Université de Tours

École doctorale MIPTIS

Laboratoire d'Informatique Fondamentale et Appliquée de Tours
LIFAT (EA 6300)

Fouille de séquences de mobilité sémantique

Sur l'élaboration de mesures pour la comparaison, l'analyse et la découverte de comportements

Soutenue et présentée par **M. Clément Moreau** le 25 novembre 2021

pour obtenir le grade de Docteur de l'Université de Tours mention : Informatique

Jury

Pr. Christophe Claramunt	Institut de rech. de l'École Navale	Rapporteur
Pr. Jérôme Gensel	Université Grenoble Alpes	Rapporteur
Dr. Cyril de Runz	Université de Tours	Examinateur
Pr. Anne Laurent	Université de Montpellier	Présidente
Dr. Marie-Jeanne Lesot	Université Paris Sorbonne	Examinatrice
Pr. Thomas Devogele	Université de Tours	Directeur
Dr. Laurent Etienne	ISEN Brest	Co-encadrant
Dr. Verónika Peralta	Université de Tours	Co-encadrante

Remerciements

À bien des égards, la tâche de rédaction du présent manuscrit fut une sinécure comparée à celle des *Remerciements* tant l'envergure du soutien reçu et des rencontres faites est importante. Par ces quelques lignes j'espère réussir à rendre hommage à l'ensemble des personnes, expériences, et souvenirs marquants qui auront laissé (et laisseront) une trace au-delà même de cette période que fut celle de la thèse.

Pour débuter, je remercie tous les lecteurs et lectrices s'étend penché sur mes travaux de thèse. Merci de l'intérêt porté à mon travail et, en ce sens, je remercie les membres du jury pour leurs conseils, remarques et les riches échanges ayant conclu la soutenance de thèse et visant l'amélioration de la qualité du présent manuscrit.

Ce travail n'aurait jamais pu voir le jour sans l'ensemble des membres du département département informatique (DI) de Blois, un lieu où j'ai eu la chance de grandir intellectuellement, découvrir de nombreuses choses, cultiver ma curiosité, passer de si bons moments et trouver une bienveillance, une chaleur humaine rare. Merci à tous de m'avoir compter en tant qu'étudiant, collègue et ami.

Merci à Béatrice et Arnaud et m'avoir permis d'enseigner des matières qui me sont chères, me plaisent et m'amusent. Merci à eux pour leurs enseignements, leur confiance et leurs conseils.

Un merci particulier aux membres institutionnels du bureau Blobfish, Patrick, Nicolas et Alex de m'avoir accepté parmi eux, de m'avoir offert cafés, conseils et bien plus encore. Un remerciement tout particulier à Alex pour son aide, sa disponibilité et son enthousiasme toujours indéfectible. Bon courage pour ta thèse cher ami et collègue.

Dans cette lignée je remercie tous mes amis et collègues doctorants : Alex donc, mais aussi Faten, Faudil, Ben, Adam, Raymond et tous les autres qui débutent. Je sais à quel point ce parcours et long et parfois semé d'embûches. Courage.

Un merci plus personnel aux amis TAListes. Merci à Jean-Yves pour son savoir, son expérience et ses conseils, à Théo pour tous nos échanges riches où l'on avait tendance à refaire le monde et à Maëlle. Je suis heureux de te voir aujourd'hui épanouie. Qui aurait cru que nous en serions là 8 ans auparavant lors de notre arrivée en L1...

Enfin, je n'oublie pas Fred. Ah ! Un grand merci à toi d'avoir été présent pendant ces années de thèse. Nous avons tant partagé : nos encadrants, notre vision de la recherche, nos difficultés, nos joies, peines et questionnements. Notre voyage en

Bavière à PKDD 2019 restera parmi mes meilleurs souvenirs de thèse. :) Merci à toi cher ami et confrère Dr. Bisone.

Finalement, je remercie mes encadrants sans qui tout ce travail ça n'aurait jamais été possible. Plus que des collègues, ce fut pour moi une seconde famille. Tout d'abord, merci à Verònika pour son savoir et sa rigueur scientifique, mais aussi sa chaleur humaine, muy caliente. Merci à elle pour notre contribution majeur au magazine "4 saisons" ; maintenant je sais faire pousser les plus belles tomates de tout Paris. Merci aussi à Laurent sans qui je ne peux plus regarder une motte de beurre salé sans penser à lui. Merci d'avoir toujours été présent malgré la distance et d'avoir toujours su trouver les mots d'encouragement justes. Enfin très grand merci à Thomas sans qui rien n'aurait été possible. Merci de toute la confiance, la gentillesse dont tu as fait preuve durant ces années. Ton encadrement m'aura permis de grandir à la fois intellectuellement et humainement. Je suis heureux d'avoir eu la chance d'avoir un directeur de thèse aussi savant, sage mais surtout aussi sympa. Pour finir je n'oublie pas deux personnes un peu spéciales à mes yeux. Merci à Cyril d'avoir rejoint en cours de route l'encadrement de la thèse. Merci pour tous tes conseils, ton aide parfois précieuse et ta générosité. Et puis, merci à Evelyne sans qui je n'en serais aujourd'hui pas là. Merci à elle pour toutes ces années d'enseignement et de dévouement envers ses étudiants. Sans toi j'aurais perdu le goût pour les mathématiques à jamais. Je te dois beaucoup.

Il me reste pour finir à remercier tous mes amis qui sont trop nombreux pour que je les cite. Mais merci à Adrien, Jordan, les Cléments et toute la bande de cop'chiens lyonnais, vous vous reconnaîtrez, et je vous ai ;)

Et finalement, il ne me reste qu'à remercier ma famille. Merci à ma soeur,

Publications personnelles

Communications nationales

- C. Moreau, T. Devogele, L. Etienne *Extraction de motifs de trajectoires sémantiques similaires*. SAGEO (2018)
- S. Duroudier, S. Chardonnel, B. Mericskay, I. André-Poyaud, O. Bedel, S. Depeau, T. Devogele, L. Etienne, A. Lepetit, C. Moreau, N. Pelletier, E. Ployon, K. Tabaka *Données hétérogènes de mobilités quotidiennes : protocole de diagnostic qualité et d'apurement à partir de la base MOBI'KIDS*. SAGEO (2019)

Communications internationales

- C. Moreau, T. Devogele, V. Peralta et L. Etienne, *Contextual edit distance for semantic trajectories*. ACM SAC (2020)
- C. Moreau, V. Peralta, P. Marcel, A. Chanson et T. Devogele, *Learning analysis patterns using a contextual edit distance*. DOLAP @ EDBT/ICDT (2020)
- W. Verdeaux, C. Moreau, N. Labroche, P. Marcel, *Causality based explanations in multi-stakeholder recommendations*. ETMLP @ EDBT/ICDT (2020)
- C. Moreau, A. Chanson, V. Peralta, T. Devogele, C. de Runz, *Clustering sequences of multi-dimensional sets of semantic elements*. ACM SAC (2021)
- C. Moreau, V. Peralta *Learning Analysis Behavior in SQL Workloads*. DOLAP @ EDBT/ICDT (2021)
- C. Moreau, T. Devogele, C. de Runz, V. Peralta, E. Moreau, L. Etienne, *A Fuzzy Generalisation of the Hamming Distance for Temporal Sequences*. FUZZ-IEEE (2021)

Articles de journaux

- C. Moreau, T. Devogele, L. Etienne *Calcul de similarité sémantique entre trajectoires*. RIG (2018)
- C. Moreau, T. Devogele, L. Etienne, V. Peralta, C. de Runz *Methodology for Mining, Discovering and Analyzing Semantic Human Mobility Behaviors*, **submitted in** DMKD (Dec. 2020)
- C. Moreau, C. Legroux, V. Peralta, M. Ali Haroumi *Mining SQL Workloads for Learning Analysis Behavior*, (Extended version of *Learning Analysis Behavior in SQL Workloads*), Information Systems (2021)

Résumé

"Dites-moi ce que vous avez fait, je vous dirai qui vous êtes". Cet aphorisme, inspiré du livre Fondation de Isaac Asimov, interroge sur la prédictibilité et la compréhension actuelle de l'humain basée sur ses actions passées. Sommes-nous ce que nous faisons ? Cette question est devenue aujourd'hui un enjeu majeur pour de nombreux domaines comme le profilage d'individus ou les systèmes de recommandation qui cherchent, dans les actions passées des utilisateurs, un révélateur de leurs comportements futurs ou de leur psychologie.

Dans cette thèse, nous ancrons la précédente réflexion dans le cadre de la mobilité humaine et proposons la mise en place d'une méthodologie complète (i.e., *data pipeline*) pour l'analyse et la découverte de comportements depuis un ensemble de séquences de mobilité sémantique. Cette méthodologie se base sur un examen approfondi de la littérature concernant les propriétés de la mobilité humaine ; nonobstant, elle fournit un cadre générique pour l'étude de toute séquence à caractère sémantique. Un processus d'apprentissage non supervisé (i.e., *clustering*) est en charge de l'extraction des comportements et une phase d'explicabilité post-process est assurée afin de traduire les clusters en comportements intelligibles. En conséquence, nous avons retenu un ensemble d'indicateurs visuels et statistiques complémentaires venant renseigner les différents aspects des séquences tout en veillant à rester suffisamment concis afin d'éviter une surcharge cognitive. Cette explication est indispensable pour des raisons pratiques et éthiques, mais aussi pour inclure l'utilisateur dans le processus de découverte. Également, les séquences en jeu étant complexes de par leur caractère temporel et leur possible multi-dimensionnalité sémantique (lieux, activités, mode de déplacement, etc), nous proposons deux nouvelles mesures pour la comparaison de telles séquences nommées Contextual Edit Distance et Fuzzy Temporal Hamming distance. Celles-ci sont respectivement inspirées de la distance d'édition et de la distance de Hamming, et viennent alimenter le précédant processus de clustering. Ces nouvelles mesures s'appuient sur les ontologies et la logique floue afin de pallier les lacunes à la fois sémantiques, temporelles et structurelles des distances originelles. Ces apports ont été appliqués sur différents jeux de données réelles issus du domaine de la mobilité – physique (mobilité urbaine) et virtuelle (exploration de base de données) et ont permis d'améliorer significativement le processus d'interprétation et de découverte de comportements. Enfin, dans un but de ré-utilisabilité et de partage, une application web, SIMBA, vient parachever nos réalisations afin de permettre aux différents experts de s'approprier nos contributions au travers d'un outil interactif de fouille de données et analyse exploratoire.

Les travaux de cette thèse s'inscrivent en collaboration de deux projets ANR et régional : MOBI'KIDS qui vise à comprendre et caractériser les formes d'autonomie et conditions d'évolution des mobilités quotidiennes des jeunes enfants. Et SMARTLOIRE, dont l'objectif est d'offrir un ensemble d'outils numériques à destination des professionnels du tourisme et décideurs politiques pour la recommandation d'itinéraires et l'analyse de traces touristiques en région Centre-Val de Loire.

Mots-clés : Fouille de données, Séquence sémantique, Mobilité, Comportement humain, Mesure de similarité, Statistiques descriptives, Logique floue, IA explicable

Table des matières

Table des figures

Liste des tableaux

Chapitre 1
Introduction

C'est un lieu commun de dire que les machines en savent beaucoup sur nous. Elles nous connaissent parfois mieux que nos proches, amis ou familles [251] et sont capables d'anticiper avec une précision déroutante nos préférences, désirs et actions. Cette capacité, les machines l'ont acquise par le fait que, dans nos sociétés actuelles, nous essaimons volontiers (plus ou moins consciemment) des données de façon quotidienne et anecdotique sur ce que nous faisons, aimons, consultons, où nous allons, quand et qui nous fréquentons. Véritables "doudous" des temps modernes, nos smartphones produisent et recueillent en continu des données nous concernant sur tous les aspects précédemment cités. Ces données sont largement générées de manière passive par tous les appareils connectés qui renseignent en temps réel leur état et leurs interactions avec l'environnement physique via Internet. C'est cette interconnexion entre objets connectés et monde numérique que l'on nomme l'*Internet of Things* (IoT) et qui nous a propulsés aujourd'hui dans l'ère du *Big Data*.

Terme souvent galvaudé et fantasmé, nous définissons le Big Data, en accord avec Boyd et Crawford [33], comme un phénomène technologique et culturel qui met en coordination les trois axes suivants :

1. Technologique. Par la maximisation des ressources, de la puissance de calcul et la précision algorithmique pour rassembler, relier et comparer de grands ensembles hétérogènes de données.
2. Analytique. Par l'appui de vastes ensembles de données, l'identification de phénomènes statistiques et la création de modèles prédictifs pour la prise de décision. Nous détaillons plus loin cet axe étroitement lié au concept de l'*Intelligence Artificielle* (IA).
3. Mythologique. Par la croyance répandue selon laquelle les grands ensembles de données offrent une forme supérieure d'intelligence et sont capables de générer de la connaissance inattendue et de qualité.

Comme d'autres phénomènes socio-techniques, le Big Data, couplé à l'IA, déclenche auprès des foules une rhétorique duale tantôt de rêves technologiques tantôt de visions dystopiques et orwelliennes [1] qui masquent souvent les changements plus nuancés et subtils qui sont en cours.

1. Relatif à George Orwell et son oeuvre

Qu'elles soient synonyme de peur ou d'espoir, les performances liées au Big Data fascinent et interrogent. Pourtant, comme le notent Boyd et Crawford : "*Bigger data are not always better data*". Ainsi, la notion de Big Data ne peut, *ipso facto*, être dissociée de celle de *data science*, de l'analyse statistique ou de façon plus générale de l'IA des données. Si l'on reprend les trois axes du phénomène Big Data, on comprend que le premier axe forme la mise en place, le substrat du phénomène, que le troisième apporte la légitimation psychologique de cette mise en place, mais que toute la valeur créée est intégralement contenue dans l'axe analytique.

Aussi, plaçons-nous dans le contexte de l'étude de la mobilité. Supposons que, grâce à un ensemble de données IoT (e.g., smartphone), un sondage ou via l'exploitation de données issues des réseaux sociaux, nous soyons en capacité de récupérer un ensemble d'informations sur la mobilité et les activités des individus de la nature suivante :

"Alice part de la fac à pied, elle passe à la boulangerie acheter une gourmandise puis décide de rendre visite à une amie. Celles-ci se rendent en voiture en ville boire un café. Enfin, Alice rentre en bus jusqu'à chez elle."

"Bob part aussi de la fac, il prend le bus jusqu'au supermarché, puis rentre en bus chez lui. Il retourne ensuite en ville, à pied, pour travailler dans un café, il rentre à pied chez lui."

Pris isolément, ces énoncés de vie forment des séquences d'activités qui renseignent seulement les habitudes individuelles de chaque personne. Cependant, d'un point de vue collectif, ils forment une matière précieuse pour l'*analyse de la mobilité*, des comportements et des habitudes de vie. Par exemple, ces informations sont très utiles pour les urbanistes à des fins de développements d'infrastructures ou politiques de régulation, aux épidémiologues pour la modélisation de transmission de maladies ou aux sociologues pour l'étude des comportements et modes de vie urbain. Toutefois, on remarque que de nombreux traitements et analyses sont nécessaires pour extraire de ces données une information de qualité, exploitable et compréhensible quant aux différents besoins soulevés dans les précédents exemples. De fait, la tâche menée dans l'axe analytique afin de justifier les prouesses de l'IA (et du Big Data) est considérable.

C'est au niveau de cet axe que nous posons nos différentes contributions. Conscients de nos limites et en accord avec le fait que les données ne sont pas génériques et s'inscrivent dans un contexte précis, nous ancrons notre travail au sein du domaine de la fouille de séquences sémantiques, et plus particulièrement dans le contexte de la mobilité humaine en vue de l'extraction automatique de comportements interprétables. Ainsi, nos problématiques s'ancrent dans le domaine de la fouille de données et visent (i) à l'élaboration de mesures tenant compte des spécificités temporelles, sémantiques et structurelles inhérentes à la mobilité et aux actions humaines et qui, une fois couplées à un algorithme de clustering adéquat, sont capables de segmenter un ensemble de séquences sémantiques en comportements intelligibles. Également, nous proposons (ii) la mise en oeuvre d'un processus de découverte de connaissances

(*data pipeline*) collaboratif entre experts humains et machine et qui vient pallier les différents manques liés à la comparaison, l'analyse, l'interprétation et la transparence contextuelle pour l'étude et l'extraction de comportements à partir d'ensembles de séquences sémantiques.

Contexte scientifique et enjeux

La notion de *séquence sémantique* peut être rendue intuitive par la définition de chacun des termes pris indépendamment. Par *séquence* nous entendons toute suite d'événements chronologiquement ordonnés représentant un processus quelconque. Le terme *sémantique* quant à lui fait référence au sens de ces événements ; ces derniers conservent une représentation riche, haut niveau et clairement intelligible. En conséquence, les informations liées à cette sémantique forte sont généralement de nature qualitative et forment le pendant des données collectées par l'IoT qui sont généralement quantitatives. Une telle abstraction permet alors la représentation formelle de tout type de phénomène évoluant à travers le temps et formant un historique tel que la navigation sur des pages web, l'achat régulier d'objets (e.g., liste de courses), l'écoute de playlists ou encore la mobilité et les activités quotidiennes effectuées par un individu. C'est dans ce dernier cas d'utilisation que s'illustrent nos travaux en collaboration avec les projets SMARTLOIRE [2] et MOBI'KIDS [3] qui visent tous deux l'étude thématique de la mobilité individuelle. Cette thèse a été financée par la région Centre-Val de Loire via l'initiative SMARTLOIRE et l'Agence Nationale de la Recherche (ANR) via le projet MOBI'KIDS.

Le projet SMARTLOIRE coordonne différents experts informaticiens issus des domaines de la recommandation, de l'optimisation, de l'ingénierie des connaissances et de la fouille de données dans le but de concevoir un ensemble d'outils numériques de pointe dédié à l'analyse et la recommandation en temps réel de parcours touristiques en région Centre-Val de Loire. Dans le cadre de la fouille de données, un des enjeux pour les acteurs politiques et professionnels du tourisme est la compréhension des dynamiques territoriales et touristiques afin de saisir les intérêts et opportunités en fonction des profils socio-démographiques des individus, leurs préférences, la période de l'année, la météo, voire d'autres données contextuelles. Une meilleure compréhension de la mobilité des touristes et de leurs envies permettrait à la région le déploiement de stratégies pour une meilleure gestion du territoire au niveau touristique telles que le développement de lignes de transport, d'infrastructures urbaines ou d'événements culturels d'ampleur. Basée sur des séquences sémantiques de visites et d'activités, la définition de profils touristiques prototypiques offre la possibilité d'adapter les offres de recommandation selon les préférences détectées des utilisateurs et d'améliorer la fréquentation de certains sites en ciblant les individus les plus susceptibles d'y être intéressés. Une telle tâche réclame une modélisation des séquences sémantiques en adéquation avec la richesse descriptive potentielle des entités touristiques à notre dis-

2. `https://smartloire.univ-tours.fr`
3. `https://www.pacte-grenoble.fr/programmes/mobi-kids`

position. De fait, SMARTLOIRE s'appuie sur l'initiative DATAtourisme [4], une ontologie nationale pour la représentation des données touristiques et du patrimoine.

Parallèlement, le projet MOBI'KIDS vise à étudier *in situ* l'évolution de la mobilité quotidienne et de l'autonomie des enfants dans un milieu urbain et péri-urbain. L'hypothèse centrale étant que le degré d'autonomie des jeunes individus est corrélé à quelques facteurs contextuels socio-démographiques et familiaux (i.e., cultures éducatives) qu'il convient de découvrir. L'originalité de MOBI'KIDS est qu'il rassemble des experts issus des sciences formelles (informatique, statistiques) et humaines (psychologie, sociologie, géographie) et mêle des méthodes d'investigation et données mixtes entre ces disciplines. Des capteurs GPS sont mis à disposition des participants (parents et enfants) afin de collecter leurs trajectoires spatio-temporelles quotidiennes. Ces données GPS sont ensuite segmentées, apurées, et enrichies sémantiquement par un procédé de map-matching ; un ultime enrichissement et contrôle de la qualité sémantique est assuré lors d'une phase d'entretien individuel entre un membre du projet et les parents et enfants enquêtés. Un des défis majeurs de MOBI'KIDS est la capacité à comparer et définir des types de comportements au sein d'un ensemble de séquences sémantiques dans un environnement non supervisé, c'est-à-dire sans avoir de connaissance préalable sur la nature des différents profils comportementaux à découvrir. Ainsi, il est indispensable de proposer une méthodologie d'analyse explicative afin de permettre aux différents experts, techniques ou issus des SHS, de réfuter ou valider la pertinence des comportements découverts. L'intégration de données contextuelles telles que les variables socio-démographiques des individus doit être possible à l'issue du processus d'extraction des connaissances. Un tel cadre de travail s'inscrit dans une perspective pluridisciplinaire et *human in the loop* où l'expertise humaine et les découvertes liées aux méthodes d'intelligence artificielle viennent s'enrichir mutuellement offrant du même coup une transparence et une intelligibilité quant aux comportements extraits.

