2}[$ tandis que θ_2 est dans l'intervalle $[\frac{3\pi}{2}, 2\pi[$. Ainsi, v_1 est supprimé afin de ne plus avoir de détour (cf. Figure 3.18c).

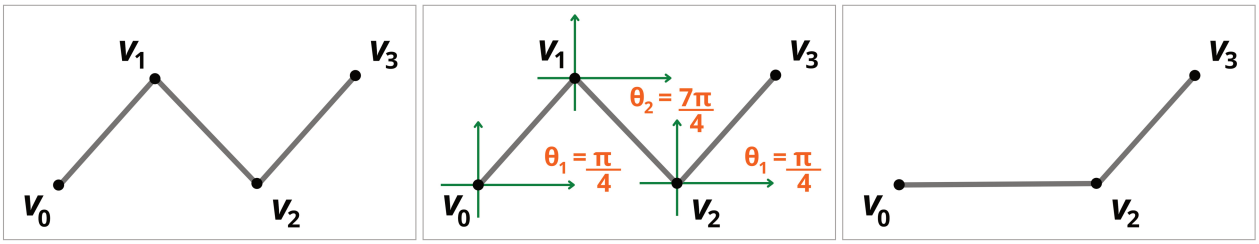


(a) Arbre initial (b) Suppression des détours (c) Smooth splitting



(d) Repositionnement des nœuds (e) Suppression des nœuds qui ne sont pas significatifs

FIGURE 3.17 – Étapes de la simplification. o est la racine de l'arbre (le point d'origne des flux). t_1 et t_2 sont les feuilles de l'arbre (points d'arrivée des flux).



(a) nœuds et arêtes initiaux. On a 4 nœuds : v_0, v_1, v_2, v_3, et 3 arêtes : (v_0, v_1), (v_1, v_2) et (v_2, v_3).

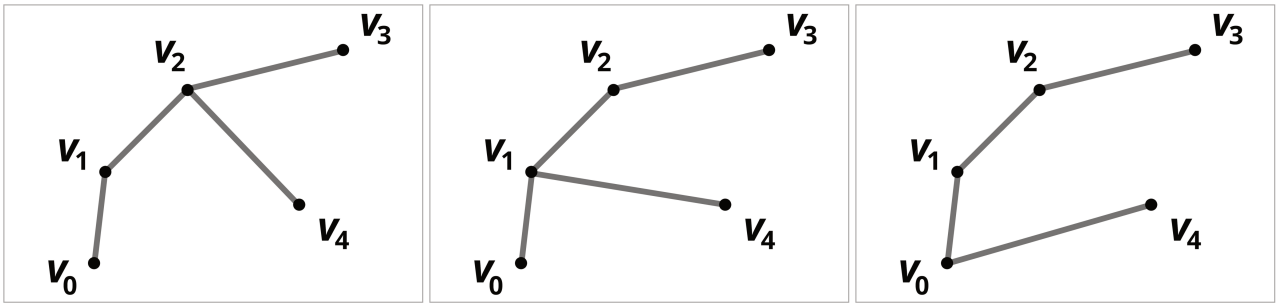
(b) Définition des coordonnées polaires des nœuds v_1, v_2 et v_3. On a $\theta_1 \in [0, \frac{\pi}{2}[$, $\theta_2 \in [\frac{3\pi}{2}, 2\pi[$ and $\theta_3 \in [0, \frac{\pi}{2}[$.

(c) Le nœud v_1 est supprimé car θ_1 et θ_3 sont dans le même intervalle et θ_2 dans un autre.

FIGURE 3.18 – Méthode pour la suppression des détours.

La deuxième étape pour améliorer le tracé des chemins consiste à diminuer les angles lorsqu'un nœud a un degré supérieur à 1 c'est-à-dire si on a une nouvelle séquence (cf. Section 3.3.1 et Figure 3.13 où on donne la définition de ce que l'on nomme séquence). L'idée principale est donc la même que celle de l'étape précédente étant donné que l'on veut supprimer les détours ; sauf que cette fois-ci il s'agit de redéfinir les nœuds de début et fin des séquences. Si l'angle formé par l'avant dernier nœud d'une séquence avec son dernier nœud, qui est aussi le premier nœud de ses séquences enfants, et le premier nœud d'une nouvelle séquence est inférieur à $\frac{\pi}{2}$ il faut alors corriger le début de la séquence (comparer la Figure 3.17b avec la Figure 3.17c). Considérons 5 nœuds v_0, v_1, v_2, v_3, v_4 et 4 arêtes (v_0, v_1), (v_1, v_2), (v_2, v_3) et (v_2, v_4). On a donc une séquence formée par les nœuds v_0, v_1 et v_2. Cette séquence a deux séquences enfants : une première avec v_2 et v_3 et une seconde formée par v_2 et v_4 (cf. Figure 3.19a). Si la succession des nœuds v_1, v_2, v_4 forme un angle inférieur à $\frac{\pi}{2}$ (c'est le cas dans la Figure 3.19a), nous supprimons l'arête (v_2, v_4) et nous ajoutons l'arête

(v_1, v_4) (cf. Figure 3.19b). Cela revient à modifier le début de la séquence enfant : elle débute alors au nœud v_1 et se poursuit avec le nœud v_4. Ensuite, l'angle formé par v_0, v_1, and v_4 est inférieur à $\frac{\pi}{2}$, alors on répète la procédure en supprimant (v_1, v_4) et en ajoutant une arête entre v_0 et v_4 (cf. Figure 3.19c). Cette démarche est répétée tant qu'il existe un nœud parent et jusqu'à ce que l'angle soit supérieur à $\frac{\pi}{2}$.



(a) nœuds et arêtes initiaux : les nœuds v_1, v_2, v_4 forment un angle inférieur à $\frac{\pi}{2}$.

(b) L'arête (v_2, v_4) est supprimée et remplacée par (v_1, v_4). Les nœuds v_0, v_1, v_4 forment un angle inférieur à $\frac{\pi}{2}$.

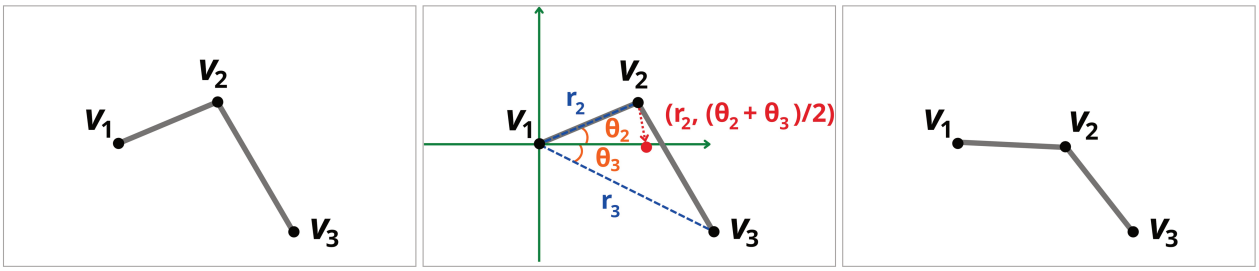
(c) L'arête (v_1, v_4) est supprimée et remplacée par (v_0, v_4).

FIGURE 3.19 – Amélioration des séquences.

Pour éliminer les effets de "demi-tour" sur les chemins, on repositionne les nœuds de degrés 2 pour lequel l'angle formé est inférieur à un seuil (pour passer de la Figure 3.17c à la Figure 3.17d). Considérons trois nœuds successifs v_1, v_2, et v_3 ainsi que deux arêtes (v_1, v_2) et (v_2, v_3). Si la ligne polygonale formée par v_1, v_2, v_3 a un angle inférieur à $\frac{\pi}{2}$ (cf. Figure 3.20a), on repositionne le nœud v_2. Pour cela, on calcul les coordonnées polaires de v_2 et v_3 en considérant v_1 comme le centre du cercle. Les coordonnée polaires calculées sont dans l'interval $]-\pi, \pi]$, elles sont respectivement notées (r_2, θ_2) et (r_3, θ_3) pour v_2 et v_3. Ensuite, on modifie les coordonnées de v_2 telles que $(r_2, \frac{\theta_2+\theta_3}{2})$ (cf. Figure 3.20b). Ces nouvelles coordonnées polaires peuvent être converties en coordonnées cartésiennes pour modifier la position du nœud (cf. Figure 3.20c).

La dernière étape de simplification consiste à supprimer les nœuds de degrés 2 qui ne sont pas significatifs. Autrement dit, cela revient à supprimer les nœuds dont la suppression ne changera pas de façon significative le tracé général de la ligne polygonale (passage de la Figure 3.17d à la Figure 3.17e)[4]. Considérons trois nœuds successifs v_1, v_2, et v_3 et deux arêtes (v_1, v_2) et (v_2, v_3) (cf. Figure 3.21a). D_1 est la distance entre v_1 et v_2. D_2 est la distance entre v_2 et v_3. D_{min} est la distance entre v_2 et le nœud le plus proche d'une entité géographique. Notons D la valeur maximale entre D_1 et D_2 présentée à la Figure 3.21b. Si $D < D_{min}$, on supprime le nœud v_2 comme illustré dans la Figure 3.21c.

4. En utilisant la méthode FMD, nous recommandons très fortement de procéder à cette dernière étape de simplification après la première distorsion, juste avant de dessiner les courbes. En effet, les nœuds qui ne sont pas significatifs en terme d'apparence, de chemin à suivre, peuvent en revanche s'avérer être significatifs dans le cadre de cette première étape de distorsion. En résumé, si l'objectif est de crée uniquement une carte de flux un-à-plusieurs sans déformation du fond de carte, cette dernière étape de simplification doit être effectuée pour lisser les chemins avant de dessiner les courbes et ajouter les épaisseurs aux flux (comme dans la Figure 3.16d).

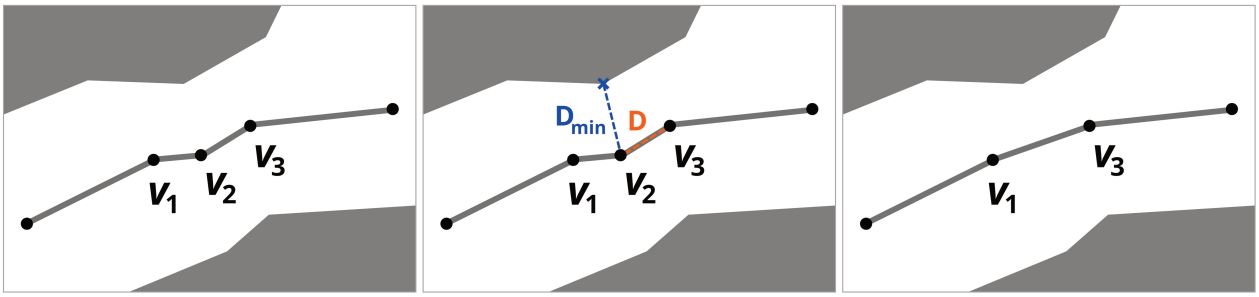


(a) Positions initiales des nœuds v_1, v_2, v_3. La ligne polygonale v_1, v_2, v_3 forme un angle inférieur à $\frac{\pi}{2}$: le nœud v_2 va être repositionné.

(b) Définition des coordonnées polaires de v_2 et v_3 en sa basant sur v_1 dans l'intervalle $]-\pi, \pi]$. Les coordonnées sont notées (r_2, θ_2) et (r_3, θ_3). La nouvelle position de v_2 est représentée en rouge.

(c) Nouvelle ligne polygonale formée par v_1, v_2 et v_3.

FIGURE 3.20 – Supprimer les angles en dessous d'un seuil.



(a) Chemin initial : trois nœuds successifs v_1, v_2 et v_3.

(b) La distance D, représentée en orange, est l'arête de longueur maximale entre (v_1, v_2) et (v_2, v_3). La distance D_{min}, en bleu, est la distance entre v_2 et le sommet le plus proche d'une entité géographique. Comme on a $D < D_{min}$, on supprime alors v_2.

(c) Chemin final sans le nœud v_2 considéré comme non significatif.

FIGURE 3.21 – Suppression des nœuds non significatifs.

3.4.4 Dessin des flux

Après l'étape de simplification, les flux peuvent être dessinés à partir de l'arbre. On dessine chaque séquences avec une courbe B-spline et on ajoute une épaisseur selon la somme totale des quantités des chemins qui y passent. Un exemple de résultat est présenté dans la Figure 3.16d.

Dans cette section, nous avons proposé une nouvelle méthode pour créer des cartes de flux un-à-plusieurs. Utiliser cette méthode n'est pas obligatoire pour appliquer la méthode de distorsion du fond de carte FMD mais très fortement recommandé. En effet, les chemins

longent les axes médians entre les entités géographiques. Cela permet notamment de minimiser le nombre et les surfaces de chevauchements et minimise donc le nombre d'itérations de FMD. De plus, cette méthode garantit de n'avoir aucun croisement entre un chemin et une entité géographique. Elle peut bien évidemment être utilisée pour créer une carte de flux un-à-plusieurs simple, sans distorsion.

3.5 Discussion

Dans cette section, nous examinons les résultats obtenus sur trois ensembles de données différents. Tout d'abord, nous présentons les caractéristiques principales des jeux de données et les temps de calcul. Ensuite, nous analysons les cartes créées par la méthode des cartes de flux un-à-plusieurs combinée à la méthode FMD, ainsi que l'impact du paramètre p. Enfin, nous présentons les commentaires d'un expert en cartographie sur les points forts et les limites de notre méthode.

3.5.1 Jeux de données

Le premier jeu de données concerne les exportations de thé de la Chine (CHN), le deuxième les exportations de soja des États-Unis (USA) et le troisième les exportations de vanille de Madagascar (MDG). Les données sont issues de FAOSTAT (dernière consultation : 11/2022). Pour chaque jeu de données d'entrée, le nombre de destinations (les pays dans ces exemples) est indiqué dans le Tableau 3.2. Ce Tableau indique également le minimum, le maximum, la moyenne, la médiane et l'écart type des valeurs associées aux flux.

Tableau 3.2 – Caractéristiques principales des jeux de données en entrée.

Jeu de données	Nb de dest.	Min	Max	Médiane	Moyenne	Écart-type
CHN	115	1	74260	280	3189.57	8134.21
USA	74	1	22575972	23183.5	707951.15	2717691.15
MDG	23	1	625	9	64.35	145.24

3.5.2 Temps de calcul

Les cartes de flux initiales pour chaque jeu de données ont été créées comme indiqué dans la Section 3.4. Pour cela, nous avions initialement 6996 sommets formant les polygones des entités géographiques. Le squelette créé à partir de ces polygones comportait 14002 sommets, 42006 arêtes et 6890 faces. Le graphe dérivé comptait 3697 nœuds et 3808 arêtes. La création du squelette et du graphe a pris environ 5 minutes (suppression des nœuds et des arêtes inutiles). Quel que soit le jeu de données, la création de la carte de flux initiale à partir du graphe et de la source de données a pris environ 2 minutes. Tous les traitements ont été effectués sur un ordinateur portable équipé d'un processeur quadricœur de 2,3 GHz et d'une mémoire de 32 Go. Le processus a été mis en œuvre avec des scripts Python, à l'exception de la création du squelette, qui a été réalisée à l'aide de la bibliothèque en C++ mentionnée dans la Section 3.4.

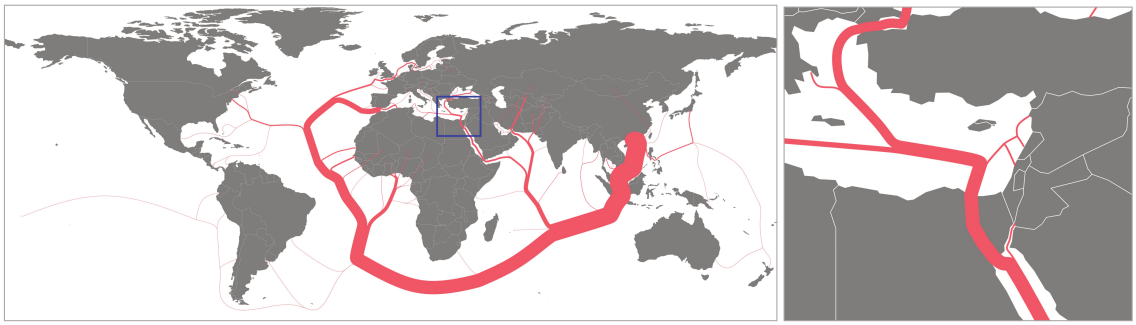
Nous avons ensuite exécuté la méthode FMD : les temps sont indiqués dans le Tableau 3.3.

Tableau 3.3 – Caractéristiques de l'arbre et temps de calcul de la méthode FMD.