De ces deux projets naît un objectif commun d'*extraction et de découverte automatique de comportements depuis un ensemble de séquences de mobilité sémantique*. Pour ce faire, ne disposant pas de labels (i.e., profils comportementaux) *a priori*, nous nous plaçons dans un cadre d'apprentissage non supervisé qui nécessite de découvrir les différents schémas et distributions qui décrivent les données afin d'en faire ressortir les régularités intéressantes ou inattendues. Afin de définir des groupes de comportements similaires, il est nécessaire d'avoir recours à la notion de mesure de similarité (resp. dissimilarité) permettant la comparaison de nos séquences de mobilité sémantique. L'élaboration d'une telle mesure est un enjeu majeur car c'est elle qui conditionne en grande partie le résultat et l'intelligence du processus de segmentation en comportements. Le défi consiste alors à créer une ou des mesure.s capable.s d'agréger à la fois les dimensions sémantique et temporelle tout en s'accommodant des spécificités relatives à la mobilité et psychologie de l'habitude. De fait, il est démontré une inclination naturelle de l'humain à répéter des schémas de comportement

4. `https://framagit.org/datatourisme/ontology/`

et conserver une certaine cohérence sémantique, dans l'exécution de ses actions / activités, un même *habitus* selon les termes de Bourdieu [32], conformément à certains principes ou exigences de vie.

Enfin, par leur nature complexe, les séquences sémantiques supposent l'élaboration d'un cadre d'étude s'appuyant sur une pluralité d'indicateurs statistiques afin d'accomplir une analyse exhaustive permettant de révéler et de bien comprendre la nature, les lois statistiques et corrélations qui sous-tendent les données. Quant aux comportements extraits, ceux-ci sont expliqués en extrayant leurs singularités, traits distinctifs relativement à chaque cluster. Cet objectif d'honnêteté et de compréhension est majeur dans notre démarche et fonde le caractère éthique de notre approche de la fouille de données. Un ensemble de visualisations est en charge de la bonne intégration des connaissances par l'utilisateur qui doit venir ensuite valider ou réfuter la pertinence des profils découverts par nos algorithmes. La création d'un tel cadre de travail est un défi majeur car il combine de nombreuses problématiques relevant de la fouille de données, des statistiques, de la visualisation et de l'interaction homme-machine. En outre, le médium de transmission des connaissances doit trouver un juste équilibre entre simplicité et exhaustivité qui sont deux exigences ardûment conciliables. Toutefois, nous croyons que la définition d'un tel cadre explicatif, où un jugement analytique humain vient se confronter aux découvertes automatiques de la machine et en valider la justesse, permet une forme de validation de nos résultats. En outre, cette méthodologie descriptive permet également une transparence, une accessibilité et une compréhension des résultats de ces processus, parfois opaques, pour les utilisateurs non-experts en statistiques et sciences des données.

Démarche et réalisations

Dans un premier temps, il convient de définir un formalisme de représentation des séquences sémantiques capable de tenir compte de l'ensemble des informations mises à notre disposition et de respecter les principes de vie privée et de singularité de l'individu. Ainsi, la modélisation sous forme de séquences de labels sémantiques permet de concilier une richesse d'expression, une intelligibilité forte pour l'humain et une forme de respect de l'intimité de l'utilisateur, ces données étant moins sensibles que des positions GPS.

Toutefois, comme abordé précédemment, les données sémantiques sont complexes à manipuler et réclament souvent un éclairage métier contextuel supplémentaire. Une structuration des données sous la forme de graphe de connaissances est nécessaire afin de les rendre pleinement opérationnelles et utilisables. L'usage d'ontologies ou de taxonomies permet à la fois l'obtention d'une vision métier unifiée et la communication entre experts techniques et métiers, mais également de disposer d'une abstraction sur laquelle la machine peut effectuer des opérations de comparaisons entre entités ou des raisonnements. Ces opérations de comparaison sont subordon-

nées par l'usage de mesures de similarité qu'il convient d'étudier afin de choisir la plus pertinente pour notre usage.

Dans le cadre des séquences sémantiques, nous avons étudié les mesures couramment utilisées dans la littérature pour la comparaison de séquences et séries temporelles symboliques (i.e., qualitatives). Parallèlement, une étude minutieuse des propriétés intrinsèques de la mobilité et des habitudes humaines est menée au sein d'un corpus rassemblant sciences humaines et formelles afin de traduire ces propriétés en spécificités formelles. Au meilleur de nos connaissances, nous avons relevé de nombreux manques à la fois à la fois sémantiques, temporels et structurels des mesures de l'état de l'art quant à la vérification de ces spécificités qu'il nous convient de combler. Pour ce faire, nous proposons d'adapter plusieurs mesures existantes de l'état de l'art selon une approche basée sur la logique floue. Les mesures obtenues forment une extension floue de l'originale et permettent de remplir l'ensemble des spécificités requises pour l'étude des comportements liés spécifiquement à la mobilité mais généralisable à toute séquence d'actions humaines.

Pour finir, comme nous évoluons dans un cadre non-supervisé manipulant des objets complexes, nous proposons la mise en place d'un cadre méthodologique générique pour l'étude des séquences sémantiques. Conformément aux manques à la fois éthiques et analytiques relevés par rapport à l'existant, nous proposons la mise en place d'un ensemble d'indicateurs statistiques complémentaires pour le profiling et la description de jeux de données de séquences sémantiques. Un processus de clustering permet la segmentation des données et une phase d'explicabilité post-process est assurée par les précédents indicateurs afin de traduire les clusters découverts en comportements intelligibles. Nous avons retenu un ensemble d'indicateurs et de visualisations associées simples, capables de décrire avec efficacité et concision la structure, les caractéristiques et les singularités liées à un ensemble de séquences sémantiques tout en épargnant l'individu de la surcharge cognitive. Ce cadre méthodologique s'inscrit dans une démarche de compréhension et de collaboration entre experts humains et la machine dans le but d'accroître la réactivité et la communication dans la chaîne de découverte de connaissances et donc d'améliorer avec plus de facilité et de rapidité le choix des différents paramètres des algorithmes. Une application web opérationnelle nommée SIMBA (SemantIc Mobility Behavior Analysis) vient concrétiser nos différentes réalisations au sein d'une plate-forme dynamique dédiée à la fouille et à l'analyse interactive de séquences sémantiques et à l'extraction de comportements. SIMBA se présente comme un environnement permettant la ré-appropriation, le partage et la communication de nos travaux auprès d'un public d'experts thématiciens non-avertis aux techniques de l'intelligence artificielle et permet, au travers d'un environnement sobre et *user-friendly*, l'exploration, l'analyse et la découverte de comportements au sein d'un ensemble de séquences de mobilité sémantique.

La thèse se compose de 9 chapitres. Après cette introduction, une première partie d'État de l'art se divise en trois chapitres : Le chapitre 2 propose une réflexion sur le

temps et les différents modes de représentation des trajectoires et séquences sémantiques au sein de la littérature. Le chapitre 3 aborde la comparaison des séquences sémantiques à travers le prisme de la mobilité. En outre, ce chapitre traite la comparaison de concepts, des propriétés universelles liées à la mobilité et aux habitudes humaines et des mesures existantes pour la comparaison de séquences qualitatives. Le chapitre 4 est dédié à l'analyse, l'extraction et la découverte de connaissances au sein d'un ensemble de séquences sémantiques. Une première section aborde la définition comportementale de la mobilité orientée selon l'école de pensée Behavioriste et de la psychologie comportementale. Une vue pragmatique des processus d'extraction est ensuite présentée en confrontant les approches de fouille de motifs séquentiels et de clustering. Enfin, des axes pour l'analyse et l'explicabilité des comportements au sein des clusters sont présentés en dernière partie de ce chapitre. La seconde partie liée aux Contributions se scinde en quatre chapitres : Le chapitre 5 introduit les problématiques et spécificités liées à la comparaison des séquences de mobilité sémantique avant d'exposer une première mesure, la Contextual Edit Distance (CED), issue des travaux de cette thèse. Une expérimentation appliquée à l'étude et l'extraction de comportements d'exploration de bases de données est présentée. Le chapitre 6 vient améliorer les premiers aspects développés dans CED au sein d'une nouvelle mesure, Fuzzy Temporal Hamming distance (FTH), tenant compte des durées des activités ainsi que des spécificités liées à la mobilité selon une approche continue et floue du temps. Le chapitre 7 incorpore la gestion d'ensembles d'éléments sémantiques multidimensionnels au sein des séquences. Des exemples et un cas d'étude sont proposés à travers une tâche d'extraction de comportements touristiques prototypiques sur des séquences artificielles issues d'instances de l'ontologie DATAtourisme. Le chapitre 8 développe nos contributions méthodologiques en termes d'analyse et d'extraction de comportements depuis un ensemble de séquences sémantiques. Un cas d'étude issu des données réelles de l'Enquête Ménage-Déplacement (EMD) 2018 est exposé suivant notre méthodologie comportant une phase de pré-étude du jeu de données et une phase d'extraction de comportements intelligibles. L'application SIMBA est présentée dans la dernière section qui décrit son implémentation effective et les perspectives de développement. Le dernier chapitre résume et conclut les différents travaux, réalisations et apports de cette thèse et pointe les perspectives.

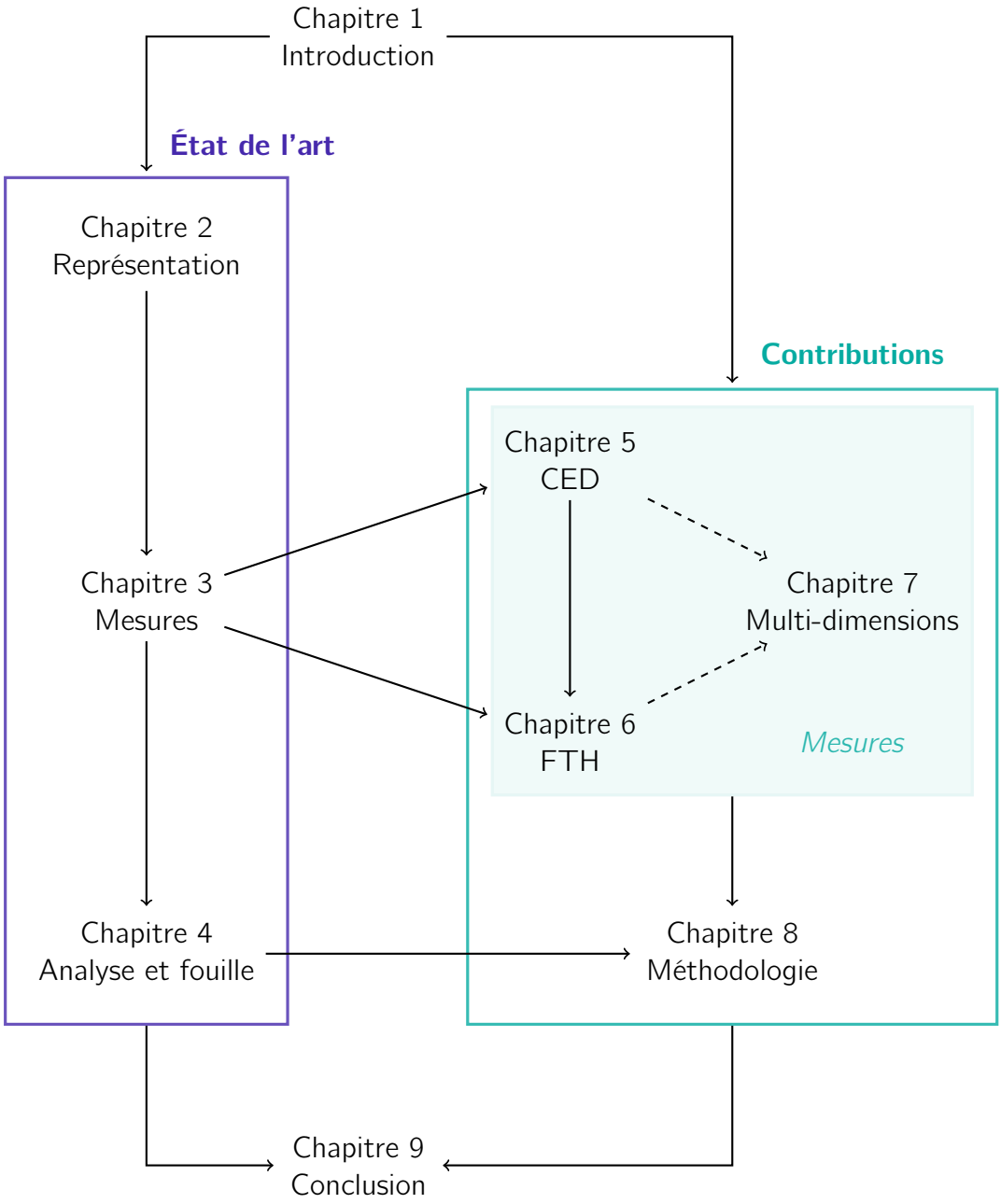


Figure 1.1 – Structure de la thèse sous forme de graphe. Les arcs représentent les liens d'influence entre les différents chapitres et parties

Première partie

État de l'art

Chapitre 2
Représentation des activités humaines dans le temps

2.1 Réflexion sur le temps

Abondamment étudié en philosophie et en physique, le *temps* est une notion fondamentale et structurante du monde qui peut-être ramenée au concept plus essentiel d'ordre dans lequel des évènements se produisent. Cette définition conventionnelle est généralement justifiée selon une vision relativiste du temps en physique [200, 68]. Le concept de l'ordre temporel des évènements imprègne ainsi notre réflexion sur la modélisation des systèmes et objets. Par exemple, nous disons que quelque chose s'est produit à 3h15 si cela s'est produit après que notre horloge ait lu 3h15 et avant qu'elle ne lise 3h16. De même, le repas du midi est situé entre le petit-déjeuner du matin et le souper du soir. De cette façon, on voit que le temps possède une direction particulière, un écoulement, une orientation d'où émerge la relation de causalité entre évènements [184].

En informatique, la notion d'ordre temporel a donc été représentée très tôt. Dans [127], Leslie Lamport propose un cadre de représentation pour les systèmes distribués basé sur la notion de *précédence*. On dit alors qu'un évènement a précède un évènement b si a s'est produit à un moment antérieur à b et on note $a \rightarrow b$ pour signifier cette relation de précédence temporelle. Notons ici que $\rightarrow$ est une relation asymétrique et transitive au sein d'un même système. Ainsi, on dit que les n évènements $e_{i \in \{1,\dots,n\}}$ d'un processus forment une *séquence* $\langle e_1, \dots, e_n \rangle$ si $\forall i, j \in [\![1, n]\!], i < j \Rightarrow e_i \rightarrow e_j$. En d'autres termes, une séquence est définie comme un ensemble d'évènements pourvu d'un ordre total.

De cette notion de précédence découle alors une dimension périssable de l'évènement ; celui-ci a une fin, il est de durée limitée. Ainsi, un second emploi du terme temps a trait cette fois-ci dans la *durée* de ces évènements. Naïvement, la durée d'un évènement e_i est définie comme l'intervalle mesuré par une horloge entre le début de e_i et sa fin [1].

1. Pour des raisons de concision et de clarté du discours, nous négligerons ici les aspects physiques et quantiques du temps. Nous renvoyons à [200] pour plus de précision sur ces sujets.

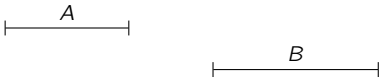
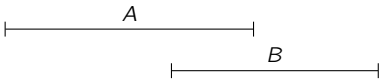
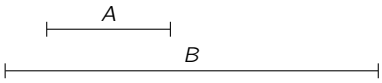
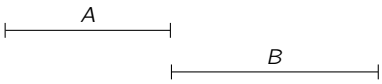
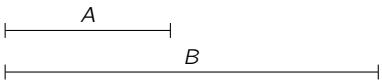
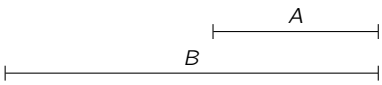
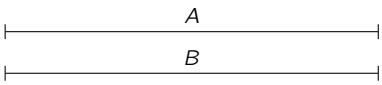
Nom			Définition	Image
before	$b(A,B)$	=	$a^+ < b^-$	
overlaps	$o(A,B)$	=	$a^- < b^- \wedge b^- < a^+ \wedge a^+ < b^+$	
during	$d(A,B)$	=	$b^- < a^- \wedge a^+ < b^+$	
meets	$m(A,B)$	=	$a^+ = b^-$	
starts	$s(A,B)$	=	$a^- = b^- \wedge a^+ < b^+$	
finishes	$f(A,B)$	=	$a^+ = b^+ \wedge b^- < a^-$	
equals	$a(A,B)$	=	$a^- = b^- \wedge a^+ = b^+$	
after	$bi(A,B)$	=	$b(B,A)$	
overlapped-by	$oi(A,B)$	=	$o(B,A)$	
contains	$di(A,B)$	=	$d(B,A)$	
met-by	$mi(A,B)$	=	$m(B,A)$	
started-by	$si(A,B)$	=	$s(B,A)$	
finished-by	$fi(A,B)$	=	$f(B,A)$	

Table 2.1 – Relations entre intervalles temporels de l'algèbre d'Allen. $A = [a^-, a^+]$ et $B = [b^-, b^+]$

Dès lors, le concept d'intervalle temporel introduit précédemment et qui sert à définir la notion de durée apparaît comme une base efficace permettant de représenter les raisonnements temporels. L'algèbre définie dans [10] par Allen propose un modèle logique permettant de formaliser et automatiser les raisonnements à caractère temporel. Le temps y est alors abstrait comme une demi-droite continue et isomorphe aux nombres réels positifs. L'algèbre d'Allen vise à considérer les exigences suivantes : (i) Permettre l'imprécision et les raisonnements en temps relatif, c'est-à-dire en usant des notions de précédence et non pas seulement de dates absolues. (ii) Modéliser les contraintes temporelles entre évènements (iii) Être indépendant d'une granularité temporelle donnée. (iv) Supporter la persistance, c'est-à-dire faciliter l'inférence d'évènements. Ainsi, Allen définit 13 types de relation résumés dans la Table 2.1.

Néanmoins, même si l'algèbre d'Allen fournit un cadre très efficace pour la modélisation et l'étude des évènements dans le temps, elle ne propose pas de traiter le cas d'intervalles temporels disjoints qui peut être utile pour représenter un aspect cyclique du temps et, par exemple, caractériser la récurrence de certains évènements (e.g., les week-ends). C'est pour combler cette lacune que Ladkin propose dans [126] une extension de la taxonomie précédemment dressée par Allen. Comme considérer une

énumération exhaustive des façons dont les intervalles peuvent être liés les uns autres est impossible car exponentielle, Ladkin propose une quantification qualitative des relations binaires d'Allen, indépendante du nombre de sous-intervalles, basées sur les qualificatifs : "mostly", "always", "partially" et "sometimes". Une relation de disjonction est également introduite pour permettre de combiner les relations d'Allen. Un dernier qualificatif "bars" est introduit pour indiquer si l'union des intervalles est connexe. Des exemples et une généralisation de l'algèbre de Ladkin pour la dimension spatiale est donnée par Claramunt dans [46].

Cette qualification qualitative donnée par Ladkin met en lumière une faiblesse dans le formalisme d'Allen et qui, pourtant, faisait partie des exigences de départ : *permettre l'imprécision*. En effet, les intervalles temporels définis par Allen sont stricts et s'étendent d'un instant (i.e., point temporel) à un autre. Or l'humain s'accommode également de perceptions vagues, floues et approximatives dans sa vision du temps. Par exemple, le terme *soirée* dénote un intervalle temporel imprécis que l'on pourra aisément situer entre 18h et 20h mais pas avec le même degré de conviction pour chaque instant. En effet, il nous semble absurde de considérer qu'à 20h nous sommes dans la soirée mais plus une fois les aiguilles indiquant 20h01.

Les connaissances dont nous disposons sur un évènement sont alors souvent imparfaites : soit parce que nous doutons sur leur validité (par exemple lorsque nous estimons un horaire sans disposer d'une horloge), soit parce qu'elles portent en elles une notion intrinsèquement imprécise (comme le terme de soirée). La première notion d'imperfection, celle qui concerne l'*incertain*, fut rapidement envisagée sous l'angle probabiliste dès le XVIIe siècle par Pascal et Fermat. La seconde notion, l'*imprécision*, fut considérée principalement à partir de 1965 lorsque Zadeh introduit le concept d'*ensemble flou* qui généralise la théorie des ensembles et logique classique en admettant des résultats intermédiaires entre le tout et le rien [252]. Le développement d'une telle notion permet alors de traiter l'idée d'appartenance partielle à une classe, de catégories aux limites mal définies et de gradualité dans le passage d'une situation à une autre. C'est dans cette optique que [207] propose une généralisation floue de l'algèbre d'Allen. Cette généralisation préserve la majorité des propriétés initiales tout en permettant de traiter de façon réelle et efficace l'imperfection au sein du temps ainsi que de quantifier les durées de façon imprécise : un évènement s'est déroulé il y a *longtemps* par rapport à une date donnée, *approximativement au même moment*, etc. On voit ici que, malgré son caractère omniscient et implacable, le temps est une notion qui peine à être pleinement saisi et où l'humain l'appréhende de façon imprécise.

Ainsi, de nombreuses disciplines modélisent leurs objets d'étude comme des séquences chronologiques d'évènements. Sans chercher l'exhaustivité, on peut citer l'Histoire et l'archéologie pour la représentation et l'analyse de périodes historiques imprécises [207, 55], la sociologie dans l'étude et comparaison des trajectoires de vie [1, 135], en médecine pour modéliser l'évolution des maladies [169] ou en musique où les partitions forment une représentation temporelle particulièrement complexe des objets

du discours musical [161]. Enfin, en informatique les problématiques liées au temps sont aussi variées que l'ordonnancement de processus [127] à la représentation, formalisation et analyse automatique de telles séquences d'évènements. On voit alors en un clin d'oeil que le temps, de par sa nature universelle, trouve sa place dans une pléthore de disciplines.

Concernant la représentation des activités humaines dans le temps, c'est en géographie et dans le domaine des Systèmes d'Information Géographique (SIG) que des cadres de travail particulièrement féconds ont émergé, parvenant à capter à la fois les dimensions temporelle, spatiale et sémantique de la mobilité dans une recherche d'étude exhaustive de l'individu et de son environnement. La contribution la plus significative dans le domaine de la géographie analytique qui a fait un usage explicite du temps comme variable dans l'étude des processus spatiaux est sans doute celle de la *Time-Geography* d'Hägerstrand [92]. Aujourd'hui, par extension, l'approche dite "activité-centrée" vise à étudier les comportements de mobilité en fonction de l'enchaînement temporel et spatial des actions qui motivent les déplacements. Ceci permet de montrer comment les formes de mobilité varient au sein d'une population selon des facteurs sociaux (styles de vie) et selon des contraintes spatio-temporelles (budgets-temps, réseau d'offre, services). Ce prisme révèle à quel point le temps peut entretenir un lien intime avec la dimension spatiale mais également avec la dimension sémantique / qualitative de *ce que fait l'humain* et met l'accent sur les ressorts de la mobilité plutôt que sur ses manifestations spatiales [37].

2.2 La Time-Geography

Dans la suite de cette section, nous détaillons le modèle de la Time-Geography et les contributions théoriques que ce cadre de pensée a apportées à la modélisation des activités quotidiennes des individus. Dans un premier temps nous définissions les concepts clés avant de nous intéresser aux apports sur de la Time-Geography sur la représentation et l'incorporation de la sémantique au sein du temps et de l'espace.

2.2.1 Cadre classique de la Time-Geography

La *Time-Geography*, définie par Hägerstrand en 1970 [92], est un cadre d'étude permettant de clarifier les relations, activités et processus spatio-temporels d'un individu avec son environnement. Un des objets d'analyse de la Time-Geography est la notion de *contrainte* qu'exerce l'environnement et l'individu sur lui-même en réflexion à son comportement. Trois types sont identifiés : (i) les *contraintes de capacité* (capacity contraints) qui limitent les activités de l'individu par ces simples capacités physiques. Par exemple, un individu se doit de remplir certains besoins fondamentaux comme manger et dormir, ce qui le contrait à allouer du temps et un lieu particulier à ce type d'activité. (ii) les *contraintes d'interaction* (coupling contraints) traduisent la nécessité pour l'individu d'interagir avec ses pairs dans l'espace-temps afin qu'ils puissent mener à bien une activité particulière, par exemple le travail, les réunions ou toute

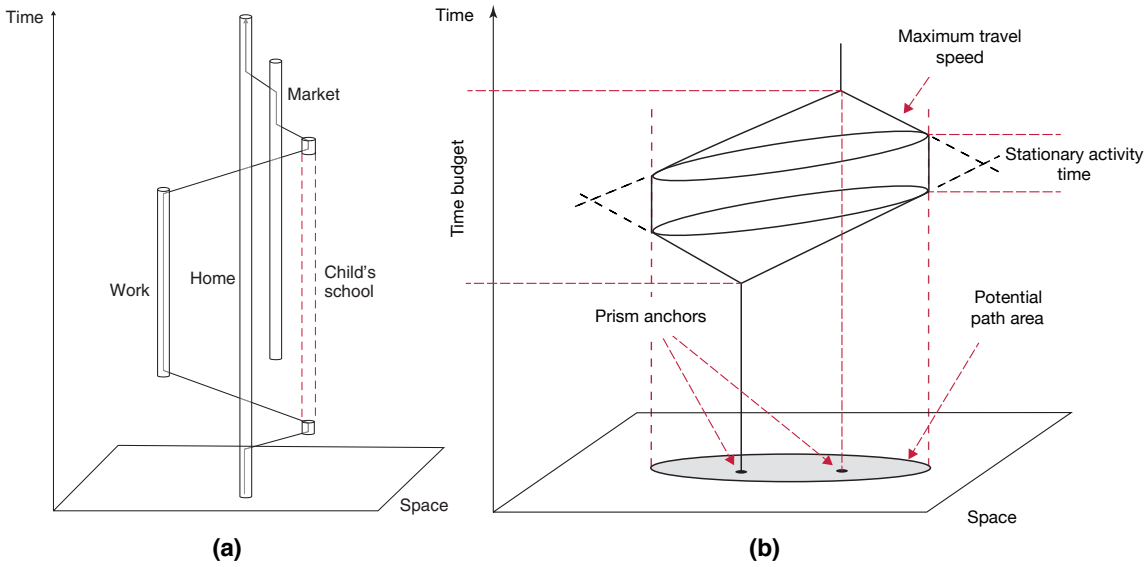


Figure 2.1 – Concepts fondamentaux de la Time-Geography : (a) le parcours spatio-temporel et (b) le prisme spatio-temporel [158]

activité relative à un besoin de sociabilité. Enfin, (iii) les *contraintes d'autorité* (authority contraints) font référence à des restrictions d'ordre légal sur des domaines spatio-temporels particuliers. Par exemple, l'accès à un centre commercial ou un parc peut être interdit à certaines heures.

La Time-Geography est cohérente avec certaines idées fondamentales dans des domaines tels que la géographie, les transports, les sciences urbaines et les sciences sociales. Il s'agit notamment d'une perspective intégrée des phénomènes humains et physiques, de la nécessité de construire des explications au niveau macro à partir d'un traitement au niveau micro et de situer les activités humaines dans leur contexte [189], notamment à l'aide de *budle*, c'est-à-dire par la superposition de plusieurs parcours spatio-temporels. Les concepts de base de la Time-Geography, tels que la faible distribution des évènements dans le temps et l'espace, la disponibilité limitée du temps et l'échange du temps contre l'espace pour accéder aux activités semblent banals, car ils sont courants et correspondent à l'expérience quotidienne. C'est pourtant la raison pour laquelle la Time-Geography est nécessaire : ces facteurs apparemment banals mais tout à fait cruciaux dans nos explications scientifiques du comportement humain ne doivent pas être négligés. La Time-Geography fournit un cadre qui exige la reconnaissance des contraintes fondamentales qui sous-tendent l'expérience humaine et fournit également un système conceptuel efficace pour suivre ces conditions.

Une des grandes contributions de la Time-Geography est le système de notations qu'elle propose qui s'accompagne d'une modélisation graphique conçue comme méthode de visualisation des phénomènes spatio-temporels. En outre, les deux grands concepts définis sont le *prisme spatio-temporel* (space-time prism) et le *parcours spatio-temporel* (space-time path) représentés sur la Figure 2.1 empruntée à [158].

Dans ce même ouvrage, Miller décrit le parcours spatio-temporel comme un ensemble d'activités réparties de façon éparse dans le temps et l'espace et disponibles pour une durée limitée dans un nombre relativement restreint de lieux. La figure 2.1 (a) illustre un parcours spatio-temporel entre des stations d'activités. Les stations sont des lieux où des activités peuvent se dérouler. D'un point de vue spatio-temporel, celles-ci sont représentées comme des tubes qui décrivent leur emplacement dans l'espace et leur disponibilité dans le temps (par exemple, les heures de travail, les horaires d'ouverture d'un magasin, etc.). Le prisme spatio-temporel est décrit quant à lui comme l'enveloppe de tous les parcours spatio-temporels possibles entre deux lieux et horaires connus. La figure 2.1 (b) illustre un prisme spatio-temporel. On remarque que deux ancres spatio-temporelles encadrent le prisme (prism anchors). Ces ancres spatio-temporelles sont souvent des lieux à une période de temps donnée où des activités fixes obligent à la présence de l'individu, par exemple le travail, un commerce, etc. Ainsi, étant donné une vitesse maximale de voyage, conditionnée par le mode de mobilité (à pied, en voiture, etc.), et les points d'ancrage spatio-temporels, il est possible de définir l'enveloppe de tous les parcours spatio-temporels possibles entre ces deux ancres. L'empreinte spatiale du prisme est la zone de trajectoire potentielle (potential path area). Il s'agit de la région de l'espace qui est accessible à l'individu. Si une contrainte préalable est donnée quant à cette zone, par exemple si l'on sait que l'individu ne peut sortir d'un certain périmètre spatial, alors cette contrainte se traduit par un cylindre [2] d'activité stationnaire (stationnary activity time) au niveau du prisme.

2.2.2 Une approche sémantique de l'espace et du temps

Ainsi, comme le note Siabato et al. dans leur impressionnante étude [214] [3], la Time-Geography a eu une influence considérable sur l'ensemble des modélisations au sein des SIG et SIG temporels. Miller et Han dans [159] énumèrent les différents domaines où la Time-Geography possède des applications-clé. Citons sans être exhaustif : les bases de données d'objets mobiles [102], la recommandation de services et d'activités [194, 213], l'épidémiologie [42], l'urbanisme et les transports [227, 18].

Néanmoins, comme l'indique Miller dans [158] : "*The Time geography is a* **physical not a behavioral** *theory ; it highlights the necessary spatiotemporal conditions for human activities, but does not explain the sufficient events that lead to specific activities.*".

Ainsi, la Time-Geography ne cherche pas à formuler une explication quant aux déplacements observés. Elle ausculte et sert uniquement de cadre factuel, voire naturaliste, d'observation de la mobilité des individus dans leur milieu. Il faudra attendre l'émergence d'un nouveau cadre conceptuel de l'utilisation du temps au sein des activités

2. On considère ici que la frontière spatiale de contrainte est une ellipse. On renvoie à [157] pour les aspects analytiques de la construction du prisme spatio-temporel.

3. Celle-ci s'accompagne d'une bibliographie interactive `http://spaceandtime.wsiabato.info/tGIS.html`

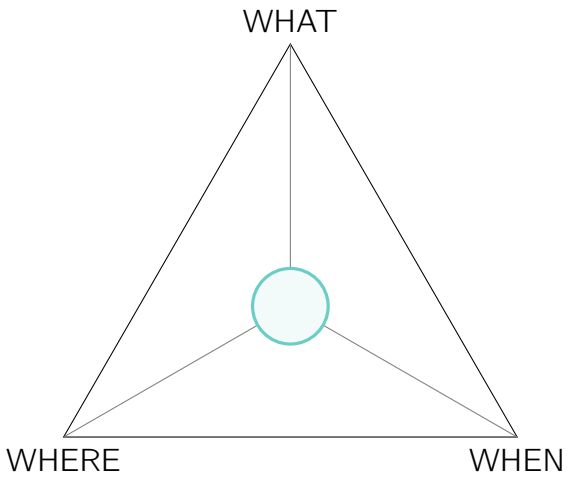


Figure 2.2 – Triade conceptuelle de Peuquet [187]

humaines, initié par Peuquet dans [187], pour s'intéresser à l'aspect *sémantique* de la mobilité et ainsi interroger réellement l'aspect comportemental de l'humain. La figure 2.2 récapitule les trois grands principes énoncés par Peuquet dans sa Triade conceptuelle. L'ambition de Peuquet était de proposer un nouveau cadre de travail en accord avec la façon dont les humains voient le monde. La Triade se base sur des concepts issus de la psychologie de la perception, de l'intelligence artificielle et des techniques de représentation géographiques. Elle permet de fournir une base conceptuelle pour la représentation des bases de données qui soient à la fois flexibles, efficaces et conformes à la façon dont les humains apprennent des concepts et stockent des informations. L'union de trois points de vue de représentation – basés sur l'objet (WHAT), l'emplacement (WHERE) et le temps (WHEN) – et l'incorporation de perspectives objectives et subjectives sur le temps et l'espace permettent une meilleure interrogation des bases de données quant aux questions de "*Ce que fait l'individu et Pourquoi ?*". Comme le souligne Kuhn dans [125], la sémantique au sein des SIG sert à *comprendre* ce qui sous-tend l'objet d'analyse et ainsi mieux l'appréhender, de façon plus naturelle mais aussi plus formelle et analytique. Nous entendons ici par *sémantique* tout type de données qui a du sens pour l'humain, est intelligible immédiatement par lui, intuitive et qui apporte du sens à la trajectoire spatio-temporel en questionnant les actions de l'individu. Or, comme le posent Claramunt et al. dans [47], la triade de Peuquet, bien que formant un changement de paradigme majeur dans la façon de penser le temps et l'espace, reste un cadre descriptif, incomplet dans l'acquisition des connaissances. Ainsi, les auteurs proposent l'intégration des nouvelles perspectives (HOW) et (WHY) qui questionnent le "Comment ?" et le "Pourquoi ?" l'individu agit et qui fondent respectivement les connaissances expérimentales et théoriques quant à un schéma de compréhension globale de l'humain mais aussi de son environnement et des évènements qui dirigent ses choix.

Cette nouvelle façon d'appréhender la mobilité, combinée à l'explosion des données mobiles, GPS, Smartphones et objets connectés ont conduit au développement de

Étape	Type	Références
Apurement & Qualité		[64, 186]
Compression		[61, 38]
Segmentation		[11, 255, 26]
Enrichissement	Réseau routier	[190, 112]
	Point d'intérêt (POI)	[180, 247, 246]
	Météo	[171]
	Agenda d'évènements	[28]
	Réseaux sociaux	[97, 243, 229]
	Inférence d'activités	[20]

Table 2.2 – Montée en sémantique de la trajectoire spatio-temporelle : indications bibliographiques

bases de données d'objets mobiles, à l'analyse et la fouille des données de mobilité orientée sur l'extraction de comportements et motifs sensés, c'est-à-dire qui indiquent une finalité claire quant au déplacement de l'individu. Ce processus d'analyse requiert toutefois une chaîne de traitement pour rendre la trace (i.e., données spatiales de l'individu) opérationnelle : de l'acquisition des données GPS, on obtient une trajectoire brute qu'il faut dans un premier temps apurer. Une seconde phase vise à la compression de la trace apurée, citons pour se faire l'algorithme de Douglas et Peucker qui permet de réduire les points de la trajectoire tout en conservant la nature [61]. Vient ensuite une étape de segmentation de la trajectoire résultante en sous-trajectoires homogènes. À cette fin, on peut citer l'algorithme de SMoT [11] qui permet de détecter les zones d'espace-temps où l'individu est en mouvement ou à l'arrêt. En outre, la trajectoire résultante se compose d'épisodes de STOP – où l'individu est à l'arrêt et/ou en activité sur un lieu donné – et MOVE – où l'individu se déplace. Un processus d'annotation et de map maching permet enfin d'enrichir la trajectoire STOP-MOVE d'informations sémantiques. On renvoie à la table 2.2 qui fournit quelques indications de lecture sur les différentes étapes de la montée en sémantique des trajectoires spatio-temporelles.

2.3 Les trajectoires sémantiques

On qualifie de *trajectoires sémantiques* les objets résultants du processus de traitement brièvement décrit dans la table 2.2 des points GPS. La figure 2.3 donne une vue synthétique des différentes étapes. On peut se référer aux articles suivants [249, 248] de Yan et al. qui présentent une méthodologie d'enrichissement complète de la trajectoire brute à la trajectoire sémantique ainsi qu'un framework, SeMiTri, dédié à l'annotation des objets mobiles [247]. Ainsi, si la plupart des auteurs s'entendent sur ce qu'est [ou n'est pas] une trajectoire sémantique, de nombreuses définitions ont été données au cours du temps.

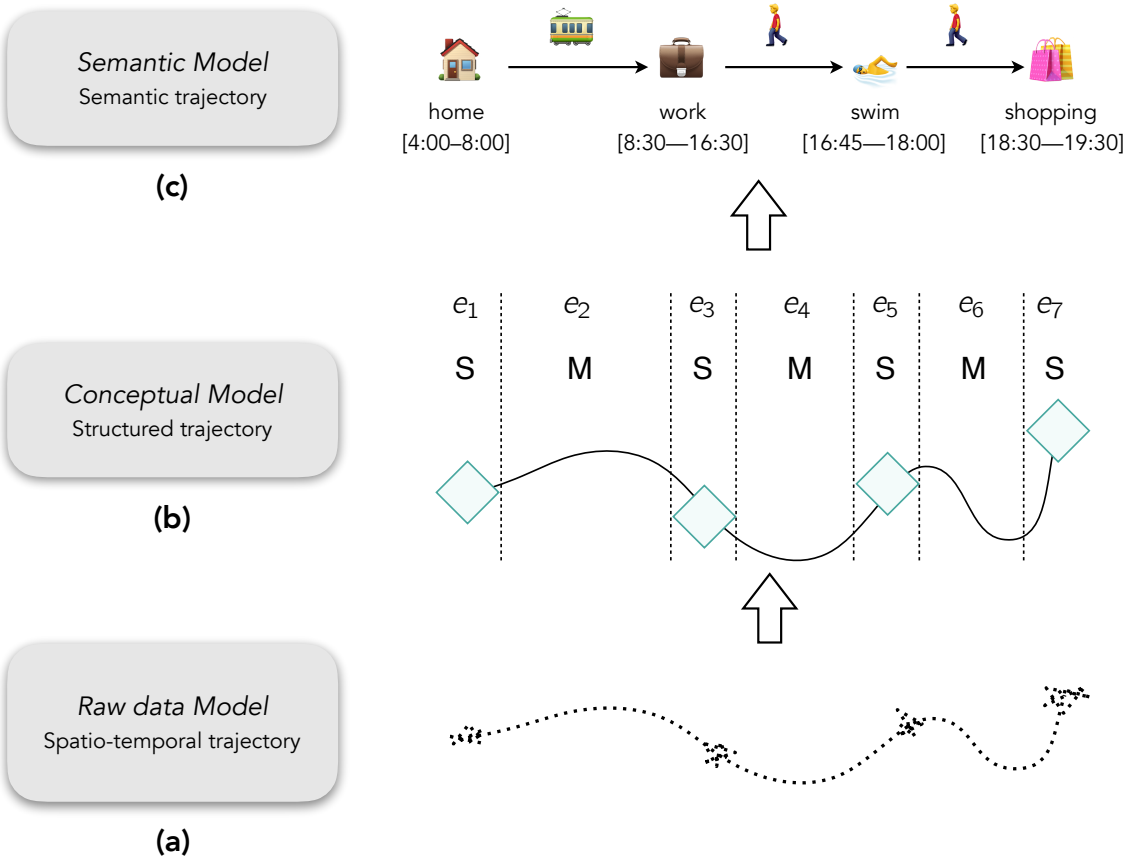


Figure 2.3 – Différentes perspectives de la trajectoire selon [248] : (a) Séquence de points GPS (b) Séquences d'épisodes STOP-MOVE (c) Séquences d'épisodes avec annotations sémantiques

Si l'on revient à la genèse des trajectoires sémantiques, la première occurrence du terme vient de Spaccapietra et al. où ils donnent dans [222] une première définition et modélisation conceptuelle de la trajectoire sémantique. Celle-ci est décrite comme une séquence chronologique constituée alternativement de STOPS et de MOVES. Un STOP est défini de façon intuitive comme une période de temps $T_{stop} = [t_{begin}, t_{end}]$, non vide, et un ensemble de points GPS $P_{stop} = \{(x, y, t) | t \in T_{stop}\}$ tel que la distance spatiale d (e.g., euclidienne, Haversine), entre tout couple de points est inférieure à un seuil $\epsilon > 0$ donné i.e., $\forall p, p' \in P_{stop}, d(p, p') < \epsilon$. Un MOVE est alors défini comme l'ensemble spatio-temporel (non vide) qui sépare deux STOPS consécutifs. Ainsi, les auteurs proposent une modélisation orientée objet des trajectoires sémantiques à l'aide d'un nouveau patron de conception (design pattern) basé sur la notion de STOP et MOVE. Ce design pattern inclut également la possibilité d'une modélisation de la trajectoire, des MOVES et des STOPS en termes d'attributs métiers, par exemple la direction de la trajectoire (e.g, Nord → Sud), la météo, l'activité menée au cours d'un stop, etc.