Jeu de données	Nb de nœuds	FMD (temps)
CHN	1243	51 seconds
USA	1108	36 seconds
MDG	619	18 seconds

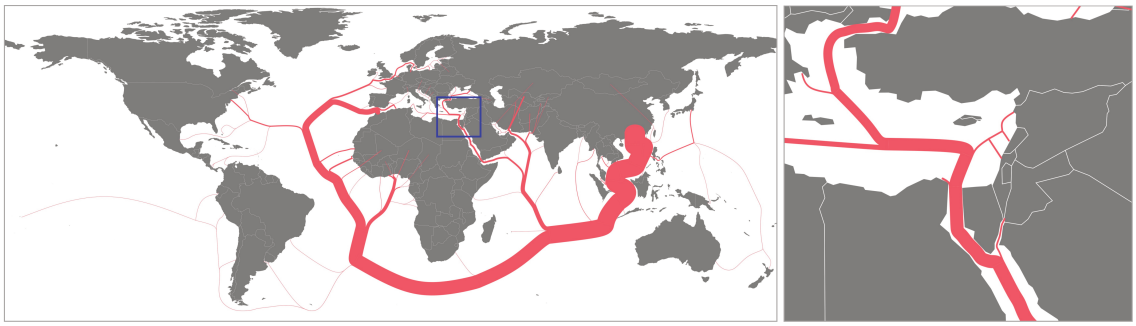
3.5.3 Résultats

Les résultats obtenus pour chacun des 3 ensembles de données avec et sans distorsion sont présentées dans Figure 3.22, Figure 3.23 et Figure 3.24 . Les résultats sans distorsion ont été obtenus par la méthode proposée dans la Section 3.4. Les résultats avec distorsion ont été obtenus en utilisant la méthode FMD de la Section 3.3 sur les résultats sans distorsion.



(a) Carte sans distorsion

(b) Zoom sur une zone de la carte sans distorsion

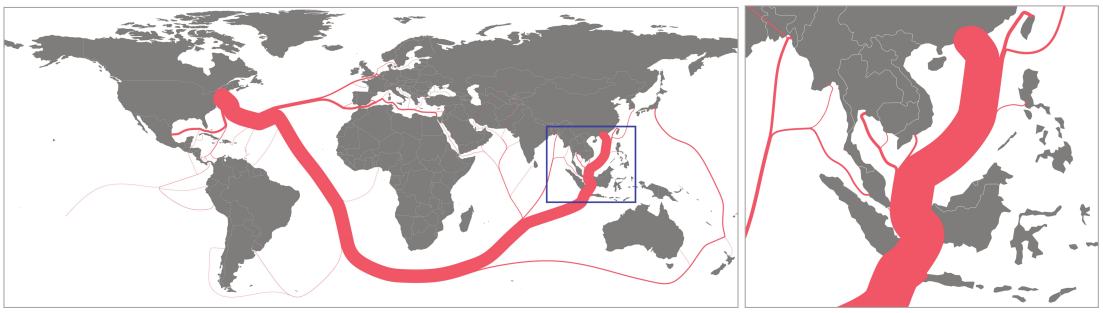


(c) Carte avec distorsion

(d) Zoom sur une zone de la carte avec distorsion

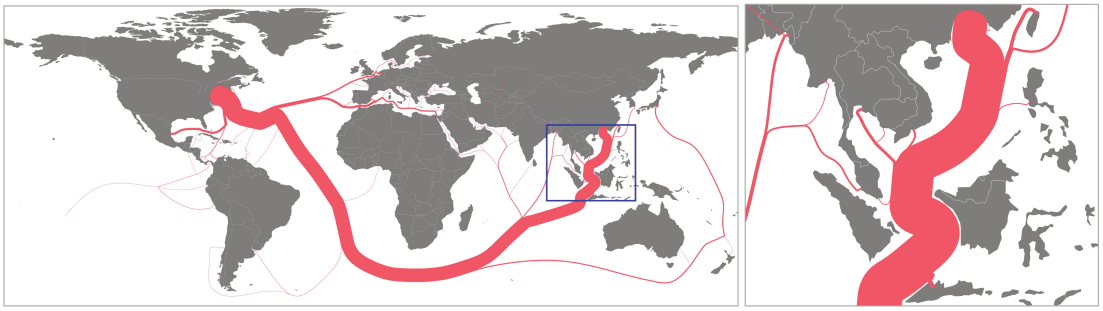
FIGURE 3.22 – Résultats avec le jeu de données CHN.

Dans certaines zones de la carte sans distorsion, un flux épais peut cacher de petits flux et des caractéristiques géographiques telles que des îles. Par exemple, la Figure 3.25 compare la zone méditerranéenne avec et sans distorsion. Sans distorsion, l'île de Crète, au sud de la Grèce, n'est pas du tout visible, tandis que la Sicile et la Sardaigne sont partiellement cachées. Avec la distorsion, les trois îles sont entièrement visibles et bien reconnaissables. De même, plusieurs petits flux à destination de divers pays d'Afrique du Nord, tels que la Tunisie et l'Algérie, sont à peine voire pas du tout visibles. Après la distorsion, on distingue beaucoup mieux ces flux.



(a) Carte sans distorsion

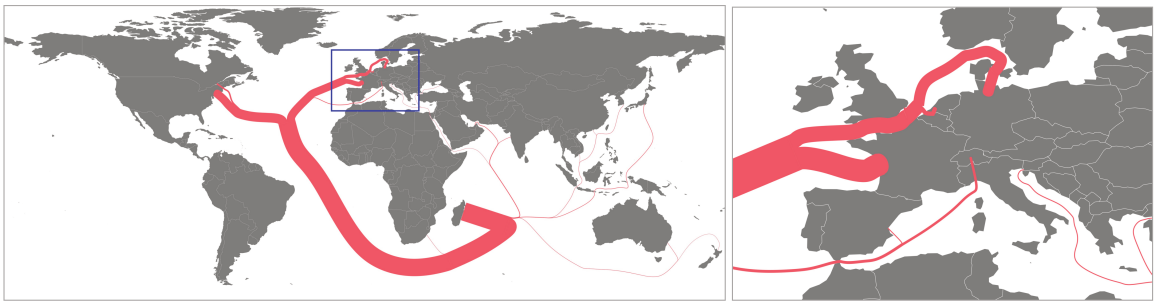
(b) Zoom sur une zone de la carte sans distorsion



(c) Carte avec distorsion

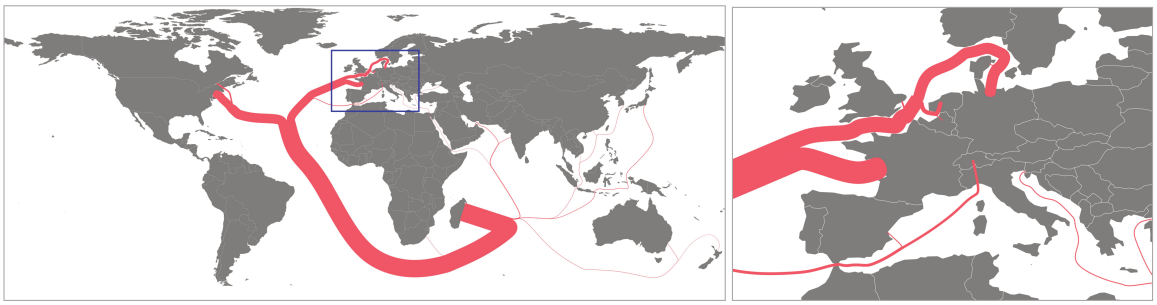
(d) Zoom sur une zone de la carte avec distorsion

FIGURE 3.23 – Résultats avec le jeu de données USA.



(a) Carte sans distorsion

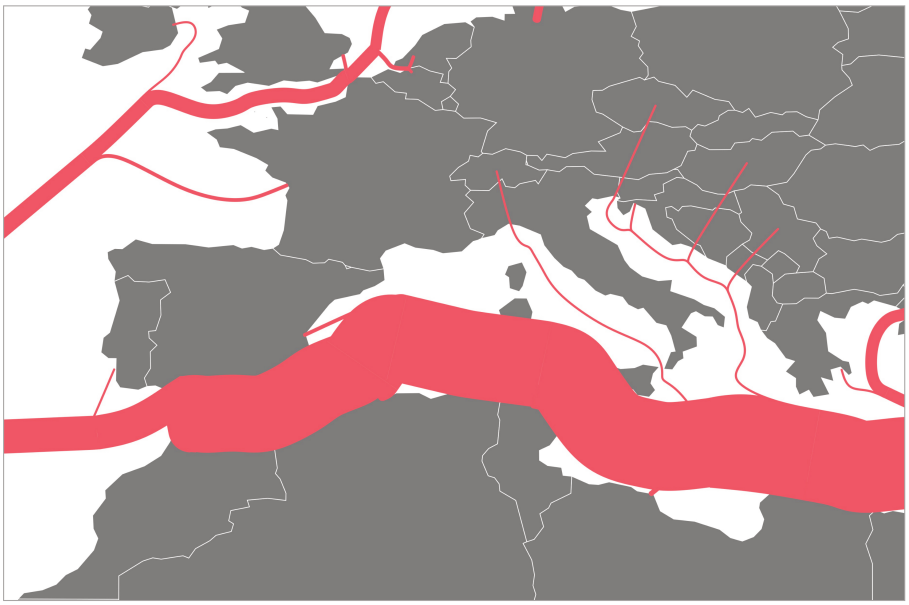
(b) Zoom sur une zone de la carte sans distorsion



(c) Carte avec distorsion

(d) Zoom sur une zone de la carte avec distorsion

FIGURE 3.24 – Résultats avec le jeu de données MDG.