Dans [11], Alvarès et al. approfondissent les premiers résultats établis par Spacca-

pietra en proposant un algorithme de segmentation STOP-MOVE, SMoT. Dans ce même article, un STOP est rattaché systématiquement à une certaine sémantique, par exemple Travail, Hôtel, etc. L'idée est, comme le proposait initialement Peuquet dans [187], de permettre un requêtage des bases de données plus simple, de haut niveau et naturel pour l'humain en s'aidant du *concept* qu'incarne le STOP de l'individu. Ainsi, si l'on emprunte le vocabulaire de Peuquet, on bascule ici du WHERE (dimension spatiale) au WHAT (dimension sémantique).

C'est grâce à l'article [182] de Parent et al. que l'étude et la conceptualisation des trajectoires sémantiques se consolide. Dans celui-ci, la trajectoire sémantique est définie comme une trajectoire ayant subi un processus d'enrichissement similaire à celui décrit Table 2.2 et figure 2.3. Ainsi, chaque segment de la trajectoire est enrichi d'annotations. Une *annotation* est décrite comme toute donnée additionnelle externe permettant de fournir une information supplémentaire quant au segment spatio-temporel qu'elle vient enrichir. Par exemple, il peut s'agir de données contextuelles externes comme le lieu, l'activité pratiquée, la météo ou un mode de transport. In fine, la trajectoire sémantique est représentée comme une séquence d'épisodes ; un *épisode* étant défini comme un sous-segment maximal homogène d'une trajectoire, c'est-à-dire telle que toutes ses positions spatio-temporelles sont conformes à l'annotation donnée. Par exemple, si une annotation TransportationMode = walk vient labéliser un segment de la trajectoire, alors toutes les positions spatio-temporelles du segment doivent correspondre à l'activité et mode de déplacement "marche à pied". Dans [182], les auteurs proposent une segmentation selon le paradigme STOP-MOVE, puis d'annoter chaque STOP avec l'activité et le lieu correspondants à l'arrêt détecté et chaque MOVE au mode de déplacement utilisé.

Le modèle CONSTAnT [28] de Borgorny et al. se présente comme une continuité du travail initié précédemment par Parent et al.. Partageant une conception similaire de la trajectoire sémantique, les auteurs proposent un modèle conceptuel complet qui définit les aspects les plus importants pour créer un concept général de trajectoire sémantique. CONSTAnT est représenté sur la Figure 2.4, on détaille les éléments clés du modèle : la trajectoire sémantique est représentée comme une séquence de *sous-trajectoires sémantiques* (SemanticSubTrajectory). À l'instar de la notion d'épisode donnée par Parent et al., une sous-trajectoire sémantique est un segment homogène de la trajectoire composé d'un ensemble de *points sémantiques* (SemanticPoint). Le point sémantique est l'élément primordial du modèle, il s'agit en substance d'un point GPS enrichi de diverses annotations sémantiques telles que l'environnement (Environment) qui permet de décrire par exemple la température ou pression de l'atmosphère, et un ensemble de lieux (Place) où le point est enregistré. Un lieu peut être lié ou non à un évènement qui se produit pendant une période de temps donnée. Par exemple, un évènement peut prendre la forme d'un spectacle musical le 20 juillet 2020, de 20 :00 à 23 :00, sur la place Louis XII de Blois.

La seconde partie du modèle (classes en gris foncé) est plus complexe et nécessite des méthodes avancées de fouille de données ou être renseignée de façon déclara-

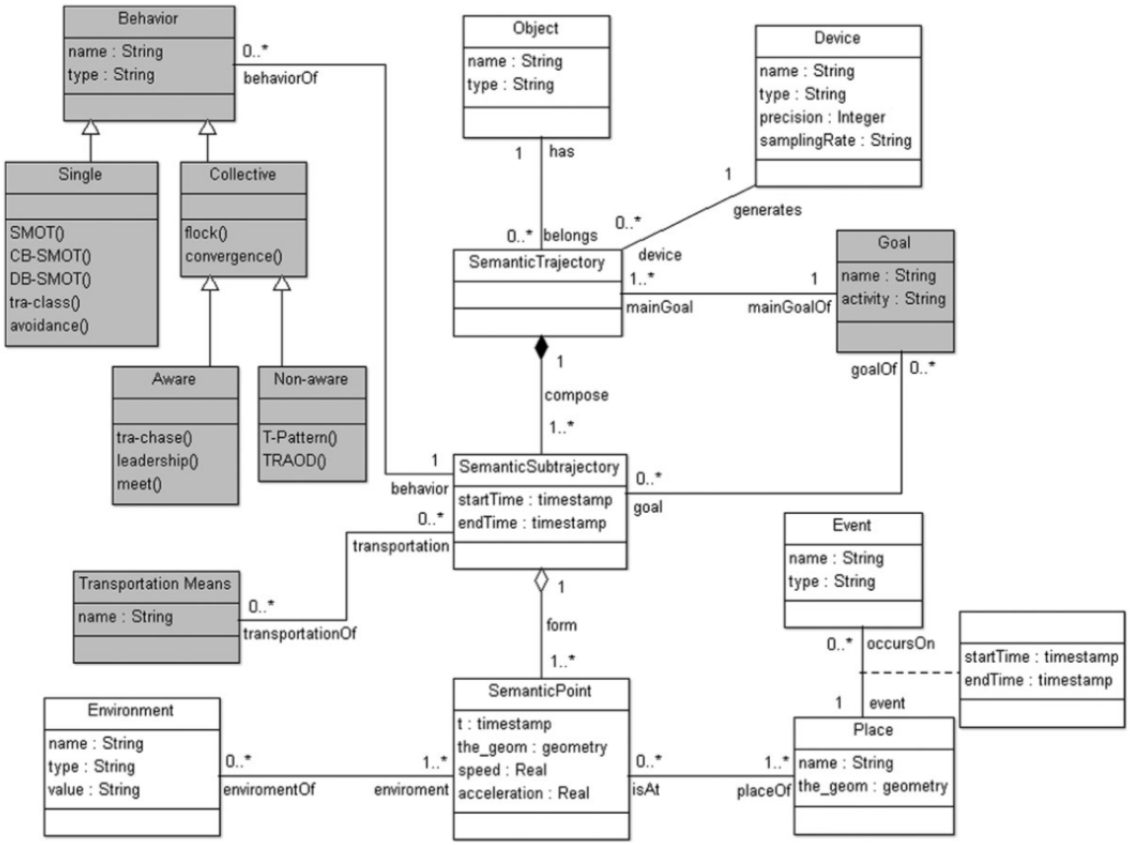


Figure 2.4 – Modèle conceptuel des trajectoires sémantiques : CONSTAnT [28]

tive par l'individu pour être instanciée. Elle concerne le but (Goal) de la trajectoire sémantique ou des sous-trajectoires sémantiques, le moyen de déplacement (TransportationMeans) et le comportement (Behavior). Un *but* représente en substance la raison, le *pourquoi*, du mouvement de l'individu, par exemple "aller au travail", ou "faire du jogging". Le *comportement* quant à lui est une caractéristique importante d'une trajectoire sémantique, préalablement introduite par Parent et al. dans [182]. Le comportement d'une trajectoire est un ensemble de caractéristiques qui identifie un comportement particulier d'un individu ou groupe d'individus en mouvement. En d'autres termes, le comportement peut montrer *comment* (HOW) et *pourquoi* (WHY) un individu se déplace, ajoutant ainsi une sémantique supplémentaire au mouvement. Nonobstant, le comportement d'une trajectoire ou d'un groupe de trajectoires ne peut pas être compris ou déduit de façon isolée et/ou triviale. En effet, l'acquisition de comportement demande bien souvent l'utilisation de techniques sophistiquées ainsi qu'une solide connaissance experte du sujet étudié. En outre, la fouille de données et l'analyse statistique sont des outils précieux pour la découverte de motifs (*pattern*) et l'inférence de comportements.

Ainsi l'étude de la dimension sémantique au sein des trajectoires, davantage porteuse de sens quant au comportement humain [76], va progressivement s'émanciper de l'as-

pect spatial. Dans [174], Noel et al. proposent un modèle de trajectoire sémantique multi-aspects permettant de décrire selon différents points le contenu de trajectoires de vie. Dans [90], Güting et al. proposent un modèle générique simple et flexible pour représenter tout type d'information sémantique que l'on pourrait vouloir associer à une trajectoire. Ce modèle, que les auteurs nomment *trajectoire symbolique*, prend la forme d'une séquence chronologiquement ordonnée de paires $\langle (i_1, L_1), ..., (i_n, L_n) \rangle$ où i_k est un intervalle de temps et L_k un label (e.g., chaîne de caractères ou emoji) ou ensemble de labels quelconques. Les intervalles de temps sont disjoints ou éventuellement adjacents. Par exemple, une trajectoire symbolique simple peut prendre la forme de la figure 2.5.

$\langle ([8:00-8:45[, 🏠), ([8:45-9:30[, 🚃), ([9:30-9:45[, 🗼) \rangle$

Figure 2.5 – Exemple de trajectoire symbolique [90]

Par sa simplicité, ce modèle se prête à une formulation élégante et expressive des requêtes à l'aide d'automates finis et d'expressions régulières telle qu'imaginée par du Mouza et al. [62], mais aussi à un large panorama de disciplines. Ainsi, cette modélisation ultra générique peut-être adoptée pour représenter et analyser toute entité qui évolue dans le temps et dont les états possèdent une durée. De plus, comme cette représentation n'incorpore par les points GPS, elle permet de s'extraire des problématiques liées à l'anonymat des données.

Dans l'article [254], Zhang et al. empruntent une modélisation des trajectoires sémantiques similaire à celle de Güting et al. mais y incorporent la dimension spatiale. Les labels sont identifiés à des lieux (POI) et la position géographique de ces lieux est prise en compte. D'ailleurs, dans [232], Valdés et Güting enrichissent leur modèle des trajectoires symboliques et améliorent l'efficacité de leur système de requêtage à l'aide d'expressions régulières pour tenir compte des aspects multi-attributs et géométriques des trajectoires.

Parallèlement à cette vision orientée base de données des trajectoires sémantiques / symbolique, des approches issues du domaine des ontologies et de la représentation des connaissances ont émergé avec pour dessein de structurer à la fois l'information et l'essence même des trajectoires sémantiques. Parmi les représentants notables de cette approche, nous pouvons citer les frameworks dédiés à la structure et l'enrichissement des trajectoires sémantiques, Baquara[2] de Fileto et al. [78] et FrameSTEP de Nogueira et al. [176] ainsi que l'ontologie datAcron[4], portée par Vourous et al. [236, 235], conçue pour fournir un modèle commun et adaptatif pour un large panel de disciplines liées à la mobilité spatio-temporelle et sémantique avec pour finalité la prédiction, l'analyse et la visualisation à différents niveaux de granularité des informations liée aux trajectoires. Mentionnons également le modèle ontologique de Noel et

4. `http://ai-group.ds.unipi.gr/datacron_ontology/`

al. [175] permettant de modéliser des trajectoires sémantiques multidimensionnelles et incluant des facteurs explicatifs pour une meilleure compréhension des évènements au sein la trajectoire. En outre, le design pattern proposé par les auteurs est appliqué au cadre de la modélisation des trajectoires de vie.

Enfin, concernant la modélisation sémantiquement riche des trajectoires, citons dernièrement le modèle MASTER de Mello et al. [151]. Centré sur une vision Big Data, MASTER propose une solution complète – du modèle conceptuel au stockage et à l'interrogation des données, exploitant l'abondance d'informations produites au cours des déplacements. Ces données, nommées *aspects* par les auteurs, sont très semblables aux annotations décrites dans [182] par Parent et al. mais concernent ici n'importe quelle information relative à l'individu ou à la trajectoire. La figure 2.6 montre un exemple de trajectoire multi-aspects issue de MASTER. On remarque une hétérogénéité forte des informations prises en considération : le lieu, l'activité, le moyen de transport mais aussi la météo, le rythme cardiaque de l'individu, ses messages postés sur les réseaux sociaux ou encore toute donnée relative aux POI visités (température, pollution, standing des établissements, etc.).

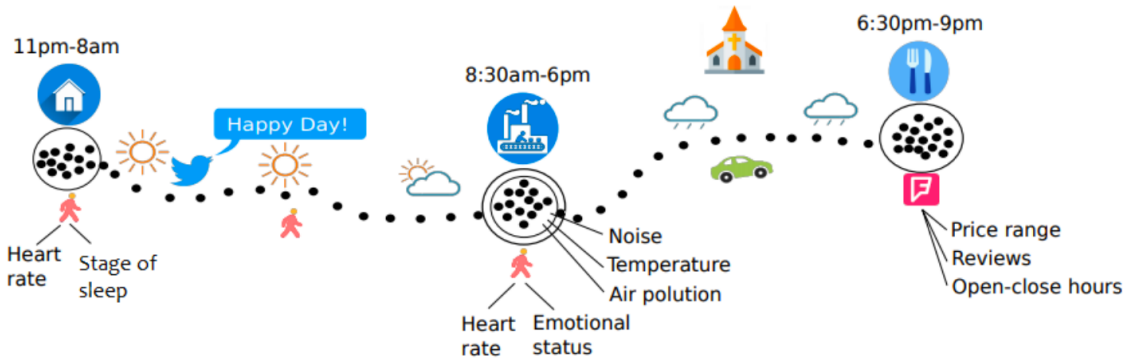


Figure 2.6 – Un exemple de trajectoire multi-aspects [151]

Malgré les avantages et forces incontestables en termes de représentation et de gestion de l'information du modèle MASTER, celui-ci pose de nombreuses questions pour la plupart relatives aux interrogations classiques concernant le Big Data [108].

Une première question peut se poser en matière de protection de la vie privée des utilisateurs [33]. En effet, le modèle MASTER tire partie du fait que les populations occidentales, aujourd'hui, fournissent généreusement nombre de données et réponses émotionnelles, psychologiques et/ou physiologiques quant à leur vie quotidienne par l'usage de leur smarphone et des réseaux sociaux. Néanmoins, le public craint beaucoup l'utilisation inappropriée des données personnelles, notamment par la mise en relation de données provenant de sources multiples. De plus, la disponibilité de ces données reste un privilège très largement exclusif aux seules entreprises et institutions qui les produisent et qui ont instauré un système de gouvernance stricte quant à qui peut les consulter de l'extérieur. En outre, les problématiques liées au consentement

éclairé sur l'utilisation des données et de protection efficace de la vie privée sont des défis à la fois technique, juridique et éthique qui doivent être abordés conjointement par ces disciplines pour concrétiser les promesses du Big Data.

Une deuxième question est relative à l'hétérogénéité et à l'abondance des données prises en compte. Si les humains tolèrent et consomment aisément des informations hétérogènes, incertaines ou imprécises – la nuance pouvant fournir une profondeur précieuse, les données récoltées doivent être remises en contexte pour tirer une information de qualité, assimilables par la machine et en adéquation avec la réalité. En outre, l'inconsistance, l'incomplétude et la véracité des données sont des problèmes endémiques aux systèmes Big Data qu'il convient de gérer. De plus, les algorithmes d'analyse automatique s'attendent généralement à des données homogènes et structurées. Par conséquent, les données doivent être soigneusement remaniées dans un format standard avant de débuter un traitement analytique des données.

La dernière question concerne d'ailleurs l'analyse des données. Outre le fait que l'on peut interroger la pertinence de mêler des données au caractère aussi hétérogène entre elles, ce qui peut conduire à la découverte de corrélations spécieuses [253], un vrai défi se pose quant à la comparaison de telles trajectoires. Si certaines solutions sont proposées dans la littérature [79, 151], l'excès d'information peut conduire à une surcharge cognitive de l'individu ce qui conduit à un tri et un recul analytique impossible [19, 145]. En outre, et même si les systèmes peuvent supporter une masse considérable d'information, il faut également s'assurer que les utilisateurs finaux – humains – puissent "absorber" correctement les résultats de l'analyse et ne pas se perdre dans un océan de données. C'est pourquoi la *perspective de l'humain* dans le processus d'acquisition des connaissances, la simplicité de restitution et de visualisation des données ainsi que la concision des analyses doivent demeurer une préoccupation centrale dans l'ensemble des choix et méthodes utilisés.

Dans la suite de cette dissertation, nous nous concentrerons exclusivement sur l'étude des dimensions temporelle et sémantique des trajectoires, la dimension spatiale étant de moins en moins pertinente pour qualifier des types de comportements similaires, c'est-à-dire adoptant les mêmes actions / activités mais se déroulant dans des lieux géographiques différents. Typiquement, on peut imaginer deux individus qui fréquentent deux supermarchés différents au même instant. Ceux-ci sont guidés par la même finalité, le même comportement : celui de réaliser un achat d'achat, pourtant la dimension spatiale est partiellement incapable de rendre compte, sans l'aide de la sémantique, de ce type d'information. De plus, occulter la dimension spatiale permet également de saper, en partie, les problématiques liées à l'anonymat des données et la vie privée des utilisateurs.

Dans une perspective de généricité, nous nous ré-approprierons partiellement les modèles de Güting et al. [90] (trajectoire symbolique) ainsi que celui de Parent et al. [182] (trajectoire sémantique STOP-MOVE) tout en conservant une approche empreinte des concepts sociologiques portés par la Time-Geography. Faisant fi de la dimension spa-

tiale, nous utiliserons plus loin le terme *séquence* plutôt que trajectoire. De plus, dans un souci d'adopter une vision thématique souple, structurée et partagée entre les différents acteurs de projet (experts et non-experts), nous proposons de modéliser les symboles de ces séquences sémantiques (POI, activités, modes de transport, etc.) à l'aide d'ontologies afin à de disposer d'une interface commune de représentation du monde. Ces ontologies ont ainsi pour but de fournir un alignement des connaissances et du vocabulaire ainsi qu'une vision du monde opérationnelle, structurée et compréhensible autant pour l'humain que la machine. En outre, elles permettent le calcul automatique de mesures de similarité sémantiques utiles pour la comparaison fine de séquences sémantiques que nous abordons dans le prochain chapitre.

Chapitre 3
Comparaison de séquences sémantiques

3.1 Comparaison de concepts : La sémantique

La sémantique peut être simplement définie comme la notion de *sens* attribué dans un contexte sens métier à un terme ou symbole par le biais d'un phénomène de conceptualisation. Dans la suite, nous donnons de premières définitions quant au phénomène de conceptualisation puis son implémentation effective en intelligence artificielle grâce aux graphes de connaissances et ontologies. Enfin, dans une seconde partie, nous abordons la problématique de la comparaison automatique de concepts au sein des graphes de connaissances par le biais de mesures de similarité sémantique. Nous concluons cette section par un tableau qui synthétise l'étude de l'ensemble des mesures abordées.

3.1.1 Concepts et ontologie

La notion de *concept*, à la croisée des chemins de la philosophie, de la psychologie, de la linguistique et de l'ingénierie des connaissances, est un terme difficile à définir et fait encore débat au sein des différentes communautés. La Stanford Encyclopedia of philosophy définit le concept [1] comme l'unité élémentaire de la pensée, le résultat d'une opération d'abstraction par laquelle l'esprit rassemble, sous un terme unique, les relations et traits communs à un ensemble d'instances [147]. Dans notre cas, nous désignerons par concept tout symbole issu d'une trajectoire symbolique [90] qui peut être conceptualisé. On glisse ici de la *trajectoire symbolique* à la *trajectoire sémantique*.

Selon Mervis et Rosh [154], le processus de conceptualisation peut prendre deux formes dimensionnelles qui donnent lieu à deux représentations majeures des connaissances :

- Horizontale i.e., une segmentation catégorielle en propriétés caractéristiques. Par exemple le concept *oiseau* est caractérisé par les propriétés [`aDesPlumes`] et [`saitVoler`].

1. Les termes *idée* ou *catégorie* sont également utilisés.

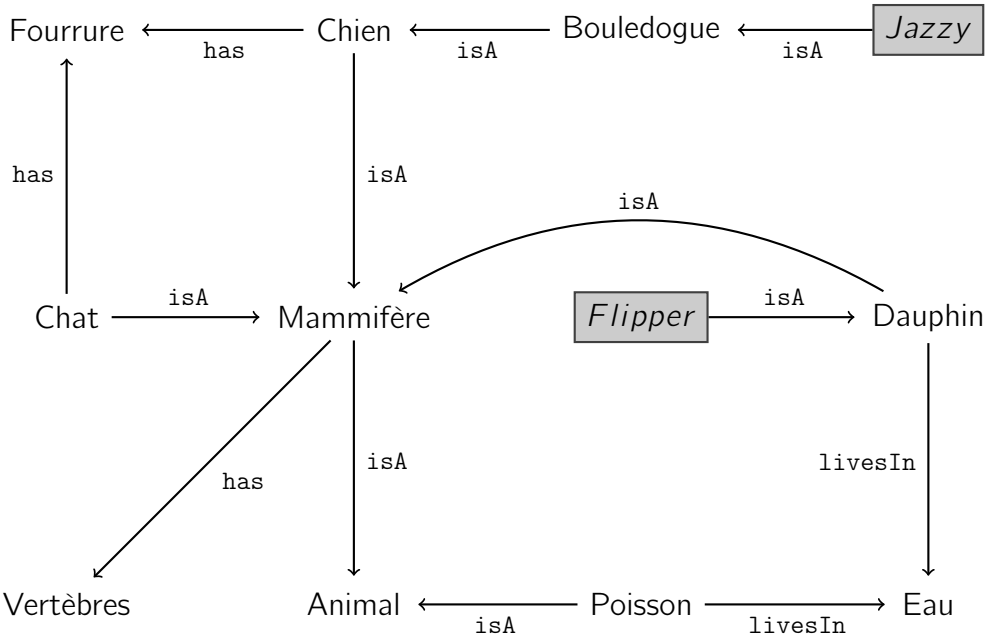


Figure 3.1 – Exemple de graphe de connaissances en RDF. Les noeuds gris désignent des instances du monde réel.

- Verticale i.e., différents niveaux hiérarchiques où l'ascendance est une inclusion d'ensembles de concepts. Par exemple Bouledogue $\subset$ Chien $\subset$ Canidé $\subset$ Mammifère.

La première forme, horizontale, hérite de la conception aristotélicienne [2] et est nommée *théorie des Conditions Nécéssaires et Suffisantes*. L'autre, plus expérimentaliste se base sur la *Théorie des prototypes* de la psychologue Eleanor Rosh [198].

L'intelligence artificielle tenta de concilier ces deux visions au sein d'un unique paradigme de représentation, celui des graphes de connaissances [221]. Un *graphe de connaissances* (ou réseau sémantique) représente formellement la sémantique en décrivant les entités et leurs relations à l'aide d'un graphe dirigé et acyclique où les noeuds du graphe sont des concepts (ou instances pour les noeuds inférieurs du graphe) et les arcs décrivent les propriétés sémantiques des concepts. Le modèle RDF (Resource Description Framework) fournit un standard de représentation pour de tels graphes à base de triplets (*sujet, prédicat, objet*) où :

- Le *sujet* est un noeud et représente la ressource à décrire.
- Le *prédicat* est un arc et représente une propriété applicable au sujet.
- L'*objet* est un noeud et représente la valeur de la propriété.

Un exemple simple de graphe de connaissances au format RDF est donné figure 3.1.

Dans [87], Gruber pose la définition théorique de ce qu'est une *ontologie* au sens informatique : "*An ontology is an explicit specification of a conceptualization.* [...] *A*

2. Voir *Les Catégories* d'Aristote

conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose.". Ainsi, une ontologie est un outil permettant de structurer, représenter et partager des corpus de connaissances rattachés à un domaine, utilisable par un ordinateur et permettant de raisonner par transitivité sur les concepts. Dans la figure 3.1, on peut conclure que *Flipper* `has` Vertèbres.

De façon opérationnelle, une ontologie peut être vue comme un graphe de connaissances qui regroupe un ensemble de concepts suffisants pour décrire un domaine. Ces concepts sont liés les uns aux autres par des relations taxinomiques type `isA` ou `partOf` (hiérarchisation des concepts) et sémantiques. Pour de plus amples explications et détails sur les ontologies, on pourra se référer par exemple à Guarino et al. [88].

Aujourd'hui, l'émergence du Web 3.0 – *Web sémantique* et du Linked Open Data (LOD), supervisée par le World Wide Web Consortium (W3C), encourage l'utilisation de données structurées au format RDF dans un but de partage des connaissances et d'interropérabilité entre les systèmes d'information [24]. L'initiateur de ce projet, Tim Berners-Lee, annonce que : "*Properly designed, the Semantic Web can assist the evolution of human knowledge as a whole*". Une vue concrète de ce fait est la croissance rapide du projet DBpedia [134] qui regroupe aujourd'hui environ (dans sa version anglaise) 4.58 millions d'objets référencés et structurés, dont 4.22 correctement classifiés à l'aide d'une ontologie cohérente [3].

Outre les projets "cathédrale" comme DBpedia qui ont pour ambition de structurer l'ensemble de la connaissance humaine, on compte de nombreuses initiatives restreintes à des domaines spécialisés permettant de structurer l'information. Un exemple issu du domaine des SIG est le projet Linked Geo Data project [4] qui fournit un graphe RDF de points d'intérêt (POI) géographiques depuis Open Street Map ; un second exemple est le réseau Wordnet [5] qui forme une base de données lexicales structurée selon des relations taxomoniques. Si nous nous concentrons dans le domaine de la mobilité quotidienne, ces ontologies et taxonomies permettent d'établir une classification de l'information afin d'évoluer à différents niveaux de granularité de l'information et de comparer les concepts en jeu. À cet effet, les *Time-Use surveys* qui enquêtent sur la structuration des emplois du temps des individus établissent une taxonomie normalisée des activités quotidiennes [6]. Les taxonomies forment une grande partie des ontologies [86]. Calquées sur le modèle biologique de classification du vivant, elles se retreignent aux relations hiérarchiques de type `isA` et ont, en conséquence, l'avantage d'être simples à construire et de permettre le calcul de mesures de similarité entre les termes, basées sur les structures hiérarchiques de façon rapide et efficace.

Nous décrivons dans la section suivante les différentes mesures de similarité séman-

3. `https://wiki.dbpedia.org/about/facts-figures`
4. `https://www.geonames.org/`
5. `https://wordnet.princeton.edu/`
6. `https://ec.europa.eu/eurostat/documents/3859598/11597606/KS-GQ-20-011-EN-N.pdf/2567be02-f395-f1d0-d64d-d375192d6f10?t=1607360062000`

tique existantes entre concepts avec une emphase particulière pour les mesures où les concepts sont décrits à l'aide d'une taxonomie.

3.1.2 Mesures de similarité sémantique

Pouvoir comparer des objets est un pré-requis pour de nombreuses techniques de machine learning. Admettons que l'on souhaite former des groupes d'individus similaires, alors il nous faut définir et qualifier exactement ce qui fait la *similarité*, la ressemblance ou encore la proximité entre deux individus.

En mathématiques, la notion de *distance* sert à quantifier l'éloignement (i.e., différence) entre deux objets issus d'un même ensemble et généralise l'idée intuitive de longueur qui sépare deux points. Plus formellement, soit un ensemble Σ d'entités, une distance d est une application $d : \Sigma \times \Sigma \rightarrow \mathbb{R}^+$ satisfaisant les axiomes suivants :

1. Symétrie : $\forall x, y \in \Sigma, d(x, y) = d(y, x)$
2. Séparabilité : $\forall x, y \in \Sigma, d(x, y) = 0 \Leftrightarrow x = y$
3. Inégalité triangulaire : $\forall x, y, z \in \Sigma, d(x, y) + d(y, z) \geqslant d(x, z)$

Plus généralement, on appellera *dissimilarité* une application à vocation d'être une distance mais qui viole un des axiomes précédents.

Une *mesure de similarité* est très analogue au concept de distance. En outre, Chen et al. montrent dans [41] que ces deux notions sont équivalentes à une bijection près. Si l'on considère un ensemble Σ, une mesure de similarité sim est une application $sim : \Sigma \times \Sigma \rightarrow \mathbb{R}^+$, toutefois, on notera qu'une large majorité des auteurs préfèrent une similarité normalisée dans $[0, 1]$. On donne les axiomes de la mesure de similarité dans ce cas précis :

1. Symétrie : $\forall x, y \in \Sigma, sim(x, y) = sim(y, x)$
2. Séparabilité : $\forall x, y \in \Sigma, sim(x, y) = 1 \Leftrightarrow x = y$
3. Inégalité triangulaire : $\forall x, y, z \in \Sigma, sim(x, y) + sim(y, z) \leqslant sim(x, z) + 1$

Enfin, précisons ici que nous utiliserons, par commodité, le terme générique de **mesure**[7] pour désigner toute application ayant pour but la comparaison d'entités. Nous userons autant que possible des termes conventionnels "similarité", "dissimilarité" et "distance" lorsque l'ambiguïté n'est pas permise.

La majorité des mesures de similarité sémantiques sont conçues dans le cadre de l'évaluation de la similarité entre paires de concepts définis dans une taxonomie. Ces mesures peuvent être utilisées pour comparer n'importe quelle paire de noeuds exprimés dans un graphe qui définit un ordre partiel, c'est-à-dire tout graphe structuré

7. Le terme est ici à dissocier de celui utilisé communément en analyse et théorie des probabilités.

par des relations transitives, réflexives et antisymétriques comme les relations hiérarchiques `isA` ou `partOf`.

Les principales approches utilisées pour comparer les concepts définis dans une taxonomie sont :

- L'*approche topologique* (ou approche structurelle) basée sur l'analyse de la structure des graphes. Ces méthodes estiment la similarité en fonction du degré d'interconnexion entre les concepts.
- L'*approche par traits* basée sur l'extraction des caractéristiques des concepts au sein du graphe. Ces méthodes estiment la similarité en fonction des caractéristiques partagées et distinctes des concepts.
- L'*approche informationnelle* basée sur l'estimation de la quantité d'information portée par un concept.

En dehors des approches reposant sur les taxonomies, on notera les approches dites *statistiques* qui reposent sur l'utilisation de moteurs de recherche ou la construction de modèles statistiques sur des corpus textuels. Ces approches s'appuient sur le fait que des termes similaires sont souvent présents de façon co-occurente [202, 156]. Parmi les mesures qui utilisent les moteurs de recherche citons par exemple la Normalized Google Distance (NGD) [45] qui se ré-approprie les concepts établis dans [139] et qui, pour deux concepts donnés, calcule un score de dissimilarité selon le nombre d'occurrences retournées par le moteur de recherche Google. Bien que ces méthodes basées sur les moteurs de recherche montrent des résultats expérimentaux intéressants [29], elles posent de nombreuses questions éthiques quant aux critères d'association d'un score de ressemblance à deux concepts ainsi qu'à leur interprétabilité. Une autre approche s'appuie sur l'analyse sémantique latente (LSA) de corpus [56] et les modèles de sémantique vectorielle de Galton pour associer une mesure de similarité à deux concepts à l'aide de la similarité du cosinus. Malheureusement, malgré leur efficacité [129] ces méthodes sont également difficilement interprétables et nécessitent d'importantes ressources textuelles pour être pleinement exploitables.

Dès lors, nous nous focaliserons sur les mesures dédiées aux taxonomies expertes préalablement établies. Précisons encore une fois que nous n'adopterons pas ici une vision holistique. Nous renvoyons au travail très complet de Harispe et al. [96] pour les lecteurs en quête d'exhaustivité sur les mesures de similarité sémantiques et à l'Encyclopedia of distances [59] de Deza et Deza pour vision plus mathématique.

3.1.2.1 Mesures par approche topologique

Avant de détailler quelques mesures classiques de la littérature, nous précisions les notations suivantes employées :

- $G = (\Sigma, A)$ désigne un graphe de connaissances dirigé acyclique où Σ est l'ensemble des noeuds (i.e., concepts) et $A \subset \Sigma \times \Sigma$ l'ensemble des arcs. On précise que le graphe est doté d'un concept "racine" nommé $\texttt{all} \in \Sigma$.

- $x, y \in \Sigma$ désignent deux concepts de G.
- $d : \Sigma \times \Sigma \rightarrow \mathbb{N}$ est la fonction qui, pour $d(x, y)$, retourne le plus court chemin entre x et y selon l'algorithme de Dijsktra. On utilisera également l'application partielle $d(x)$ pour désigner le plus court chemin entre x et le noeud racine `all`.
- $D = \max_{x \in \Sigma}\{d(x)\}$ désigne la profondeur maximale de G.
- $LCA : \Sigma \times \Sigma \rightarrow \Sigma$ est la fonction qui, pour $LCA(x, y)$, retourne le plus petit ancêtre commun (Least Common Ancestor) de x et y, c'est-à-dire le concept le plus profond qui les subsume.

Nous décrivons ici les mesures principales de l'approche topologique dans les taxonomies. Pour une étude plus exhaustive, nous renvoyons à [256].

Mesure de Rada Rada et al. définissent dans [192] la similarité entre deux concepts comme étant l'inverse du plus court chemin entre eux :

$$sim_{rada}(x, y) = \frac{1}{1 + d(x, y)} \tag{3.1}$$

L'inconvénient majeur de cette mesure est qu'elle néglige complètement la forme ainsi que l'amplitude / profondeur du graphe. L'un des principaux défis des concepteurs de mesures sémantiques au fil des ans a donc été d'affiner les mesures en tirant parti des preuves sémantiques liées à la spécificité, typicalité et connotations entre les concepts.

Mesure de Resnik En conséquence du fait précédent, Resnik dans [196] propose de prendre en compte la profondeur maximale du graphe.

$$sim_{resnik}(x, y) = 2D - sim_{rada}(x, y) \tag{3.2}$$

Cette mesure souffre du fait de ne pas être normalisée dans $[0, 1]$, or l'usage veut que les mesures de similarité soient définies dans cet intervalle et valent 1 si les concepts comparés sont identiques.

Mesure de Leacock-Chodorow Afin de normaliser la mesure et de simuler une pondération non uniforme des arcs, Leacock et Chodorow [132] introduisent une transformation logarithmique :

$$sim_{LC}(x, y) = \log_2 \left(2 - \frac{d(x, y)}{2D}\right) \tag{3.3}$$

Mesure de Wu-Palmer Des auteurs ont également proposé de prendre en compte la spécificité des concepts comparés, à l'aide de la profondeur de leur LCA. Par exemple, Wu et Palmer (1994) dans [244] ont proposé d'exprimer la similarité de

deux concepts comme un ratio prenant en compte la profondeur de chaque concept ainsi que la profondeur de leur LCA.

$$sim_{wup}(x, y) = \frac{2 \times d(LCA(x, y))}{d(x) + d(y)} \tag{3.4}$$

Mesure de Li Dans [141], Li et al. reprochent aux précédentes mesures de ne pas être paramètrables pour tenir plus ou moins compte des spécificités topologiques du graphe (profondeur et liens de parenté des concepts). Également, les auteurs suggèrent l'utilisation d'une fonction non linéaire type exponentielle afin de mieux gérer la croissance de la similarité.

$$sim_{li}(x, y) = e^{-\alpha \times sim_{rada}(x,y)} \times \tanh(\beta \times d(LCA(x, y))) \tag{3.5}$$

où $\tanh(\theta) = \frac{e^{\theta}-e^{-\theta}}{e^{\theta}+e^{-\theta}}$ désigne la tangente hyperbolique. Les coefficients $\alpha, \beta \in [0, 1]$ contribuent respectivement à la longueur du chemin et à la parenté des concepts. Les auteurs proposent les paramètres optimaux empiriques suivants : $\alpha = 0.2$ et $\beta = 0.6$.

3.1.2.2 Mesures par approche par traits

L'approche par traits fait généralement référence aux mesures et théories de la perception émises par Tversky, très influencé par la Gestalt Theory [230]. D'un point de vue informatique, l'approche s'inspire du paradigme objet où les concepts sont réifiés en une collection de caractéristiques les décrivant. Cette réduction des concepts à des collections de caractéristiques permet alors de replacer l'estimation de la similarité sémantique dans le contexte des mesures ensemblistes. Une approche couramment utilisée pour représenter les caractéristiques d'un concept au sein d'une taxonomie consiste à considérer ses ancêtres comme des caractéristiques [96]. On désigne par $\Gamma(x)$ l'ensemble des ancêtres (ou hyperonymes) du concept x (lui compris) privé de $\{\texttt{all}\}$.

Depuis l'indice de Jaccard proposé il y a plus de 100 ans [107], de nombreuses mesures ont été définies dans divers domaines. On renvoie à [44] pour une étude exhaustive.

À noter également qu'au sein de ce type d'approche on distingue deux courants : les mesures *intensionnelles* qui considèrent les concepts selon le point de vue de leurs caractéristiques et les mesures *extensionnelles* qui considèrent la ressemblance du point de vues du nombre d'instances partagées par les concepts comparés. On désigne par I l'ensemble total des instances et $\mathcal{I}(x) \subset I$ l'ensemble des instances (ou hyponymes) du concept x.

Mesure de Jaccard La mesure de Jaccard évalue la similarité entre deux concepts par la quantité d'objets partagés sur l'union globale. Elle peut être envisagée soit de façon intensionnelle en considérant les ancêtres des concepts :

$$sim_{jacInt}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{3.6}$$

Ou extensionnelle, c'est alors les instances des concepts qui sont considérés :

$$sim_{jacExt}(x,y) = \frac{|\mathcal{I}(x) \cap \mathcal{I}(y)|}{|\mathcal{I}(x) \cup \mathcal{I}(y)|} \tag{3.7}$$

Mesure de Tversky La mesure de Tversky a été proposée par le psycho-cogniticien Amos Tversky dans [230] et généralise la mesure de Jaccard afin de mieux prendre en compte certaines spécificités de l'esprit humain comme la non-symétrie et non séparabilité [231] lors des processus de comparaison :

$$sim_{tversky}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cap \Gamma(y)| + \alpha|\Gamma(x) \backslash \Gamma(y)| + \beta|\Gamma(y) \backslash \Gamma(x)|} \tag{3.8}$$

où $\alpha, \beta \in [0,1]$ représentent respectivement la contribution relative des caractéristiques uniques de x et de y. On notera que si $\alpha = \beta = 1$, alors $sim_{tversky}(x,y) = sim_{jacInt}(x,y)$

Mesure de d'Amato Si l'on s'intéresse de nouveau aux mesures extensionnelles, nous remarquons que la mesure de Jaccard disqualifie totalement les concepts qui ne partagent aucune instance en commun. D'Amato et al. estiment cependant que deux concepts peuvent être similaires sans pour autant avoir quelque instance en commun, ce qui est le cas pour une similarité de type intensionnelle [65]. En conséquence, les auteurs proposent une nouvelle mesure qui évalue, non pas l'intersection entre les ensembles d'instances de chaque concept, mais la variation de la cardinalité de l'ancêtre commun :

$$sim_{amato}(x,y) = \frac{\min\{\mathcal{I}(x), \mathcal{I}(y)\}}{\mathcal{I}(LCA(x,y))} \times \left(1 - \frac{\mathcal{I}(LCA(x,y))}{|I|}\right) \times \left(1 - \frac{\min\{\mathcal{I}(x), \mathcal{I}(y)\}}{\mathcal{I}(LCA(x,y))}\right) \tag{3.9}$$

3.1.2.3 Mesures par approche informationnelle

L'approche informationnelle s'appuie sur la théorie de l'information formulée par Claude Shannon dans [211]. Comme pour l'approche par traits, les mesures sémantiques informationnelles reposent sur la comparaison de deux concepts en fonction de leurs points communs et de leurs différences, ici considérés en termes d'information. Cette approche introduit formellement la notion de saillance des concepts par la définition du *contenu informationnel* (Information Content ou IC). Pour rappel, l'IC d'un

concept x est défini comme le logarithme de la probabilité d'apparition d'une de ces instances [8], soit :

$$IC(x) = -\log\left(\frac{\mathcal{I}(x)}{|I|}\right) \tag{3.10}$$

Mesure informative de Resnik Resnik définit dans [196] la similarité de deux concepts à partir de l'IC du concept le plus profond qui les subsume :

$$sim_{resnikIC}(x, y) = IC(LCA(x, y)) \tag{3.11}$$

Mesure de Lin Lin et al. dans [142] propose, à la manière de Wu et Palmer, de pondérer la mesure informative de Resnik par l'IC des concepts x et y :

$$sim_{lin}(x, y) = \frac{2 \times IC(LCA(x, y))}{IC(x) + IC(y)} \tag{3.12}$$

Mesure de Cross Une mesure informationnelle proche de la définition de la mesure de Jaccard est donnée par Cross et al. dans [51]. Les auteurs proposent de caractériser l'information portée par un concept en additionnant les IC de ses ancêtres.

$$sim_{cross}(x, y) = \frac{\sum_{c \in \Gamma(x) \cap \Gamma(y)} IC(c)}{\sum_{c \in \Gamma(x) \cup \Gamma(y)} IC(c)} \tag{3.13}$$

La mesure de Cross peut également être considérée comme une stratégie hybride entre une approche par traits et la théorie de l'information.