(a) Sans distortion



(b) Avec distortion

FIGURE 3.25 – Example des bénéfices de la distorsion (cartes issues du jeu de données CHN avec $p = 0$ au lieu de $p = 5$ comme dans la Figure 3.22).

Le risque de ces distorsions est que les entités géographiques ne soient plus identifiables. Dans nos exemples, malgré les distorsions plus ou moins importantes, il est encore possible de reconnaître rapidement les entités géographiques. La Figure 3.26 illustre un exemple pour l'Indonésie qui compare les entités géographiques avec et sans distorsion : elles gardent leurs positions les unes par rapport aux autres, les déformations sont localisées et les entités restent identifiables. Nous avons réalisé une petite expérience avec 5 participants. Elle consistait à faire correspondre plusieurs pays étiquetés sur une carte avant déformation à la carte après déformation (sans le dessin du flux). Deux zones ont été étudiées : la mer Méditerranée (comme dans la Figure 3.25) et la zone de la Figure 3.26. Les participants n'avaient aucune connaissance spécifique en matière de cartographie et ont tous obtenu 100% de réponses correctes.



(a) Sans distortion (b) Avec distortion

FIGURE 3.26 – Exemple de distorsion (à partir du jeu de données USA comme dans la Figure 3.23).

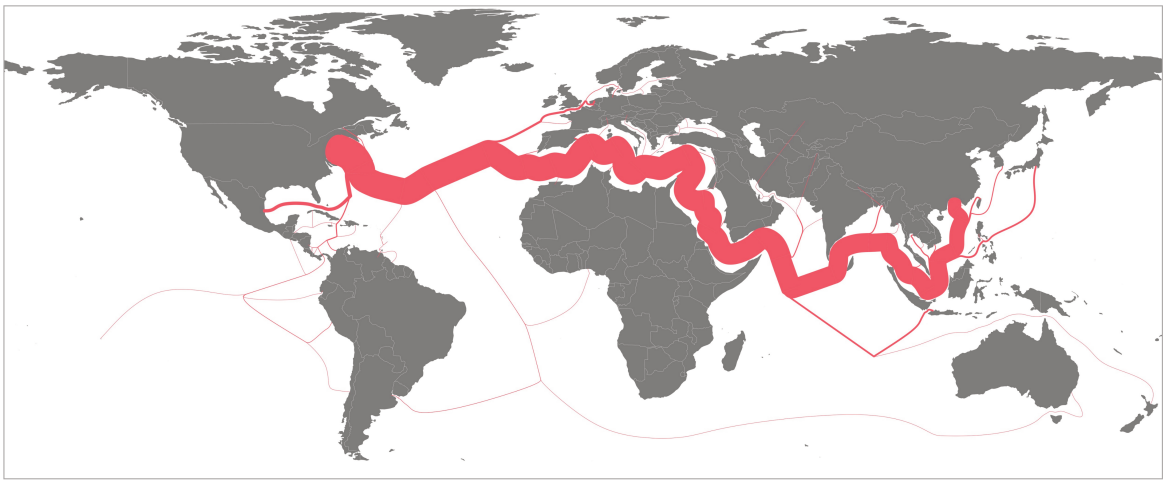
Pour obtenir toutes les cartes présentées précédemment, nous avons utilisé le paramètre $p = 5$ (sauf dans la Figure 3.25). Pour rappel, ce paramètre correspond à la pénalité que nous appliquons au poids des arêtes du graphe lors du calcul de l'arbre (cf. Section 3.4.2). Avec $p = 0.5$ et $p = 5$, les flux ne suivent pas nécessairement le chemin le plus court en termes de distances euclidiennes, mais privilégient les arêtes les plus longues puis les espaces les plus larges (cf. Figure 3.27). Cela limite les distorsions à effectuer et accélère le temps de calcul de la méthode FMD. Comme on peut le voir dans le Tableau 3.4, lorsque p augmente, le nombre de sommets relocalisés diminue et la distance totale, *i.e.* la somme des distances entre les positions initiales et finales des sommets relocalisés, est beaucoup plus petite. Quelle que soit la valeur de p, la distance moyenne parcourue par les sommets relocalisés reste la même. Par conséquent, le total des distances entre les sommets initiaux et finaux dépend principalement du nombre de sommets relocalisés et diminue donc lorsque nous avons une pénalité (avec $p > 0$). En conséquence, nous recommandons de définir l'arbre par un squelette droit tel que présenté dans la Section 3.4 et de pénaliser le poids des arêtes lors du calcul des plus courts chemins afin d'éviter les zones trop étroites (Section 3.4.2).

Tableau 3.4 – Modification des positions des sommets des polygones avec la méthode FMD.

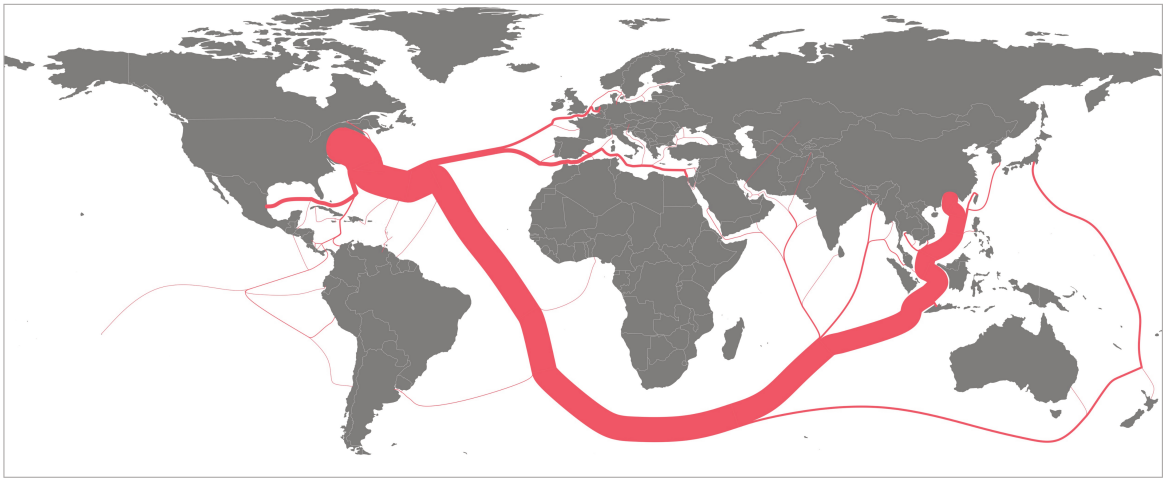
	Nb de sommets déplacés			**Distance totale**			**Distance moyenne**		
	p=0	**p=0.5**	**p=5**	**p=0**	**p=0.5**	**p=5**	**p=0**	**p=0.5**	**p=5**
CHN	2811	1242	1265	19093	8899	9157	6.79	7.16	7.24
USA	4576	219	219	38536	1453	1453	8.42	6.64	6.64
MDG	794	46	46	4425	280	280	5.57	6.06	6.06

Un expert en cartographie nous a fait un retour d'expérience, il n'a pas été impliqué dans le développement de notre méthode. Nous avons sollicité son expertise et lui avons fourni les sections précédentes ainsi qu'une application en ligne avec la possibilité d'exécuter notre méthode étape par étape. Nous lui avons ensuite demandé de rédiger un rapport sur les points forts et les limites de notre approche :

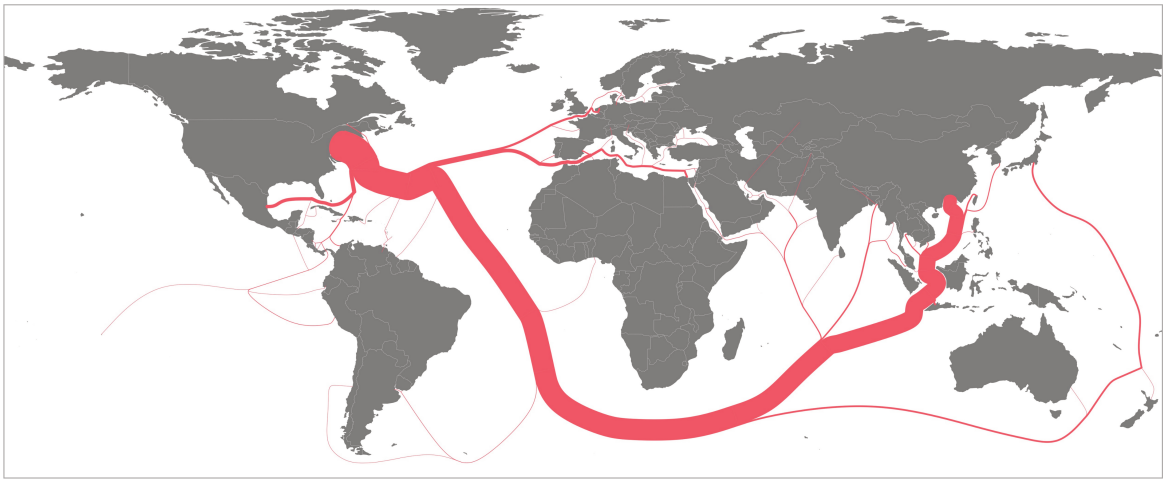
> La méthode FMD permet de créer automatiquement des cartes de flux "à la Minard" à partir d'une matrice origine-destination. La principale difficulté de la



(a) $p = 0$



(b) $p = 0.5$



(c) $p = 5$

FIGURE 3.27 – Cartes obtenues à partir avec différentes valeurs de p pour le jeu de données USA.

méthode réside dans le maintien d'un équilibre entre réalisme des déformations et visibilité des flux et des entités limitrophes. Les résultats illustrent l'efficacité et les limites de cette méthode, au niveau global et local.

Au niveau mondial, les cartes de flux présentées dans les Figures 3.22, 3.23 et 3.24 permettent de visualiser clairement les principales destinations à partir d'une origine, ainsi que les quantités associées à chaque flux. L'objectif principal de la carte de flux est donc atteint : le lecteur est en mesure (1) de hiérarchiser clairement les destinations correspondant aux flux d'une origine, et (2) d'associer une quantité à chaque flux. Graphiquement, l'utilisation des splines offre un confort en termes de lisibilité des grandes routes maritimes. Lors de la division d'un flux vers plusieurs destinations, l'orientation des courbes dans le sens du flux garantit une cohérence dans la lecture dynamique de la carte. En revanche, le réalisme général des routes proposées peut être discuté. En effet, la méthode de skeletisation construit des routes éloignées des côtes et des principales voies de navigation. La perception par le lecteur de l'organisation réelle des flux peut donc en être affectée. Néanmoins, le fait de maintenir à distance les flux des côtes présente un avantage, puisqu'il permet de limiter les conflits de chevauchement entre les flux et le fond de carte, et ainsi d'éviter les déformations géométriques qui en découlent.

Au niveau local, la méthode proposée permet d'éviter les chevauchements entre les flux importants (matérialisés par une largeur imposante) et les entités géographiques situées à proximité des côtes (qui constituent des repères importants pour le lecteur de la carte). L'exemple en mer Méditerranée dans la Figure 3.25 illustre clairement l'intérêt de la méthode pour visualiser les îles masquées par les flux symbolisés. D'un autre côté, le réalisme de certaines déformations spécifiques (associées à la proximité de flux importants avec de petites entités géographiques) peut être discuté. Par exemple, des distorsions irréalistes sont générées sur les îles de Java et de Bornéo dans la Figure 3.23d. Dans ce cas, le lecteur de la carte peut être amené à perdre des points de repère importants dans le fond de carte, en raison de la modication excessive de la forme originale des entités géographiques.

Dans ce contexte, l'intégration de contraintes visant à préserver la forme initiale des entités géographiques pourrait constituer une perspective intéressante afin de maintenir un compromis entre visibilité et réalisme. En effet, de telles contraintes permettraient de générer des déformations qui minimisent l'impact visuel sur les entités géographiques de petites tailles, qui peuvent être considérées comme des points de repère importants pour le lecteur de la carte.

(Jean-François Girres)

3.6 Conclusion

Dans ce chapitre, nous avons proposé une méthode, nommée Flow Map Distorsion, pour supprimer les chevauchements entre les flux et les entités géographiques dans une carte de flux un-à-plusieurs. Nous avons également présenté une méthode pour générer des cartes de flux un-à-plusieurs avec peu de chevauchements, qui peut être utilisée avant d'exécuter FMD. À travers plusieurs exemples, la discussion montre que notre méthode révèle des informations cachées tout en maintenant les positions des entités géographiques les unes aux autres et leur forme globale afin de les reconnaître malgré les déformations. Les résultats répondent à l'objectif principal, qui était de créer automatiquement des cartes déformées inspirées des cartes dessinées à la main par Charles Joseph Minard au 19e siècle.

Finalement, dans ce chapitre dédié à la visualisation de données quantitatives entre les entités géographiques nous avons proposé une méthode de visualisation permettant de conserver les informations géographiques tout en évitant trop d'encombrement visuel.

CHAPITRE

4

CONCLUSION ET PERSPECTIVES

4.1 Synthèse des contributions

Ce mémoire a abordé à la fois la comparaison de données quantitatives associées à des entités géographiques et la visualisation de données quantitatives associées à des relations entre des entités géographiques. Les travaux ont été présentés sous différentes formes selon la problématique et les besoins : (1) conception d'une approche de visualisation et développement d'une plateforme, (2) étude utilisateur, (3) création et implémentation d'un nouvel algorithme.

Conception d'une approche de visualisation

Dans le cadre du projet MOOD, les utilisateurs finaux ont exprimé le besoin de visualiser et d'explorer des données épidémiques afin d'identifier des corrélations entre des variables. Nous avons, dans un premier temps, généralisé la problématique à la comparaison par juxtaposition et avons proposé une approche de comparaison. Cette dernière a été intégrée et étendue pour répondre aux besoins supplémentaires des utilisateurs, au sein de la plateforme de visualisation de données épidémiologiques *Epid Data Explorer*.

Étude utilisateur

De nombreuses techniques de cartographie bivariée sont possibles mais encore trop peu de travaux ont été effectués à ce sujet. Nous avons proposé une évaluation comparant des types de cartes bivariées grâce à une étude utilisateur. L'objectif était d'apporter de nouvelles pistes pour aider dans le choix d'une technique adaptée au grand public parmi la diversité des solutions possibles.

Création et implémentation d'un nouvel algorithme

Les cartes de flux, généralement utilisées pour représenter des déplacements entre entités géographiques, par exemple des marchandises ou des migrations, présentent des problèmes d'encombrement. Des solutions ont été proposées mais aucune ne permet d'éviter de superposer les flux aux entités géographiques. Une solution à ce problème est l'utilisation de cartes de flux dans lesquelles les formes géographiques sont modifiées de façon à ne pas se superposer avec les flux, comme celles réalisées par Minard. Ces cartes étant construites à la main, nous avons proposé un algorithme (FMD) pour automatiser la déformation du fond de carte afin qu'il s'adapte aux flux. Nous avons également proposé un algorithme de positionnement des flux initiaux qui réduit la quantité de superposition et donc qui peut être utilisé avant FMD pour réduire les déformations qu'il engendre.

4.2 Perspectives

Les travaux effectués durant cette thèse se positionnent sur deux axes majeurs. Le premier, exposé dans le Chapitre 2, concerne les conceptions de visualisations par juxtaposition et superposition pour comparer des données quantitatives associées à des entités géographiques. Le second, décrit dans le Chapitre 3, est consacré à la visualisation de données quantitatives associées à des relations entre des entités géographiques à l'aide de cartes de flux avec déformation du fond de carte.

Les perspectives se divisent donc naturellement en deux premières parties, une pour la conception d'approches de visualisation pour la comparaison, une pour les améliorations de notre approche de distorsion de fond de carte d'une carte de flux. Une troisième partie concerne la comparaison dans le cadre d'une visualisation de flux. En effet, nous avons proposé une solution pour visualiser et comparer des flux d'une même carte de flux représentant les valeurs d'une seule variable à une seule date mais nous n'avons pas traité la comparaison à différentes dates ni de différentes variables.

4.2.1 Conception de nouvelles approches de comparaison

Nous commencerons par présenter des améliorations possibles pour l'approche de comparaison par juxtaposition que nous avons proposée. Ensuite, nous évoquerons les perspectives concernant la comparaison par superposition. Nous discuterons également d'autres approches de visualisation dédiées à la comparaison de différentes variables. Enfin, nous montrerons qu'il pourrait également être possible de combiner plusieurs solutions entre elles.

Juxtaposition

L'approche de comparaison par juxtaposition peut être améliorée en ajoutant de nouvelles fonctionnalités telles que le filtrage des données. Nous avons utilisé uniquement la technique de cartographie choroplèthe mais ce système de synchronisation peut finalement s'appliquer à bien d'autres types de cartes et d'autres techniques de visualisation qui peuvent même sortir du cadre de la cartographie.