Conclusion sur les mesures de similarité sémantique

De toutes les mesures étudiées, il est difficile de trancher sur une approche ou mesure qui supplanterait les autres notamment car l'utilisation d'une mesure, comme d'une ontologie, est rattachée à un contexte précis. Les tables 3.1 et 3.2 fournissent une vue synthétique des mesures étudiées. Nous pouvons résumer l'utilité de chacune des trois approches étudiées de la façon suivante :

- L'*approche topologique* s'appuie sur la structure même de l'ontologie. La force de cette approche est qu'elle est très facile à mettre en oeuvre comparée aux deux car elle ne nécessite pas de disposer des instances pour calculer un score. Toutefois, la qualité des résultats dépend fortement de la granularité (i.e., niveau de détail) avec lequel la taxonomie est construite, si celle-ci n'est pas uniforme dans toute la hiérarchie, les liens n'ont pas la même intensité ce qui peut conduire à fausser les résultats. Ainsi, ce type d'approche est applicable

8. La version normalisée de IC est : $IC(x) = 1 - \log_2\left(1 + \frac{\mathcal{I}(x)}{|I|}\right)$

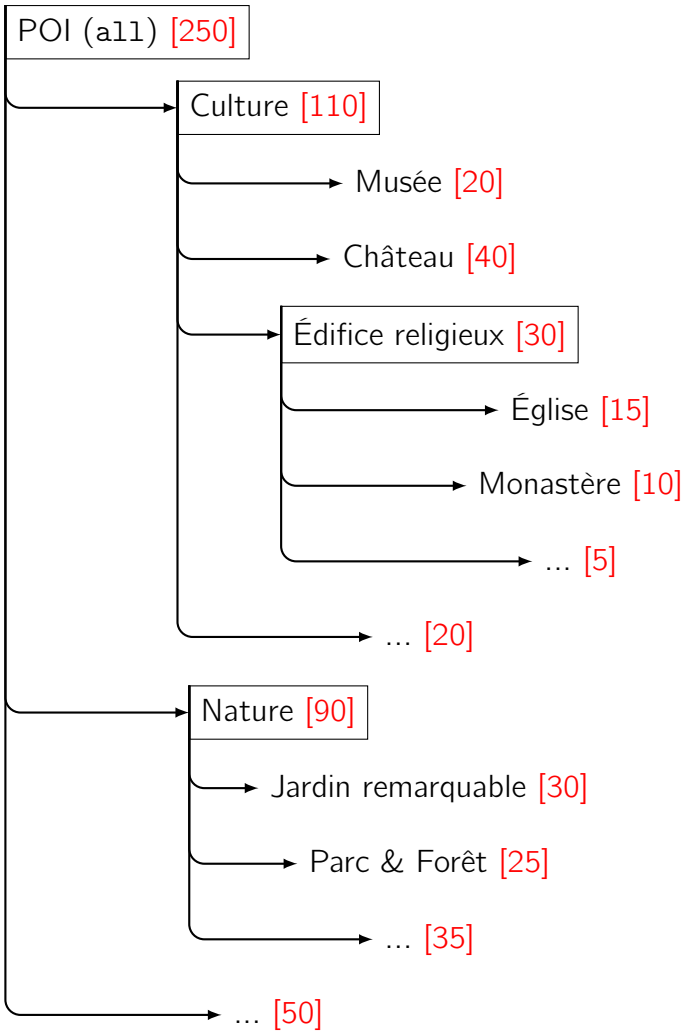


Figure 3.2 – Exemple de taxonomie de POI touristiques. L'étiquette rouge indique le nombre d'hyponymes et/ou instances du concept

dans n'importe quel contexte mais il est préférable de disposer de taxonomies profondes et détaillées pour de meilleurs résultats.

- L'*approche par traits* s'appuie sur la théorie psycho-cognitive de la Gestalt et d'une vision aristotélicienne des concepts afin d'évaluer une similarité. Plus les concepts comparés ont d'attributs en commun, plus ceux-ci seront jugés semblables. Deux visions sont possibles : l'intentionnelle (on compare les ancêtres en commun) ou extensionnelle (on compare les enfants en commun). Selon [7], cette approche est la plus à même de rendre compte de l'intuition, néanmoins elle demande une description fine des qualités des concepts pour être réellement utilisable. Également, si l'on souhaite adopter une mesure de similarité intentionnelle, alors il est indispensable de disposer des instances de chaque concept.
- Enfin, l'*approche informationnelle* mêle, en quelque sorte, les deux sortes d'ap-

proches précédentes. Les mesures détaillées s'appuient majoritairement sur une vision structurelle de la taxonomie tandis que l'IC possède intrinsèquement une visée intentionnelle par la prise en compte des instances du concept. Tout comme l'approche par traits, celle-ci exige une description détaillée de la taxonomie pour être utilisable.

La figure 3.2 présente un exemple de taxonomie de POI touristiques. La table 3.3 fournit une comparaison des scores pour l'évaluation de la similarité entre les concepts "Église, Château" et "Église, Nature" pour les mesures normalisées étudiées ci-dessus. On constate une variabilité importante des scores, notamment pour la comparaison entre les concepts "Église" et "Château". Pour les besoins des exemples sur les similarités extensionnelles, on supposera que $\mathcal{I}(\text{Église}) \cap \mathcal{I}(\text{Château}) = 5$ et $\mathcal{I}(\text{Église}) \cap \mathcal{I}(\text{Nature}) = 1$. On constate également sur cet exemple qu'aucune mesure ne possède un score proche du score moyen obtenu. Néanmoins, pour les deux exemples, la mesure de Wu-Palmer est celle dont le score semble être le plus proche de la médiane.

Dans la seconde partie II – Contributions – de la thèse, nous utiliserons la *mesure de Wu-Palmer* pour la comparaison sémantique de concepts. Ses avantages étant qu'elle soit normalisée, simple, tienne compte des liens de parenté entre les concepts et ne nécessite pas de disposer des instances. Également, sur les tests menés et de façon très empirique, ce score nous a semblé donner les meilleurs résultats en termes d'adéquation générale sur les autres mesures (voir table 3.3) mais aussi dans la restitution du ressenti humain.

Approche	**Nom**	**Type**	**Co-domaine**	**Utilise**				
				Plus grand ancêtre commun $LCA(x, y)$	Hyperonyme $\Gamma(x)$	Hyponyme $\mathcal{I}(x)$	Plus court chemin $d(x, y)$	Profondeur $d(x)$
Topologique	sim_{rada}	sim	$[0, 1]$				×	
	sim_{resnik}	sim	$\mathbb{R}^+$				×	×
	sim_{LC}	sim	$[0, 1]$				×	×
	sim_{wup}	sim	$[0, 1]$	×				×
	sim_{li}	sim	$[0, 1]$	×			×	
Par traits	$sim_{JaccardInt}$	sim	$[0, 1]$		×			
	$sim_{jaccardExt}$	sim	$[0, 1]$			×		
	$sim_{tversky}$	sim	$[0, 1]$	×				
	sim_{amato}	sim	$[0, 1]$	×		×		
Informationnelle	$sim_{resnikIC}$	sim	$\mathbb{R}^+, [0, 1]$	×				
	sim_{lin}	sim	$[0, 1]$	×				×
	sim_{cross}	sim	$[0, 1]$		×			
Statistique	NGD	dist.	$\mathbb{R}^+$		×	×		
	LSA	sim	$[0, 1]$		×	×		

Table 3.1 – Résumé synthétique des mesures de similarité sémantiques étudiées – 1

Approche	Nom	Description et commentaires
Topologique	sim_{rada}	Inverse du plus court chemin entre x et y.
	sim_{resnik}	Ajoute la notion de profondeur maximale de la taxonomie à la mesure de Rada.
	sim_{LC}	Ajoute un log afin de pénaliser davantage les valeurs importantes de plus court chemin.
	sim_{wup}	Prise en compte du plus grand ancêtre commun des concepts (le plus spécifique qui les subsume) et de la profondeur des concepts dans la taxonomie.
	sim_{li}	Similaire à Wu - Palmer mais utilise des fonctions non linéaires. Paramètres α, β contribuant respectivement au plus court chemin et au degré de parenté entre les concepts.
Par traits	$sim_{jaccardInt}$	Ratio du nombre de concepts ancêtres en commun dans la taxonomie.
	$sim_{jaccardExt}$	Ratio de nombre instances en commun dans la taxonomie.
	$sim_{tversky}$	Généralisation du modèle ensembliste de Jaccard. Paramètres α, β contribuant respectivement aux concepts uniques de y et x.
	sim_{amato}	Estime que deux concepts peuvent être similaires sans avoir aucune instances en commun.
Informationnelle	$sim_{resnikIC}$	IC du plus grand ancêtre commun. Ne respecte pas l'identité des indiscernables
	sim_{lin}	Équivalent de Wu - Palmer basé sur l'IC.
	sim_{cross}	Équivalent de Jaccard basé sur la somme des IC.
Statistique	NGD	Basée sur la fréquence des termes relatifs à x et y retourner par le moteur de recherche Google. Ne respecte pas les axiomes de la métrique.
	LSA	Utilisée avec la similarité du cosinus. Basée sur un ensemble de corpus pour représenter les concepts à partir des documents. S'appuie sur la notion de co-occurence de termes.

Table 3.2 – Résumé synthétique des mesures de similarité sémantiques étudiées – 2

	Nom	**Comparaison**			
		Église - Château		Église - Nature	
	$LCA(x,y)$ / $d(LCA(x,y))$	Culture / 1		All / 0	
	$\lvert\Gamma(x,y)\rvert$	4		0	
	$\lvert\mathcal{I}(x,y)\rvert$	5		1	
	$d(x,y)$	3		4	
	$d(x)$	2		2	
	$d(y)$	3		1	
Approche		**Détails des calculs**	**Score**	**Détails des calculs**	**Score**
Topologique	sim_{rada}	$\frac{1}{1+3}$	0.25	$\frac{1}{1+4}$	0.2
	sim_{LC}	$\log_2\left(2-\frac{3}{6}\right)$	0.58	$\log_2\left(2-\frac{4}{6}\right)$	0.42
	sim_{wup}	$\frac{2\times1}{3+2}$	0.4	$\frac{d(\text{POI})}{3+2}$	0
	sim_{li}	$e^{-0.2\times\frac{1}{4}}\times\tanh(0.6\times1)$	0.51	$e^{-0.2\times\frac{1}{5}}\times\tanh(0.6\times0)$	0
Par traits	$sim_{jaccardInt}$	$\frac{\lvert\{\text{Culture}\}\rvert}{\lvert\{\text{Eglise, Édifice religieux, Château, Culture}\}\rvert}$	0.25	$\frac{\lvert\varnothing\rvert}{\lvert\{\text{Église, Édifice religieux, Nature}\}\rvert}$	0
	$sim_{jaccardExt}$	$\frac{5}{15+40-5}$	0.1	$\frac{1}{15+90-1}$	$1/104$
	sim_{amato}	$\frac{15}{110}\times\left(1-\frac{110}{250}\right)\times\left(1-\frac{15}{110}\right)$	0.07	$\frac{15}{250}\times\left(1-\frac{250}{250}\right)\times\left(1-\frac{15}{250}\right)$	0
Informationnelle	$sim_{resnikIC}$	$1-\log_2\left(1+\frac{110}{250}\right)$	0.47	$1-\log_2\left(1+\frac{250}{250}\right)$	0
	sim_{lin}	$\frac{2\times\log_2\left(\frac{110}{250}\right)}{\log_2\left(\frac{15}{250}\right)+\log_2\left(\frac{40}{250}\right)}$	0.56	$\frac{2\times\log_2\left(\frac{250}{250}\right)}{\log_2\left(\frac{15}{250}\right)+\log_2\left(\frac{90}{250}\right)}$	0
	sim_{cross}	$\frac{0.47}{0.92+0.84+0.79+0.47}$	0.16	$\frac{0}{4.06+3.06+1.18+1.47}$	0
	Score Moyen		**0.34**		**0.06**
	Écart-type		**0.19**		**0.13**
	Médiane		**0.33**		**0**

Table 3.3 – Comparaison de concepts selon les mesures de similarité sémantique étudiées

3.1.2.4 Mesures entre ensembles de concepts

Une question subsidiaire se pose lorsque l'on ne cherche plus à comparer seulement deux concepts mais des *ensembles de concepts*. Dans ce cas, la mesure de similarité possède la signature $sim_{set} : \mathcal{P}(\Sigma) \times \mathcal{P}(\Sigma) \rightarrow \mathbb{R}^+$.

De telles mesures se basent en général sur des fonctions d'agrégation (min, max, moyenne) ou des stratégies issues de la théorie de la décision. Pour la fin de la section, on considère une mesure de similarité sim parmi celles décrites dans la table 3.1.

Agrégation par la moyenne Une approche commune est d'effectuer simplement la moyenne des similarités pour chacun des éléments des deux ensembles.

$$sim_{set-agg}(X, Y) = \frac{\sum_{x \in X} \sum_{y \in Y} sim(x, y)}{|X| \times |Y|} \tag{3.14}$$

Néanmoins, cette approche peut se montrer pénalisante lorsque les ensembles ont un cardinal déséquilibré.

Mesure de Hausdorff La distance de Hausdorff (notée d_H) [99] définit une application entre parties d'ensembles. Habituellement utilisée pour comparer des images [105], nous pouvons l'utiliser dans le cadre d'ensembles de concepts comme suit :

$$d_H(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} \{1 - sim(x, y)\},\ \max_{y \in Y} \min_{x \in X} \{1 - sim(x, y)\} \right\} \tag{3.15}$$

L'inconvénient de la distance de Hausdorff, comme la majorité des mesures basées sur les fonctions min/max, est qu'elle est sensible aux valeurs aberrantes.

Mesure de Halkidi Un compromis entre les deux précédentes mesures est la similarité définie par Halkidi et al. (notée ζ) dans [94]. Celle-ci moyenne par le cardinal de chaque ensemble, puis globalement la similarité maximale entre X et Y.

$$\zeta(X, Y) = \frac{1}{2} \left(\frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} \{sim(x, y)\} + \frac{1}{|Y|} \sum_{y \in Y} \max_{x \in X} \{sim(x, y)\} \right) \tag{3.16}$$

Selon les auteurs, cette mesure donne des résultats expérimentaux proches du ressenti humain. De plus, comme elle est basée sur une combinaison de moyennes, elle est plus robuste, moins sensible aux données aberrantes (*outliers*) que les deux précédentes mesures.

3.2 Propriétés universelles de la mobilité et des habitudes humaines

Rarement envisagée sous sa dimension sémantique, la mobilité a pourtant été extrêmement étudiée depuis la perspective spatio-temporelle afin de cerner au mieux les grands principes qui guident les déplacements des humains. Pour autant, nous avons noté dans le chapitre 2 que les dimensions sémantique, spatiale et temporelle sont loin de s'ignorer mutuellement. Dès lors, un approfondissement des notions communes à ce sujet semble indispensable pour décoder en partie les propriétés structurelles intrinsèques des séquences de mobilité sémantique.

De nombreuses études sur la mobilité humaine ont montré une remarquable hétérogénéité locale dans la mobilité qui pourtant coexiste avec un haut degré de prédictibilité globale [8]. En outre, les individus présentent un large spectre de comportements de mobilité tout en répétant des activités aux horaires quotidiens dictés par la routine. González et al. ont montré que les individus sont caractérisés par une distance de déplacement caractéristique indépendante du temps et que ceux-ci ont une probabilité significative de retourner à quelques endroits fréquemment visités [85]. Les auteurs ont notamment souligné, en accord avec Brockmann et al. [34], que la distance de déplacement caractéristique (radius of gyration) r_g des individus suit une distribution correspondant à une loi puissance d'équation $P(r_g) = (r_g + r_g^0)^{\beta_r} \exp(-r_g/\kappa)$ (avec $\beta_r = 1.65 \pm 0.15, r_g^0 = 5.8, \kappa = 350$) visible sur la figure 3.3 **(a)** mettant en avant que l'immense partie des déplacements sont effectués sur de courtes distances.

Dans la même lignée d'analyse, Song et al. s'intéressent dans [219] à la fréquentation des lieux par les individus. En outre, les auteurs montrent que le nombre de lieux distincts visités par un individu au cours du temps $\delta(t)$ (voir graphique **(b)**), peut être approximé par l'équation $\delta(t) \sim t^{\mu}$ (avec $\mu = 0.6 \pm 0.02$ et t en heure) indiquant par exemple sur la figure que $\delta(\textit{une semaine}) \approx 5.5$ soit qu'un individu visite entre 5 et 6 lieux différents par semaine. Par ailleurs, le fait que $\mu < 1$ indique bien un ralentissement sur de grandes échelles de temps, une tendance décroissante de l'individu à visiter des lieux précédemment non visités ; de plus les auteurs notent que l'équation est indépendante du radius of gyration r_g ce qui traduit le fait que même des individus parcourant de grandes distances ont tendance à fréquenter peu de lieux nouveaux. Un phénomène analogue, déjà relevé par González et al., est que les individus ont une inclination très forte à retourner aux mêmes endroits visités qui se traduit par une fréquence de visite caractérisée par un phénomène Zipfien. Le graphique **(c)** montre que la fréquence f_k du k-ème lieu le plus visité suit le modèle $f_k \sim k^{\xi}$ (avec $\xi = -1.2 \pm 0.1$). Récemment, Schlapter et al. dans [205] ont déterminé que la loi qui régit le nombre de visiteurs ρ_i d'un lieu i dépend à la fois de la distance au domicile r et de la fréquentation f (au sens donné par Song et al. [220]) et peut être modélisée très simplement par l'équation $\rho_i = \frac{\mu_i}{(r \times f)^{\eta}}$ (avec μ_i, une constante propre à chaque lieu et représentant l'attractivité, et $\eta \approx 2$).

Ces résultats, tous en accord, montrent que le "radius of giration" r_g dépend intrinsèquement à la fois de la distance mutuelle des lieux visités et du nombre total de visites de chaque lieu. En outre, un individu qui passe une majorité du temps dans ses lieux les plus fréquentés, par exemple son domicile et son travail, aura un r_g d'autant plus grand si ces deux lieux sont assez éloignés l'un de l'autre, comme c'est le cas par exemple pour les populations rurales. En conséquence, la notion de distance géographique, comme le r_g, semble peu adaptée pour caractériser la diversité de la mobilité sémantique d'un individu car trop dépendante du contexte environnemental (citadin vs urbain).

Afin de contrecarrer cet effet, Papparlardo et al. proposent de calculer le radius of giration $r_g^{(k)}$ uniquement en se basant sur les k lieux les plus fréquentés par l'individu [181]. Grâce à cette mesure, les auteurs ont découvert deux types de comportements de mobilité : les *returners* et les *explorers*. La figure **(d)** montre les graphes de mobilité des returners et des explorers pour $k = 2$. Les noeuds indiquent les lieux géographiques géographiques visités par l'individu, et chaque lien indique un voyage observé entre deux endroits. Lorsque le r_g total est faible, les deux lieux les plus importants (resp. en rouge et bleu) sont proches l'un de l'autre pour les explorers et les returners. Lorsque le $r_g \to \infty$, le comportement des returners et explorers se distingue : les noeuds rouge et bleu et le centre de gravité (croix grise) restent proches pour les explorers tandis qu'ils se séparent nettement pour les returners dont le r_g est dominé majoritairement par leurs deux emplacements préférés (typiquement le travail et le domicile), on a alors $r_g^{(2)} \approx r_g$. À l'inverse, si $r_g^{(2)} \ll r_g$, alors les deux lieux les plus fréquentés n'offrent pas une caractérisation précise des habitudes de déplacement de l'individu.

Si l'on observe la topologie des graphes de la figure **(d)**, on remarque que ceux-ci sont largement marqués par un mouvement pendulaire entre deux lieux (2-cycle). Ce fait a inspiré Schneider et al. qui, dans [206], déterminent les motifs topologiques de mobilité des individus appelés *daily patterns*. Le graphique **(e)** montre les 17 motifs découverts par les auteurs composent 90 % de la mobilité [9] et qui viennent corroborer les précédentes découvertes que nous avons abordées : les motifs sont extrêmement simples (moins de 7 noeuds) et majoritairement caractérisés par des mouvements cycliques. Les noeuds rouges marquent le noeud central (avec le plus d'entrées) du motif. Également, les auteurs proposent une classification en 4 classes des motifs de taille N en fonction de leurs propriétés topologique : (I) Les motifs avec un 2-cycle et cycle de composé $(N-1)$-cycle. (II) Les motifs composés uniquement d'un N-cycle. (III) Les motifs avec deux 2-cycles et un $(N-2)$-cycle. (IV) Les motifs avec un 3-cycle et un $(N-2)$-cycle.