Superposition

D'autres techniques de cartographie bivariée pourraient également être étudiées et comparées avec celles de notre expérimentation. Nous nous sommes basés sur les plus accessibles, mais bien d'autres existent, comme, par exemple, les cartes où un diagramme en barres, qui est positionné au centre de chaque entité, représente les valeurs des variables à l'aide de la hauteur des barres, ou bien la technique pour visualiser deux variables sur un seul cartogramme proposée par Nusrat et al. [69]. Il est aussi tout à fait envisageable d'en imaginer de nouvelles.

De plus, comme on peut le voir dans l'article proposant une taxonomie des tâches pour les géovisualisations [78], il existe beaucoup de types de tâches liées à la spatialité que nous n'avons pas traitées. Ces tâches peuvent aussi être déclinées à des problématiques spécifiques. Par conséquent, de nouvelles études permettant d'adresser d'autres tâches générales ou des tâches plus spécifiques à des cas particuliers peuvent fournir plus d'indications sur l'efficacité des techniques. Enfin, nous nous sommes limités à la comparaison de données quantitatives séquentielles, mais des études portant sur des données qualitatives (ordinales ou nominales) ou des données quantitatives divergentes ou cycliques pourraient aussi être utiles.

Certaines techniques de cartographie permettent de représenter plusieurs variables et pas seulement deux. Nous nous sommes centrés sur la visualisation bivariée mais de futurs travaux pour la visualisation multivariée sont à prévoir.

Les visualisations par superposition que nous avons étudiées n'ont aucune interaction contrairement à notre approche de visualisation par juxtaposition. Concevoir une approche de visualisation par superposition interactive permettrait d'améliorer son efficacité et d'explorer de nouvelles pistes.

Codage explicite pour comparer

Nous avons visualisé les données telles qu'elles étaient à l'origine pour comparer deux variables, que ce soit via l'approche de juxtaposition ou de comparaison. On note cependant juste une exception avec EDE où il est possible, durant l'importation des données, de procéder à des agrégations. Une perspective pertinente serait de pré-traiter les données de sorte à faire un codage explicite, correspondant à la troisième catégorie de comparaison (avec la juxtaposition et la superposition) déterminée par Gleicher [37]. Les possibilités sont assez nombreuses. Si, par exemple, on souhaite comparer pour une même variable sa valeur à deux dates différentes on pourrait très bien procéder à une soustraction pour calculer la différence des valeurs entre les deux dates. Cette différence serait directement représentée sur la carte et permettrait de voir comment les valeurs ont évolué. Il n'est alors pas nécessaire de comparer deux cartes juxtaposées ou d'opter pour de la cartographie bivariée.

Combinaisons et comparaison des approches

Dans les travaux futurs il serait également intéressant de combiner les deux approches de comparaison étudiées dans ce mémoire. Deux cartes bivariées pourraient être visualisées à partir de l'approche de juxtaposition que nous avons proposé. Ainsi, une paire de variables pourraient être comparées à une autre paire de variables par exemple. La relation entre deux variables pourrait également être explorée plus en détails en comparant deux dates différentes.

Nous avons comparé les techniques de cartographie bivariées mais celles-ci n'ont pas été comparées à l'approche de juxtaposition. Une perspective très utile est de mettre en

place une expérimentation qui cette fois-ci consisterait à comparer les trois approches : la juxtaposition, la superposition et le codage explicite.

4.2.2 Visualisation de flux par distorsion de fond de carte

L'algorithme de distorsion de fond de carte d'une carte de flux que nous avons proposé nécessite d'être optimisé. L'accélération du processus est suggéré en premier lieu, suivi de l'application de la méthode FMD à d'autres échelles géographiques. Ensuite, la perspective de l'étendre aux cartes de flux plusieurs-à-plusieurs (*many-to-many*) est présentée.

Améliorer la méthode FMD

Les temps de calcul de la méthode FMD peuvent beaucoup s'élever si la quantité de données est conséquente. Les travaux futurs doivent dans un premier temps se consacrer à l'accélération du processus avant d'envisager tout autre amélioration. Durant l'étape de la première distorsion, nous traitons chaque sommet des polygones constituant les entités géographiques et chaque point des séquences formant les flux, et calculons les distances pour l'ensemble des paires. Il faudrait procéder différemment pour la détection des chevauchements étant donnée que sa complexité en temps est conséquente.

Il en est de même pour la seconde distorsion où l'on calcule la distance entre chaque sommet des polygones des entités géographiques et chaque point le long de la trajectoire des flux.

De manière générale la détection de chevauchements, à elle seule, nécessite du temps et pourrait être significativement accélérée en déterminant une autre solution.

Adapter la méthode à d'autres échelles

Adapter FMD à d'autres échelles de cartes est une perspective tout à fait envisageable. Par exemple, on pourrait se concentrer à l'échelle d'une ville. Dans ce cas, les entités géographiques pourraient être les bâtiments et les zones vides par lesquelles les flux doivent passer pourraient être les rues et les routes. Notre méthode pourrait être utile pour mieux visualiser le trafic routier par exemple.

Un autre exemple d'échelle serait celle d'un pays où les flux pourraient longer les cours d'eau. Pour supprimer les chevauchements entre les entités géographiques et les flux il faudrait appliquer la même procédure qu'à l'échelle mondiale avec les espaces maritimes, en agrandissant les zones des cours d'eau.

Étendre l'approche de distorsion aux cartes de flux plusieurs-à-plusieurs

Les travaux futurs pourraient également porter sur le traitement de cartes de flux plusieurs-à-plusieurs. Cela signifie qu'il n'y aurait pas un unique point de départ. Cette perspective constituerait certainement l'amélioration la plus intéressante étant donné que beaucoup de jeux de données contiennent des relations entre différentes entités. Un application directe de notre algorithme entraînerait certainement une forte augmentation du temps de calcul puisque le nombre de séquences serait bien plus élevé. De plus, plusieurs questions se poseraient, notamment pour la définition de l'arbre : les séquences doivent-elles être regroupées ou traitées indépendamment en fonction du point de départ ? Comment gérer le chevauchement des flux entre eux ?...

4.2.3 Comparaison des flux à partir d'une même carte de flux

Nous avons évoqué la comparaison selon trois dimensions (variable, date et zone géographique) dans le cadre de la visualisation de données quantitatives associées à des entités géographiques. Il serait également pertinent d'appliquer cette idée dans le cadre de la visualisation de données entre des entités géographiques comme des cartes de flux. Plutôt que de juxtaposer des cartes de flux pour représenter une variable à différentes dates ou bien différentes variables, il est possible de tout avoir dans une même carte de flux. Deux possibilités se présentent alors : soit on positionne les flux ayant la même origine et la même destination les uns à côté des autres, soit on les imbrique. Dans le premier cas le principal avantage est au niveau de l'interprétation, on peut très facilement comparer les valeurs représentées par l'épaisseur des flux. Cependant, cette option implique d'importantes distorsions. Dans le second cas, le principal intérêt est de limiter les distorsions. En revanche, il est cette fois-ci plus complexe d'interpréter les valeurs correctement et, dans certains cas, s'il y a peu de différence entre deux flux, on risque de ne même plus les distinguer. Face à ces avantages et inconvénients, une étude comparant ces deux solutions, après les avoir implémentées, est à envisager.

BIBLIOGRAPHIE

[1] Natalia Adrienko and Gennady Adrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2) :205–219, 2010.

[2] Oswin Aichholzer and Franz Aurenhammer. Straight skeletons for general polygonal figures in the plane. In Cai Jin-Yi and Wong Chak Kuen, editors, *Proceedings of the Second Annual International Computing and Combinatorics Conference (COCOON)*, pages 117–126. Springer, 1996.

[3] Oswin Aichholzer, Franz Aurenhammer, David Alberts, and Bernd Gärtner. A novel type of skeleton for polygons. In *Journal of Universal Computer Science*, pages 752–761. Springer, 1995.

[4] Basak Alper, Selçuk Sümengen, and Selim Balcisoy. Dynamic visualization of geographic networks using surface deformations with constraints. In *Proc. of the Computer Graphics International Conference (CGI), Computer Graphics Society, Petrópolis, Brazil*, pages 7–9, 2007.

[5] Benjamin Bach, Emmanuel Pietriga, Ilaria Liccardi, and Gennady Legostaev. Ontotrix : a hybrid visualization for populated ontologies. In *Proceedings of the 20th international conference companion on World wide web*, pages 177–180, 2011.

[6] Françoise Bahoken. Représenter la mondialisation par des flux, le rôle de la distance cartographique perçue. *Mappemonde. Revue trimestrielle sur l'image géographique et les formes du territoire*, 2022.