Les graphiques précédents font abstraction de la dimension temporelle, pourtant celle-ci a été abondamment commentée et étudiée dans le domaine de la mobilité et des

9. Les différentes barres colorées montrent les résultats des données réelles issues de sondage de mobilité (cyan Paris ; bleu Chicago), de géolocalisation par antenne téléphonique (orange Paris) et les résultats selon le modèle des auteurs (vert clair Paris ; vert foncé Chicago).

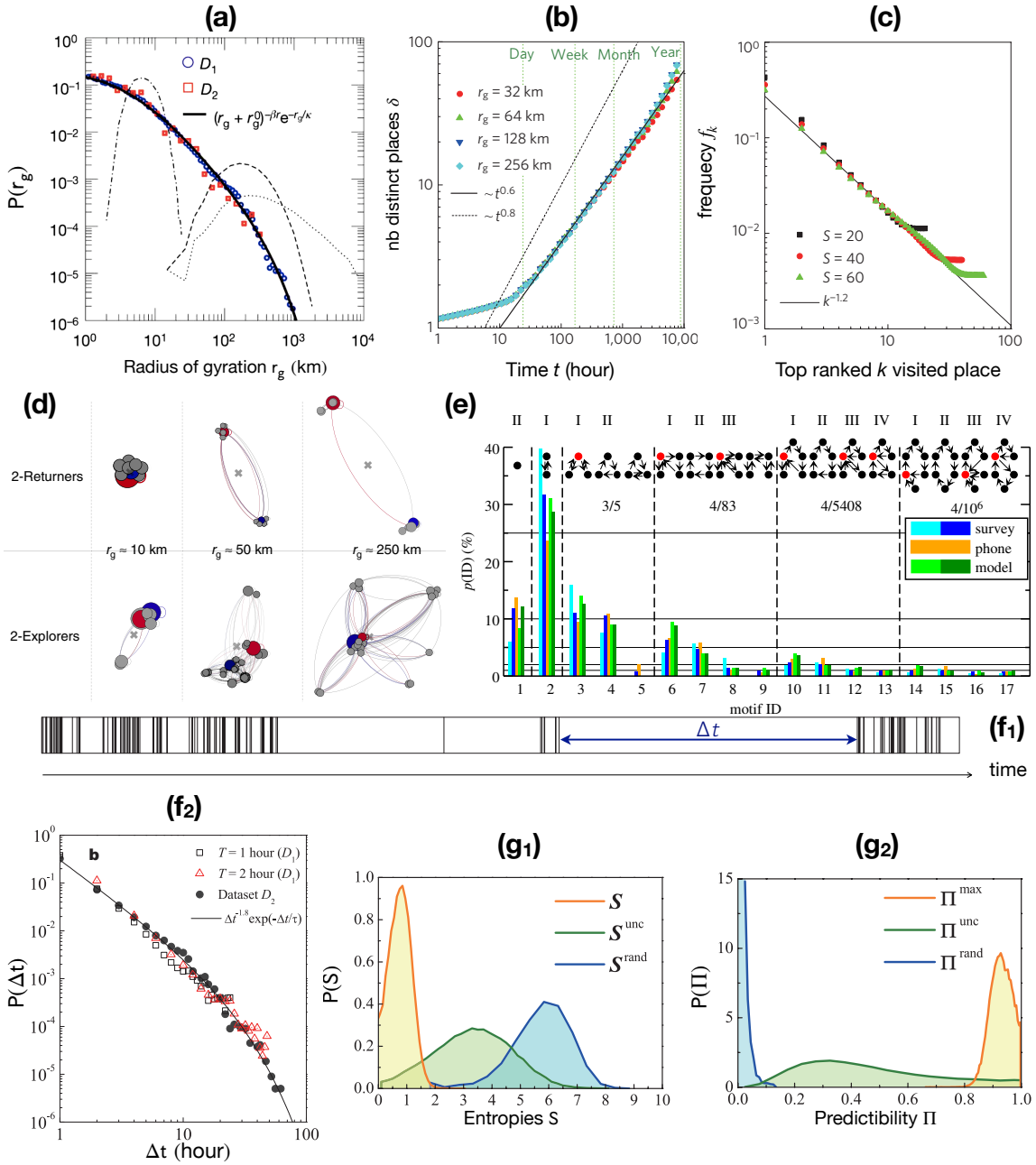


Figure 3.3 – Ensemble des graphiques résumant les propriétés universelles de la mobilité sémantique **(a)** Distribution de la distance caractéristique r_g parcourue selon une loi puissance [85] **(b)** Nombre de lieux différents $\delta(t)$ visités au cours du temps [219] **(c)** La fréquence des lieux visités suit une loi de Zipf [219] **(d)** La mobilité se scinde en deux groupes : les returners et les explorers [181] **(e)** Motifs topologiques fréquents représentant la mobilité [181] **(f_1)** Représentation des activités au cours du temps telle que l'écart temporel inter-activités Δt suit une distribution puissance [16] **(f_2)** Distribution de l'écart temporel inter-activités Δt selon une loi puissance [219] **(g_1)** Distribution de densité des entropies randomisée S^{rand}, non corrélée S^{unc} et réelle S [220] **(g_2)** Distribution de densité des prédictibilités associées aux entropies de (g_1)

activités humaines. [16] établit que la durée des activités humaines Δt suit une loi puissance de la forme $P(\Delta t) = \Delta t^{-1-\beta_t} \exp(-\Delta t/\tau)$ (avec $\beta_t = 0.8 \pm 0.1, \tau = 17$) décrite figure **($\mathbf{f}_2$)**. La figure **($\mathbf{f}_1$)** montre une telle distribution au cours du temps où chaque trait correspondant au début d'une activité. En outre, cette distribution retranscrit le fait que le rythme des activités humaines est caractérisé par des rafales d'événements rapides séparés par de longues périodes de la même activité (par exemple, travailler ou rester à la maison) et que le timing de ces activités est flou – celles-ci se produisent dans une fenêtre de temps imprécise qui dépend partiellement des précédentes activités accomplies. Nous noterons toutefois que cette distribution, à vocation universelle, demande à être nuancée et que la durée d'une activité est extrêmement conditionnée par sa nature (activité de loisir, travail, etc.) mais aussi par des facteurs socio-démographiques de l'individu qui l'effectue. C'est notamment le rôle des Time-Use survey d'éclairer ce point et nous renvoyons à [70] pour une étude exhaustive sur l'allocation du temps au sein des populations [10].

Enfin, nous noterons le travail de Song et al. sur les limites du caractère prédictible de la mobilité des individus [220]. Les auteurs ont extrait une base de données de 45.000 séquences de mobilité d'une durée de 14 semaines avec une fréquence d'enregistrement de 1 lieu par heure. Ils ont ensuite calculé l'entropie de ces séquences telle qu'énoncée par Shannon [211] afin de connaître leur quantité d'information intrinsèque. Plus celle-ci est faible, plus cela traduit un caractère prédictible de la séquence. Ainsi, les auteurs ont calculé trois niveaux différents d'entropie correspondants à trois degrés de raffinement de l'information : l'entropie randomisée S^{rand} où chaque lieu a une probabilité uniforme d'être visité, l'entropie non corrélée S^{unc} où la probabilité de visite est pondérée par sa fréquence d'apparition totale et l'entropie réelle S basée sur l'algorithme de compression de Lempel-Ziv [123] qui calcule le degré d'information d'une séquence en se basant $\forall i \in \{1, ..., n\}$ sur la sous-séquence de lieux $l_1, ..., l_i$ visités. Les figures **($\mathbf{g}_1$)** et **($\mathbf{g}_2$)** montrent respectivement les distributions d'entropies et de prédictibilité Π obtenues. En utilisant l'entropie réelle S, les auteurs ont déterminé une predictibilité record du futur emplacement à hauteur de 93% comparée à une prédictibilité d'environ 30% pour l'entropie non corrélée ce qui démontre qu'une part importante de la prédictibilité est encodée dans l'ordre temporel et la durée de visite des lieux.

Précisions ici que nous sommes focalisés sur les propriétés qui concernent *in fine* la mobilité sous sa forme symbolique, c'est-à-dire telle que représentée par une séquence de lieux / activités. Pour les lecteurs en quête d'exhaustivité sur la nature de la mobilité humaine, nous renvoyons à l'article [17] de Barbosa et al.

Outre le thème de la mobilité, l'exigence de répétition, de cohérence/homogénéité sémantique forme un verrou central et plus global pour la comparaison de séquences d'activités humaines. En tant que proverbiales créatures d'habitudes, les individus ont tendance à répéter des comportements similaires dans des contextes récurrents

10. Voir aussi : `https://ourworldindata.org/time-use`

ce qui tient à la fois de propriétés cognitives, motivationnelles et neurobiologiques. Étudié très tôt par le sociologue Pierre Bourdieu où dans [32] il créait la notion d'*habitus* qui désigne l'inclination naturelle à un style de vie, des goûts et préférences cohérents aux yeux du monde qui sont des produits de la socialisation et générateur de nouvelles pratiques sémantiquement cohérentes d'un point de vue social. De fait, l'habitus assure un déterminisme, une cohésion et une cohérence au sein de l'individu singulier et entre les individus de même groupe social en définissant la matrice des comportements individuels. Toutefois, Bourdieu insiste sur le fait que l'habitus n'est pas un phénomène de reproduction d'un comportement inculqué (car celui-ci peut évoluer dans le temps) mais plutôt un mécanisme fondamental, au coeur de la reproduction des pratiques sociales.

La pratique de l'habitude dans son sens commun, c'est-à-dire comme une pratique quotidienne répétée, est expliquée par les psychologues Wood et Rünger dans [241] où ils expliquent que les habitudes et routines naissent, certes d'exigences sociales (travail, sociabilité, etc.), mais plus généralement lorsque les individus poursuivent des objectifs en répétant les mêmes réponses dans un contexte donné. Ainsi, l'habitude se matérialise comme une réponse stéréotypée efficace à la poursuite d'objectifs similaires. Enfin, les individus ont tendance à inférer de la fréquence d'exécution des habitudes et de l'approbation (individuelle ou venant d'autrui) que leur comportement est positif, un cycle de renforcement s'opère par cet encouragement. Enfin l'habitude est également envisagée comme une alternative rassurante pour le sujet face à l'inconnu. En outre, l'application de changement dans la routine comportementale est souvent effectuée avec réticence et suppose un effort psychique [242], voire une psychosomatisation, de l'individu qui peut parfois percevoir l'hypothèse d'un changement comme angoissant [245]. Pour de plus amples détails sur le domaine de la psychologie de l'habitude, nous recommandons la lecture des travaux de la psychologue Wendy Wood, notamment de l'article [178] qui apporte un éclairage psycho-sociale sur le caractère prédictible de l'humain.

Ainsi, si nous cherchons à résumer les informations précédentes nous savons apprenons que l'apparente complexité de la mobilité humaine masque un haut degré de prédictibilité de celle-ci [220, 8] dicté principalement par des habitudes psycho-sociales [178]. En outre par le fait que les individus ont une forte tendance à retourner aux mêmes endroits [85], qu'ils répètent les mêmes activités (e.g., fréquentent les mêmes lieux) [219, 205] ce qui forme une périodicité des observations [219, 206, 43, 181]. Que les emplois du temps sont entre-coupés de quelques activités très longues mais majoritairement constitués d'activités courtes [16] ce qui conduit les activités à se tenir dans une fenêtre temporelle imprécise au sens présenté section 2.1. Enfin, dans une visée plus psychologique, cette répétition d'activités peut être également vue comme une conséquence de contextes stables qui facilitent cette propension naturelle à effectuer des comportements répétés avec une surveillance cognitive minimale, et donc l'adoption de comportements quotidiens stéréotypés [240]. La cohérence de

la vie quotidienne crée des habitudes, ou des dispositions comportementales à répéter des actions bien pratiquées dans des circonstances récurrentes.

Compte tenu de ces éléments, il est important que ces propriétés fondatrices qui signent le caractère essentiel de mobilité et de l'humain soient prises en compte dans le processus de comparaison de deux séquences de mobilité sémantique. Plus précisément, nous pensons qu'une mesure pour la comparaison de telles séquences doit pouvoir tenir compte des exigences globales suivantes :

1. Saisir le contexte, c'est-à-dire de détecter des activités sémantiquement similaires (au sens d'une ontologie métier) dans une période de temps imprécise donnée.
2. Permettre la répétition d'éléments selon une proximité temporelle et sémantique floue.
3. Permettre la permutation locale d'éléments qui traduit une notion d'homogénéité sémantique – deux séquences doivent pouvoir être semblables si elles sont composées des mêmes éléments et/ou d'éléments sémantiquement similaires mais dans un ordre différent.
4. Être robuste aux déformations temporelles légères – activités qui durent plus ou moins longtemps.

Dans la suite de cette section, nous passons en revue les différentes mesures existantes au sein de la littérature pour la comparaison de séquences de données symboliques tout en relevant celles aux propriétés adéquates pour s'appliquer au contexte de la mobilité.

3.3 Comparaison de séquences : Le temps

Fort des spécificités étudiées précédemment, nous étudions dans cette section les différentes mesures établies pour la comparaison de séquences et séries temporelles qualitatives. Une première sous-section présente les mesures les plus couramment utilisées pour la comparaison de séquences symboliques, puis nous abordons un ensemble de mesures plus spécifique à la comparaison de séquences dans les domaines de la mobilité et des sciences humaines. Nous concluons cette section en dressant un tableau récapitulatif des mesures commentées en accord avec les spécificités soulevées en section précédente.

3.3.1 Dissimilarités classiques entre séquences symboliques

Dans le chapitre 2, nous avons présenté différentes abstractions de la représentation des activités humaines au cours du temps. Nous avons vu que celles-ci prennent, de manière générale, la forme d'une séquence d'éléments $S = \langle x_1, ..., x_n \rangle$ où $x_i \in \Sigma$. On désigne alors Σ sous le nom d'alphabet et on note $X \in \Sigma^*$ une séquence de taille quelconque issue de Σ. Pour toute séquence $S \in \Sigma^n$, on note $|S| = n$ sa taille.

Dans la suite de cette section, on considère deux séquences $S_1 = \langle x_1, ..., x_n \rangle \in \Sigma^n$ et $S_2 = \langle y_1, ..., y_m \rangle \in \Sigma^m$. Dès lors, on cherche une distance $d : \Sigma^n \times \Sigma^m \to \mathbb{R}^+$ qui compare S_1 et S_2.

Lorsque $\Sigma = \mathbb{R}$, c'est-à-dire que les séquences sont composées de nombres réels, ces dernières sont qualifiées de *séries temporelles*. De nombreux travaux ont étudié les dissimilarités et similarités portant sur les séries temporelles notamment afin de prédire les tendances futures et regrouper des scénarios d'évolution similaires [5, 73, 4].

Dans le cas où Σ n'est pas un ensemble de nombres mais un ensemble fini de symboles, alors l'étude des séquences devient analogue à celui des chaînes de caractères ou des structures d'ADN [218]. Dans notre cas néanmoins, la perspective temporelle reste une composante essentielle et si l'on cherche des distances pertinentes – qui puissent être appliquées à des données qualitatives pour l'analyse de séries temporelles qualitatives. Presque toutes peuvent être identifiées comme une variation de l'un des groupes de base suivants [69] :

- Distance de Minkowski (ou distances ℓ_p)
- Distance de Hamming
- Distance d'édition (Edit distance)
- Longest Common Subsequence (LCS)
- Dynamic Time Warping (DTW)

Distance de Minkowski

La *distance de Minkowski* d'ordre p (avec $p \geqslant 0$) forme un ensemble de métriques au sein d'un espace vectoriel. Elle est définie comme une variante de la norme ℓ_p d'un vecteur telle que :

$$\|S_1 - S_2\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}, \quad |S_1| = |S_2| = n \tag{3.17}$$

Pour l'analyse de série temporelle, la *distance de Manhattan* ($p = 1$) ou *distance euclidienne* ($p = 2$) sont souvent utilisées, éventuellement combinées à la distance d'édition (voir ci-dessous) [39], comme pour l'Edit Distance on Real sequences (*EDR*) [40].

L'inconvénient de cette famille de distances est qu'elle n'opère que sur des vecteurs numériques ($\Sigma = \mathbb{R}$) et dont la taille des vecteurs comparés doit être égale. Une possibilité alors pour comparer deux séries qualitatives et d'en extraire une liste de caractéristiques (features). Néanmoins, ce type de représentation élude souvent en grande partie les dimensions sémantique et temporelle intrinsèques de la séquence originelle.

Toutefois, les ℓ_p distances sont très largement utilisées en machine learning dès que les objets à comparer peuvent être représentés dans un espace préhilbertien (c'est-à-dire un espace euclidien muni du produit scalaire) [11]. Ainsi, quelques travaux [113, 53] se sont intéressés à transférer les séquences dans de tels espaces en s'aidant de techniques de réduction de dimensionalité ou en les représentant dans un espace de caractéristiques (*features*) [67].

Distance de Hamming

La variante qualitative la plus connue de la distance de Minkowski est la *distance de Hamming* qui reprend les concepts essentiels de la distance ℓ_1. Initialement introduite pour la comparaison de séquences binaires [95], la distance de Hamming H peut être définie de façon intuitive comme le nombre de symboles à modifier pour changer la séquence S_1 en S_2. Plus formellement, celle-ci prend la forme générale :

$$H(S_1, S_2) = \sum_{i=1}^{n} \delta(x_i, y_i), \quad |S_1| = |S_2| = n \tag{3.18}$$

où $\delta : \Sigma \times \Sigma \rightarrow [0, 1]$ est une distance sur Σ (voir section 3.1.2). La distance de Hamming a l'avantage d'être très simple et rapide à calculer. Néanmoins, elle ne cherche pas un appariement optimal des séquences et est en conséquence peu robuste aux décalages et distorsions temporels.

Distance d'édition

La *distance d'édition* élargit la distance de Hamming pour tenir compte des séquences de longueurs différentes.

La *distance de Levenshtein* est l'une des distances d'édition les plus célèbres, largement utilisée pour la comparaison de chaînes de caractères ou l'analyse de séquences. La distance d'édition pratique un appariement optimale des séquences (Optimal Matching OM) et est définie intuitivement comme le nombre minimum d'opérations d'édition nécessaires pour convertir une séquence S_1 en une séquence S_2 [138]. Les trois opérations d'édition étant la **suppression**, l'**ajout** et la **modification** de symboles. Par exemple, la distance d'édition entre les mots "*chats*" et "*chiens*" est de 3. On part de "*chats*", puis on remplace successivement le *a* par *i* (chits), le *t* par *e* (chies), enfin on insère un *n* (chiens) [12].

Dans [237], Wagner et Fisher posent un formalisme de la distance d'édition. Une *opération d'édition e* est définie comme un couple $e = (a, b) \neq (\varepsilon, \varepsilon)$, où $a, b \in \Sigma \cup \{\varepsilon\}$ (on précise que ε désigne le symbole vide). L'ensemble des opérations d'édition est

11. Lorsque l'espace est de grande dimension, on préféra utiliser la similarité du cosinus telle que $\cos(X, Y) = \frac{X \cdot Y}{\|X\| \|Y\|}$

12. Notons que la distance est symétrique et que processus peut être envisagé en partant du mot *chien*.

noté E. Une opération d'édition e est alors écrite sous la forme $a \rightarrow b$ faisant ainsi référence à la transformation du symbole a en b selon le modèle des grammaires formelles. Les trois opérations précédentes sont formalisées telles que :

- Modification $a \rightarrow b$.
- Ajout $\varepsilon \rightarrow b$.
- Suppression $a \rightarrow \varepsilon$

Afin de quantifier le coût d'une opération d'édition, on peut définir une *fonction de coût* $\gamma : E \rightarrow \mathbb{R}^+$ qui attribue un coût positif à une opération d'édition.

Un *chemin d'édition* $(e_1, ..., e_N)$ de S_1 vers S_2 est alors défini comme la composition successive des opérations d'édition permettant de transformer S_1 en S_2. On note $\mathscr{P}(S_1, S_2)$ l'ensemble de ces chemins.

La distance d'édition $ED : \Sigma^n \times \Sigma^m$ est alors définie comme la somme minimale parmi les chemins d'édition de $\mathscr{P}(S_1, S_2)$. Plus formellement :

$$ED(S_1, S_2) = \min_{(e_1,...,e_N) \in \mathscr{P}(S_1,S_2)} \left\{ \sum_{i=1}^{N} \gamma(e_i) \right\} \tag{3.19}$$

La contribution principale de Wagner et Fisher est le développement d'un algorithme du calcul de la distance d'édition par programmation dynamique admettant une complexité temporelle en $O(n \times m)$.