[7] Jacques Bertin. *Sémiologie graphique, Les diagrammes, Les réseaux, Les cartes.* Paris : Gauthiers-Villars, 1967.

[8] Quirijn W Bouts, Tim Dwyer, Jason Dykes, Bettina Speckmann, Sarah Goodwin, Nathalie Henry Riche, Sheelagh Carpendale, and Ariel Liebman. Visual encoding of dissimilarity data via topology-preserving map deformation. *IEEE Transactions on Visualization and Computer Graphics*, 22(9) :2200–2213, 2015.

[9] Cynthia Brewer and Andrew J Campbell. Beyond graduated circles : Varied point symbols for representing quantitative data on maps. *Cartographic Perspectives*, 29 :6–25, 1998.

[10] Cynthia A Brewer, Alan M MacEachren, Linda W Pickle, and Douglas Herrmann. Mapping mortality : Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3) :411–438, 1997.

[11] Cynthia A Brewer and Linda Pickle. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4) :662–681, 2002.

[12] Kevin Buchin, Bettina Speckmann, and Kevin Verbeek. Flow map layout via spiral trees. *IEEE Transactions on Visualization and Computer Graphics*, 17(12) :2536–2544, 2011.

[13] Sergio Cabello, Herman Haverkort, Marc Van Kreveld, and Bettina Speckmann. Algorithmic aspects of proportional symbol maps. *Algorithmica*, 58 :543–565, 2010.

[14] Thomas Candela, Matthieu Péroche, Arnaud Sallaberry, Nancy Rodriguez, Christian Lavergne, and Frédéric Leone. Visualising post-disaster damage on maps : a user study. *International Journal of Geographical Information Science*, 36(7) :1364–1393, 2022.

[15] Sheelagh Carpendale, John Light, and Eric Pattison. Achieving higher magnification in context. In Feiner Steven and Landay James A., editors, *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 71–80. Association for Computing Machinery, 2004.

[16] Filip Claes, Dmitry Kuznetsov, Robin Liechti, Sophie Von Dobschuetz, Bao Dinh Truong, Anne Gleizes, Daniele Conversa, Alessandro Colonna, Ettore Demaio, Sabina Ramazzotto, et al. The empres-i genetic module : a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *Database*, 2014, 2014.

[17] William S Cleveland and Robert McGill. Graphical perception : Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387) :531–554, 1984.

[18] Jeremy Crampton. Gis and geographic governance : reconstructing the choropleth map. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 39(1) :41–53, 2004.

[19] Weiwei Cui, Hong Zhou, Huamin Qu, Pak Chung Wong, and Xiaoming Li. Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6) :1277–1284, 2008.

[20] Alberto Debiasi, Bruno Simões, and Raffaele De Amicis. Force directed flow map layout. In Laramee Robert S., Kerren Andreas, and Braz José, editors, *Proceedings of the 5th International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 170–177. SciTePress, 2014.

[21] Borden D Dent. Communication aspects of value-by-area cartograms. *The American Cartographer*, 2(2) :154–168, 1975.

[22] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(6) :533–534, 2020.

[23] James A Dougenik, Nicholas R Chrisman, and Duane R Niemeyer. An algorithm to construct continuous area cartograms. *The Professional Geographer*, 37(1) :75–81, 1985.

[24] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11) :e113114, 2014.

[25] Martin E Elmer. Symbol considerations for bivariate thematic maps. In *Proceedings of 26th International Cartographic Conference*, 2013.

[26] David Eppstein and Jeff Erickson. Raising roofs, crashing cycles, and playing pool : Applications of a data structure for finding pairwise interactions. *Discrete & Computational Geometry*, 22(4) :569–592, 1999.

[27] Ozan Ersoy, Christophe Hurter, Fernando Paulovich, Gabriel Cantareiro, and Alex Telea. Skeleton-based edge bundling for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12) :2364–2373, 2011.

[28] Esri. Histoire des SIG. https://www.esrifrance.fr/produits/SIG-historique.aspx. Dernière consultation le 12/09/2023.

[29] Petr Felkel and Stepan Obdrzalek. Straight skeleton implementation. In László Szirmay Kalos, editor, *Proceedings of the 14th Spring Conference on Computer Graphics (SCCG)*, pages 210–218. Comenius University, 1998.

[30] Kelvin LT Fung, Simon T Perrault, and Michael T Gastner. Effectiveness of area-to-value legends and grid lines in contiguous area cartograms. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[31] Michael Ganslmeier, Davide Furceri, and Jonathan D Ostry. The impact of weather on covid-19 pandemic. *Scientific reports*, 11(1) :22027, 2021.

[32] Emden R. Gansner, Yifan Hu, Stephen North, and Carlos Scheidegger. Multilevel agglomerative edge bundling for visualizing large graphs. In Giuseppe Di Battista, Jean-Daniel Fekete, and Huamin Qu, editors, *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis)*, pages 187–194. IEEE Computer Society, 2011.

[33] Peichao Gao, Zhilin Li, and Zhe Qin. Usability of value-by-alpha maps compared to area cartograms and proportional symbol maps. *Journal of Spatial Science*, 64(2) :239–255, 2019.

[34] Michael T. Gastner and M. E. J. Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences (PNAS)*, 101(20) :7499–7504, 2004.

[35] Michael T Gastner, Vivien Seguy, and Pratyush More. Fast flow-based algorithm for creating density-equalizing map projections. *Proceedings of the National Academy of Sciences*, 115(10) :E2156–E2164, 2018.

[36] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE symposium on information visualization*, pages 17–24. Ieee, 2004.

[37] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4) :289–309, 2011.

[38] Diansheng Guo and Xi Zhu. Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12) :2043–2052, 2014.

[39] Abel Guy J. and Sander Nikola. Quantifying global international migration flows. *Science*, 343(6178) :1520–1522, 2014.

[40] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain : real-time tracking of pathogen evolution. *Bioinformatics*, 34(23) :4121–4123, 2018.

[41] Lars E. Harrie, Tiina Sarjakoski, and Lassi Lehto. A variable-scale map for small-display cartography. *Proceedings of the Symposium on GeoSpatial Theory, Processing, and Applications*, 34(4) :8–12, 2002.

[42] Mark Harrower and Cynthia A Brewer. Colorbrewer.org : an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1) :27–37, 2003.

[43] Jan-Henrik Haunert and Leon Sering. Drawing road networks with focus regions. *IEEE Transactions on Visualization and Computer Graphics*, 17(12) :2555–2562, 2011.

[44] Stefan Hennemann. Information-rich visualisation of dense geographical networks. *Journal of maps*, 9(1) :68–75, 2013.

[45] Danny Holten. Hierarchical edge bundles : Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5) :741–748, 2006.

[46] Danny Holten and Jarke J. Van Wijk. Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, 28(3) :983–990, 2009.

[47] Bernhard Jenny, Daniel M. Stephen, Ian Muehlenhaus, Brooke E. Marston, Ritesh Sharma, Eugene Zhang, and Helen Jenny. Design principles for origin-destination flow maps. *Cartography and Geographic Information Science*, 45(1) :62–75, 2016.

[48] Bernhard Jenny, Daniel M. Stephen, Ian Muehlenhaus, Brooke E. Marston, Ritesh Sharma, Eugene Zhang, and Helen Jenny. Force-directed layout of origin-destination flow maps. *International Journal of Geographical Information Science*, 31(8) :1521–1540, 2017.

[49] Thienne M Johnson, Carlos Acedo, Stephen G Kobourov, and Sabrina Nusrat. Analyzing the evolution of the internet. In *EuroVis (Short Papers)*, pages 43–47, 2015.

[50] Naftali Kadmon and Eli Shlomi. A polyfocal projection for statistical surfaces. *The Cartographic Journal*, 15(1) :36–41, 1978.

[51] S Kaspar, Sara I Fabrikant, and P Freckmann. Empirical study of cartograms. In *Proceedings of the 25th International Cartographic Conference*. International Cartographic Association, 2011.

[52] Brett Kelly. Review of unclassed choropleth mapping. *Cartographic Perspectives*, 86 :30–35, 2017.

[53] Antoine Lambert, Romain Bourqui, and David Auber. 3d edge bundling for geographical data visualization. In *2010 14th international conference information visualisation*, pages 329–335. IEEE, 2010.

[54] Antoine Lambert, Romain Bourqui, and David Auber. Winding roads : Routing edges into bundles. In *Computer graphics forum*, volume 29, pages 853–862. Wiley Online Library, 2010.

[55] Nicolas Lambert and Christine Zanin. *Manuel de cartographie : principes, méthodes, applications*. Armand Colin, 2016.

[56] Anne Le Fur. *Pratiques de la cartographie*. Armand Colin, 2000.

[57] Anna Leonowicz. Two-variable choropleth maps as a useful tool for visualization of geographical relationship. *Geografija*, 42 :33–37, 2006.

[58] Rui Li. Visualizing covid-19 information for public : Designs, effectiveness, and preference of thematic maps. *Human Behavior and Emerging Technologies*, 3(1) :97–106, 2021.

[59] Yang Liu and Jeffrey Heer. Somewhere over the rainbow : An empirical assessment of quantitative colormaps. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[60] María-Jesús Lobo, Emmanuel Pietriga, and Caroline Appert. An evaluation of interactive map comparison techniques. In *Proceedings of the 33rd ACM conference on human factors in computing systems*, pages 3573–3582, 2015.

[61] Poulami Majumder and Partha Pratim Ray. A systematic review and meta-analysis on correlation of weather with covid-19. *Scientific Reports*, 11(1) :1–10, 2021.

[62] Hannah McClymont and Wenbiao Hu. Weather variability and covid-19 transmission : a review of recent research. *International journal of environmental research and public health*, 18(2) :396, 2021.

[63] Liam McNabb, Robert S. Laramee, and Max Wilson. When size matters : Towards evaluating perceivability of choropleths. In Gary K. L. Tam and Franck Vidal, editors, *Computer Graphics and Visual Computing (CGVC)*. The Eurographics Association, 2018.

[64] Tamara Munzner. *Visualization Analysis and Design*. A.K. Peters visualization series. A K Peters, 2014.

[65] Elisabeth Nelson. Using selective attention theory to design bivariate point symbols. *Cartographic Perspectives*, 32 :6–28, 1999.

[66] Elisabeth Nelson. The impact of bivariate symbol design on task performance in a map setting. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 37(4) :61–78, 2000.

[67] Sabrina Nusrat, Jawaherul Alam, and Stephen Kobourov. Recognition and recall of geographic data in cartograms. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9, 2020.

[68] Sabrina Nusrat, Md Jawaherul Alam, and Stephen Kobourov. Evaluating cartogram effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 24(2) :1077–1090, 2016.

[69] Sabrina Nusrat, Muhammad Jawaherul Alam, Carlos Scheidegger, and Stephen Kobourov. Cartogram visualization for bivariate geo-statistical data. *IEEE Transactions on Visualization and Computer Graphics*, 24(10) :2675–2688, 2017.

[70] Sabrina Nusrat and Stephen Kobourov. The state of the art in cartograms. *Computer Graphics Forum*, 35(3) :619–642, 2016.

[71] Sabrina Nusrat and Stephen Kobourov. The state of the art in cartograms. *Computer Graphics Forum*, 35(3) :619–642, 2016.

[72] Gilles Palsky. *Des chiffres et des cartes : naissance et développement de la cartographie quantitative française au XIXe siècle*. Comité des travaux historiques et scientifiques Paris, 1996.

[73] Vanessa Peña-Araya, Emmanuel Pietriga, and Anastasia Bezerianos. A comparison of visualizations for identifying correlation over space and time. *IEEE Transactions on Visualization and Computer Graphics*, 26(1) :375–385, 2019.

[74] Doantam Phan, Ling Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pages 219–224, 2005.

[75] Sandra Rendgen. *Le système Minard : Anthologie des représentations statistiques de Charles-Joseph Minard*. Princeton Architectural Press, 2018.

[76] Kent A Rittschof, William A Stock, Raymond W Kulhavy, Michael P Verdi, and Janet T Johnson. Learning from cartograms : The effects of region familiarity. *Journal of Geography*, 95(2) :50–58, 1996.

[77] Hans Rosling and Zhongxing Zhang. Health advocacy with gapminder animated statistics. *Journal of epidemiology and global health*, 1(1) :11, 2011.

[78] Robert E Roth. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12) :2356–2365, 2013.

[79] Robert E Roth, Andrew W Woodruff, and Zachary F Johnson. Value-by-alpha maps : An alternative technique to the cartogram. *The Cartographic Journal*, 47(2) :130–140, 2010.

[80] Bahador Saket, Arjun Srinivasan, Eric D Ragan, and Alex Endert. Evaluating interactive graphical encodings for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(3) :1316–1330, 2017.

[81] Manojit Sarkar and Marc H. Brown. Graphical fisheye views of graphs. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, page 83–91, 1992.

[82] Sarah Schöttler, Yalong Yang, Hanspeter Pfister, and Benjamin Bach. Visualizing and interacting with geospatial networks : A survey and design space. *Computer Graphics Forum*, 40(6) :5–33, 2021.

[83] Terry Slocum, Robert Sluter, Fritz Kessler, and Stephen Yoder. A qualitative evaluation of maptime, a program for exploring spatiotemporal point data. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 39(3) :43–68, 2004.

[84] Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. *Thematic cartography and geovisualization*. CRC Press, 2022.

[85] Georgianna Strode, John Derek Morgan, Benjamin Thornton, Victor Mesev, Evan Rau, Sean Shortes, and Nathan Johnson. Operationalizing trumbo's principles of bivariate choropleth map design. *Cartographic Perspectives*, 94 :5–24, 2019.

[86] Hui Sun and Zhilin Li. Effectiveness of cartogram for the representation of spatial data. *The Cartographic Journal*, 47(1) :12–21, 2010.

[87] Shipeng Sun. A spatial one-to-many flow layout algorithm using triangulation, approximate steiner trees, and path smoothing. *Cartography and Geographic Information Science*, 46(3) :243–259, 2019.

[88] Naohisa Takahashi. *An Elastic Map System with Cognitive Map-based Operations*, pages 73–87. Springer Berlin Heidelberg, 2008.

[89] Martijn Tennekes and Min Chen. Design space of origin-destination data visualization. In *Computer Graphics Forum*, volume 40, pages 323–334. Wiley Online Library, 2021.

[90] Waldo R Tobler. Choropleth maps without class intervals. *Geographical analysis*, 5(3) :262–265, 1973.

[91] Waldo R. Tobler. Experiments in migration mapping by computer. *The American Cartographer*, 14(2) :155–163, 1987.

[92] Waldo R. Tobler. Thirty five years of computer cartograms. *Annals of the Association of American Geographers*, 94(1) :58–73, 2004.

[93] Jeffrey S Torguson. Choropleth map. *International Encyclopedia of Geography : People, the Earth, Environment and Technology : People, the Earth, Environment and Technology*, pages 1–9, 2016.

[94] Bruce E Trumbo. A theory for coloring bivariate statistical maps. *The American Statistician*, 35(4) :220–226, 1981.

[95] Laëtitia Viau, Arnaud Sallaberry, Nancy Rodriguez, Jean-François Girres, and Pascal Poncelet. Polygon vector map distortion for increasing the readability of one-to-many flow maps. *International Journal of Geographical Information Science*, 37(6) :1288–1314, 2023.

[96] Colin Ware. *Information visualization : perception for design*. Morgan Kaufmann, 2019.

[97] Colin Ware, Terece L Turton, Francesca Samsel, Roxana Bujack, David H Rogers, K Lawonn, N Smit, and D Cunningham. Evaluating the perceptual uniformity of color sequences for feature discrimination. In *EuroRV^3@ EuroVis*, pages 7–11, 2017.

[98] Jo Wood, Jason Dykes, and Aidan Slingsby. Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, 47(2) :117–129, 2010.

[99] Ningchuan Xiao and Yongwan Chun. Visualizing migration flows using kriskograms. *Cartography and Geographic Information Science*, 36(2) :183–191, 2009.

[100] Daisuke Yamamoto, Shotaro Ozeki, and Naohisa Takahashi. Focus+glue+context : An improved fisheye approach for web map services. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*, page 101–110. Association for Computing Machinery, 2009.

[101] Yalong Yang, Tim Dwyer, Sarah Goodwin, and Kim Marriott. Many-to-many geographically-embedded flow visualisation : An evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 23(1) :411–420, 2016.

[102] Yalong Yang, Tim Dwyer, Bernhard Jenny, Kim Marriott, Maxime Cordeil, and Haohui Chen. Origin-destination flow maps in immersive environments. *IEEE Transactions on Visualization and Computer Graphics*, 25(1) :693–703, 2018.

[103] Cem Yuksel, Scott Schaefer, and John Keyser. On the parameterization of catmull-rom curves. In *Proceedings of the SIAM/ACM Joint Conference on Geometric and Physical Modeling (GD/SPM)*, page 47–53, 2009.

[104] Meng-Jia Zhang, Kang Zhang, Jie Li, and Yi-Na Li. Visual exploration of 3d geospatial networks in a virtual reality environment. *The Computer Journal*, 61(3) :447–458, 2018.

[105] Xi Zhu and Diansheng Guo. Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS*, 18(3) :421–435, 2014.