Soit une matrice $C_{0...n,0...m}$ telle que $C_{i,j}$ représente le coût minimal afin de transformer la sous-séquence $\langle x_1, ..., x_i \rangle$ en $\langle y_1, ..., y_j \rangle$. Ces coefficients sont calculés tels que :

$$C_{i,j} = \begin{cases} i + j & \text{Si } i = 0 \text{ ou } j = 0 \\ \min \left\{ \begin{array}{c} C_{i-1,j-1} + \gamma(x_i \rightarrow y_j), \\ C_{i-1,j} + \gamma(x_i \rightarrow \varepsilon), \\ C_{i,j-1} + \gamma(\varepsilon \rightarrow y_j) \end{array} \right\} & \text{Sinon} \end{cases} \tag{3.20}$$

Enfin, $ED(S_1, S_2) = C_{n,m}$.

Concernant la fonction γ, Wagner et Fischer montrent que si (γ, E) forme un espace métrique, alors la distance d'édition ED est une métrique elle-même. En outre, les auteurs suggèrent une paramétrisation de γ telle que :

- $\gamma(a \rightarrow \varepsilon) = 1$
- $\gamma(\varepsilon \rightarrow b) = 1$
- $\gamma(a \rightarrow b) = \delta(a, b)$ où $\delta : \Sigma \times \Sigma \rightarrow [0, 1]$ est une distance sur Σ.

La définition de γ est un point très commenté dont la variation donne lieu à de nouvelles propriétés de l'Edit distance [135].

Longuest Common Subsequence – LCS

La définition de la *Longuest Common Subsequence* a été initiée à partir du concept de distance d'édition. L'intuition de LCS est de considérer que deux séquences sont similaires si elles contiennent une sous-séquence extraite (c'est-à-dire dont les symboles ne sont pas nécessairement consécutifs) suffisamment longue.

Contrairement aux autres mesures, LCS est plus robuste car elle permet à certains symboles de rester non appariés ce qui est utile lorsque les données contiennent des valeurs aberrantes. Hélas, elle ne tient pas pleinement compte d'une potentielle mesure de similarité entre les symboles et donc qu'il est difficile d'établir une fonction de coût non constante pour les opérations. Ce fait implique que la comparaison de deux séquences au contenu sémantique proche mais néanmoins différent sera fortement pénalisée. Notons que LCS est utilisée pour la comparaison entre trajectoires sémantiques dans la mesure *MSTP-Similarity* qui calcule un ratio des LCS entre les deux séquences pour estimer leur similarité [250].

D'un point de vue calculatoire, LCS partage de nombreux points communs avec la distance d'édition et peut d'ailleurs être vue comme une distance d'édition pourvue uniquement des opérations d'ajout et de suppression [237].

LCS peut elle aussi être calculée par programmation dynamique [234] selon l'équation :

$$C_{i,j} = \begin{cases} 0 & \text{Si } i = 0 \text{ ou } j = 0 \\ \begin{cases} C_{i-1,j-1} + 1 & \text{Si } x_i = y_j \\ \max\{C_{i,j-1}, C_{i-1,j}\} & \text{Sinon} \end{cases} & \text{Sinon} \end{cases} \tag{3.21}$$

Enfin, $LCS(S_1, S_2) = \max\{n, m\} - C_{n,m}$ si l'on souhaite obtenir une distance entre S_1 et S_2.

Dynamic Time Warping – DTW

La *Dynamic Time Warping* est une mesure de dissimilarité classique utilisée en traitement du signal et particulièrement en reconnaissance de la parole. Elle a été introduite pour la première fois par Berndt et Clifford [23] pour la reconnaissance de motifs dans les séries temporelles.

La caractéristique de cette mesure est qu'elle permet de dilater ou de contracter temporellement la série. Plus précisément, elle recherche un appariement optimal entre les deux séries temporelles tout en autorisant une déformation par transformation non linéaire de la variable de temps. Néanmoins, cette spécificité conduit DTW à violer les axiomes de séparabilité et d'inégalité triangulaire de la métrique.

Comme les précédentes mesures d'Optimal Matching, son calcul peut être réalisé par programmation dynamique selon l'équation :

$$C_{i,j} = \begin{cases} \infty & \text{Si } i = 0 \text{ ou } j = 0 \\ 0 & \text{Si } i = 0 \text{ et } j = 0 \\ \delta(x_i, y_i) + \min \left\{ \begin{array}{c} C_{i-1,j-1} \\ C_{i-1,j} \\ C_{i,j-1} \end{array} \right\} & \text{Sinon} \end{cases} \tag{3.22}$$

Enfin, $DTW(S_1, S_2) = C_{n,m}$.

3.3.2 Dissimilarités entre séquences symboliques spécifiques aux sciences humaines et à la mobilité

Si les précédentes mesures rendent bien compte de l'aspect temporel par la notion de précédence, elles éludent en grande partie la notion de durée telle que décrite en section 2.1. Ce fait tient en partie d'une conception discrète du temps où l'on évalue la durée quantitativement à une unité atomique temporelle (par exemple la minute, l'heure). Ainsi, pour appliquer les distances en section 3.3, la durée dans les séquences doit être d'abord discrétisée en répétant les symboles autant de fois que leur durée. Par exemple, si l'on reprend le modèle de trajectoire symbolique de Güting décrit figure 2.5 avec pour unité temporelle atomique le 1/4 d'heure, la séquence sera représentée comme ⟨🏠, 🏠, 🏠, 🚃, 🚃, 🚶⟩. Les conséquences d'un tel étalonnage est qu'*in fine*, toutes les séquences sont pourvues du même nombre de symboles, par exemple 24 si l'on choisit une durée d'une journée avec pour unité atomique de temps l'heure. Cependant, pour des séquences temporelles définies sur de larges intervalles de temps (par exemple plusieurs jours) et avec une petite unité atomique (par exemple la seconde), les temps de calcul des algorithmes peuvent devenir considérables. Pourtant, dans les domaines de la mobilité et des sciences humaines et sociales, les actions humaines sont étalées dans le temps et associées à une durée qu'il est indispensable de prendre en compte. Nous détaillons ici les quelques mesures de référence pour ces applications.

Multidimensional Similarity Measuring for Semantic Trajectories – MSM

La mesure de similarité *MSM* a été créée par Furtado et al. [79] dans l'optique de comparer les trajectoires sémantiques multi-dimensionnelles issues du modèle CONSTaNT de Bogorny et al [28].

Dans ce modèle, une trajectoire sémantique multi-dimentionnelle est représentée comme une séquence $S = \langle x_1, ..., x_n \rangle$ où $\forall i \in [\![1, n]\!], x_i \in \Sigma$ tel que $\Sigma = \times_{i=1}^{p} E_i$ et où chaque élément x_i est un vecteur de dimension p engendré par le produit cartésien des ensembles d'annotations respectifs $E_1, ..., E_p$.

Ainsi, les auteurs supposent disposer d'un ensemble de dissimilarités $D = \{d_1, ..., d_p\}$ où $d_i : E_i \times E_i \to \mathbb{R}^+$, d'un ensemble de seuils $S = \{\sigma_1, ..., \sigma_p\}$ permettant de dé-

terminer si une paire d'éléments (x_i, y_i) "matche", c'est-à-dire s'ils sont suffisamment proches / similaires et d'un ensemble de poids $\Omega = \{\omega_1, ..., \omega_p\}$ tel que $\sum_{i=1}^{p} \omega_i = 1$ requis afin d'attribuer une importance relative à chaque dimension.

Une fonction $score : \Sigma \to [0, 1]$ est définie telle que :

$$score(x, y) = \sum_{k=1}^{p} match_k(x, y) \times \omega_k \tag{3.23}$$

où $match_k : \Sigma \to \{0, 1\}$ est une fonction booléenne de la forme :

$$match_k(x, y) = \begin{cases} 1 & \text{Si } d_k(\pi_k(x), \pi_k(y)) \leqslant \sigma_k \\ 0 & \text{Sinon} \end{cases} \tag{3.24}$$

où $\pi_k : \Sigma \to E_k$ désigne l'opérateur de projection sur la k-ème coordonnée.

Grâce à la fonction de score, une fonction non-symétrique $parity : \Sigma^n \times \Sigma^m \to \mathbb{R}^+$ est ensuite définie telle que la parité de la trajectoire sémantique S_1 avec S_2 est égale à la somme des scores les plus élevés de tous les éléments $x \in S_1$ par rapport à tous les éléments de S_2, soit :

$$parity(S_1, S_2) = \sum_{i=1}^{n} \max_{j \in [\![1,m]\!]} \{score(x_i, y_j)\} \tag{3.25}$$

Finalement, la mesure de similarité $MSM : \Sigma^n \times \Sigma^m \to [0, 1]$ est définie comme la moyenne des parités :

$$MSM(S_1, S_2) = \begin{cases} 0 & \text{Si } |S_1| = 0 \text{ ou } |S_2| = 0 \\ \frac{parity(S_1,S_2)+parity(S_2,S_1)}{|S_1|+|S_2|} & \text{Sinon} \end{cases} \tag{3.26}$$

Notons que récemment, les auteurs ont proposé une version étendue de MSM (nommée SMSM) [133] permettant de prendre en compte avec plus de force les relations STOP-MOVE au sein de trajectoires sémantiques.

Dans sa conception générale, MSM pose de nombreuses questions se posent quant aux choix de modélisation des auteurs. Notamment dans l'agrégation successive de plusieurs fonctions : $match, score$ et $parity$. Le choix de leurs opérateurs respectifs n'est pas commenté par les auteurs pourtant, il est intéressant de questionner la sémantique derrière ces opérations et leurs conséquences finales sur la mesure. En outre, le choix d'effectuer un max doublé d'une somme dans la fonction $parity$ fait l'hypothèse la plus favorable quant aux matching des scores, sans considérer de notion de temporalité dans la séquence et autorise le matching avec n'importe quel élément (le matching est autorisé en début comme fin de séquence sans différence) ce qui semble conduire à une sensibilité aux valeurs aberrantes et donc une faible robustesse.

L'agrégation finale sous forme de moyenne pondérée peut sembler parcimonieuse, néanmoins, nous pourrions proposer également l'utilisation d'un min, plus pessimiste, et donc faisant plus ressortir la ressemblance entre éléments.

Une deuxième interrogation se pose sur la gestion du temps au sein de MSM. Les auteurs proposent de traiter la dimension temporelle comme une dimension à part entière ; nous interrogeons fortement ce choix quant à la symbolique qu'il porte. Le temps réclame en général d'être combiné à une autre dimension (sémantique ou spatial) pour être pleinement porteur de sens et son analyse isolée semble peu pertinente.

Enfin, MSM demande de nombreux paramètres d'initialisation, notamment ceux des dissimilarités et des seuils qui peuvent être difficiles à régler sans expertise métier. En particulier, la réduction des distances à une valeur de binaire dans la fonction $match$ semble être un point sensible de la mesure proposée.

Toutefois, parmi l'ensemble des mesures étudiées, l'approche des auteurs effectuée dans MSM est l'une des seules permettant la gestion multi-dimensionnelle au sein des trajectoires sémantiques et la prise en compte la dimension temporelle par la notion de durée, bien que celle-ci reste perfectible.

Analyse de séquences symboliques humaines : le package TraMineR

Au meilleur de nos connaissances, peu d'auteurs ont proposé des solutions alternatives pour prendre en compte la notion de durée dans les séquences autrement qu'en dupliquant les symboles de façon consécutive. Une vue pratique de cette tendance est le format d'entrée standard de la bibliothèque d'analyse de séquences TraMineR [13] sous R. Le package TraMineR [80] est dédié à l'exploration, l'analyse et la visualisation des trajectoires symboliques majoritairement issues des sciences humaines et sociales comme par exemple les données décrivant les trajectoires de vie ou la mobilité. Le TraMineR possède également une bibliothèque de dissimilarités permettant la comparaison de trajectoires symboliques ; la plupart ont été décrites dans la section précédente (Hamming, LCS, ED, DTW), nous détaillons à présent quelques unes utilisées typiquement dans le cadre de la mobilité sémantique. Nous renvoyons aux articles [80, 223] de Studer, Ritchard et Gabadinho pour une description plus exhaustive de l'ensemble des mesures disponibles dans TraMineR.

Dynamic Haming Distance – DHD La dissimilarité *Dynamic Haming Distance* imaginée par Lesnard dans [135] s'appuie sur l'hypothèse qu'à chaque moment de la journée est associé un type d'activité particulier, par exemple l'intervalle [9 :00-17 :00] représente un horaire classique de bureau. Pour Lesnard, des horaires distincts reflètent d'importantes différences de comportement. En conséquence, il propose une formule du coût de substitution δ basé sur la fréquence des transitions à l'instant i. Ainsi δ est dépendant à la fois des symboles et du temps (i.e., de l'index) i où le

13. `http://traminer.unige.ch/`

remplacement est effectué. On note X_i, la variable aléatoire qui représente le symbole à l'index i dans une séquence. Le coût δ_i est alors estimé par la série de probabilités conditionnelles décrivant les transitions entre le symbole x et y considérés entre les index $i-1$ et i, et i et $i+1$:

$$\delta_i(x,y) = \begin{cases} 4 - [P(X_i = x|X_{i-1} = y) + P(X_i = y|X_{i-1} = x) & \text{si } x \neq y \\ \quad +P(X_{i+1} = x|X_i = y) + P(X_{i+1} = y|X_i = x)] & \\ 0 & \text{sinon} \end{cases} \tag{3.27}$$

Notons que dans [80], Gabadinho et al. suggèrent de prendre uniquement en compte les deux derniers coefficients de la formule tel que :

$$\delta_i(x,y) = 2 - P(X_{i+1} = x|X_i = y) + P(X_{i+1} = y|X_i = x) \tag{3.28}$$

Cette version est actuellement implémentée dans le TraMineR. La probabilité $P(X_{i+1} = x|X_i = y)$ est estimée telle que :

$$P(X_{i+1} = x|X_i = y) = \frac{\sum_{i=1}^{n-1} \eta_{i,i+1}(x,y)}{\sum_{i=1}^{n-1} \eta_i(x)} \tag{3.29}$$

où $\eta_i(x)$ désigne le nombre de séquences tel que le symbole x est à l'indice i et $\eta_{i,i+1}(x,y)$ est le nombre de séquences tel que x est à l'indice i et y à l'indice $i+1$.

En conséquence, la vision probabiliste de DHD lui permet d'être très efficace pour la détection de comportements anormaux. Cependant, sa construction reposant sur un modèle probabiliste sans mémoire possède intrinsèquement une forte sensibilité aux décalages temporels. De plus, l'estimation des probabilités suppose l'observation d'un nombre important de transitions qui peut conduire à un risque de surapprentissage.

Optimal matching between sequences of spells – OMspell La méthode *OMspell* de Studer et Ritchard [223] modifie la fonction de coût de la distance d'édition, décrite Équation 3.19, afin d'introduire la possibilité de considérer une durée t associée à un symbole x :

$$\begin{cases} \gamma(\varepsilon \to x_t) = \gamma(x_t \to \varepsilon) = \lambda + \alpha(t-1), & \\ \gamma(x_t \to y_{t'}) = \begin{cases} \alpha|t - t'| & \text{Si } x = y \\ \delta(x,y) + \alpha(t + t' - 2) & \text{Sinon} \end{cases} & \end{cases} \tag{3.30}$$

La première ligne décrit le coût pour l'ajout ou la suppression d'un symbole x_t (symbole x durant t unité de temps). Un coût constant $\lambda \in [0,1]$ est considéré pour ces opérations. La seconde ligne évalue le coût dans le cas d'une modification. On notera également la présence d'un paramètre $\alpha \in [0,1]$ qui contrôle le coût d'extension et de compression de la séquence d'une unité de temps.

Le défaut majeur de cette méthode est qu'elle repose sur une construction d'alphabet tenant compte des durées tel que $\Sigma_t = \{x_t | x \in \Sigma, t \in [\![1, n]\!]\}$. Autrement dit, pour des séquences d'une longueur n et un alphabet initial Σ, on aura $|\Sigma_t| = |\Sigma|^n$ ce qui rend en pratique la construction de l'alphabet Σ_t impossible pour des séquences de taille importante ou des alphabets avec de nombreux symboles.

Conclusion sur les mesures entre séquences

Les tables 3.4 et 3.5 proposent un résumé des mesures étudiées dans les sections 3.3.1 et 3.3.2. Nous faisons également figurer les mesures CED et FTH qui sont deux contributions produites durant la thèse.

Parmi les mesures entre séquences symboliques étudiées au sein de la littérature, nous pointons ici plusieurs manques, en particulier dans le contexte de la mobilité et des activités humaines, qui nous convient de combler :

1. Premièrement, la prise en compte de certaines caractéristiques intrinsèques aux comportements de mobilité humaine comme la répétition de certaines activités, l'homogénéité sémantique ou le fait de pouvoir permuter localement certains éléments. Plus généralement, nous proposons la création d'une mesure qui prennent en compte la globalité du contenu de la séquence lors du processus de comparaison afin de diminuer la dissimilarité dans le cas où les séquences sont composées d'éléments similaires, au sens des mesures présentées en section 3.1.2, mais dans un ordre / timing différent.
2. Le caractère temporel de ces mesures reste très incomplet tant au niveau de la notion de précédence que de la notion de durée. Nous pensons en effet que la prise en compte d'un aspect de tolérance temporelle grâce notamment à la logique floue permettrait de rendre certaines mesures étudiées plus robustes mais aussi de rendre mieux compte de certains aspects intuitifs du temps tels qu'il a été commenté section 2.1.
3. Malgré la solution proposée par MSM, de nombreux écueils restent à pallier dans la prise en compte de l'aspect multi-dimensionnel dans les trajectoires sémantiques, notamment sur les aspects liés au temps mais également sur l'agrégation et la prise en compte d'ensembles de valeurs et mesures sémantiques.

Ainsi, nous proposons dans la partie Contributions deux nouvelles mesures : la Contextual Edit Distance (chapitre 5) et la mesure Fuzzy Temporal Hamming (chapitre 6) qui résolvent les lacunes précédentes.

Dans la suite de cette dissertation, nous proposons d'expliciter en quoi la définition des modèles de trajectoires sémantiques et les mesures appliquées à celles-ci sont importantes pour la découverte et la caractérisation de comportements de mobilité. En outre, nous détaillons le processus de découverte de tels comportements : de la définition à l'opérationnalisation et la restitution des connaissances.

Mesures		Mikowski	Hamming	Edit. dist	LCS	DTW	DHD	OMspell	MSM, SMSM	CED	FTH
Propriétés	**Métrique**	×	×	×†	×			×‡			
	Semi-métrique								×‡	×	×
	Seq. taille fixe	×	×								×
	OM			×	×	×		×		×	
	ED			×				×		×	
	Disto. temporelle					×				×	×
	Ctxt						×		×	×	×
	Permut.			×†					×	×	×
	Répetition					×				×	×
	Sim		×	×				×		×	×
	Durée							×	×		×
	Multi. dim								×	×⋆	×⋆

† Variante Damerau-Levenshtein [52]. Sacrifie l'inégalité triangulaire et autorise les transpositions adjacantes.

‡ Dépend de certains paramètres précisés par les auteurs.

⋆ Prise en compte de l'aspect mutli-dimensionnel sémantique par les auteurs dans [162].

Table 3.4 – Résumé synthétique des mesures entre séquences étudiées – 1

Mesures	**Complexité temporelle**	**Description**
Mikowski	$O(n)$	Distance selon la norme ℓ_p.
Hamming	$O(n)$	Somme des distances entre symboles à la même position.
Edit. dist	$O(n \times m)$	Somme minimale des coûts d'édition pour transformer S_1 et S_2.
LCS	$O(n \times m)$	Plus grande sous-séquence commune entre S_1 et S_2.
DTW	$O(n \times m)$	Appariemment optimal avec déformation temporelle non-linéaire des séquences.
DHD	$O(n)$	Somme des coûts dépendants de la probabilité de transition aux instants $t-1, t, t+1$.
OMspell	$O\left(\lvert\Sigma\rvert^n\right)$	Introduit la notion de durée à la distance d'édition par modification de la fonction de coût.
MSM, SMSM	$O\left(n \times m \times p\right)$	Similarité basée sur l'agrégation de fonctions de matching pour chaque dimension.
CED	$O\left(\max(n, m) \times n \times m\right)$	Distance d'édition avec coût pondéré par la position et la similarité des symboles à proximité.
FTH	$O\left(\max(n, m) T_{\max} \log(T_{\max})\right)^{\dagger}$	Distance de Hamming floue avec prise en compte des durées et de la similarité des symboles à proximité.

† $T_{\max}$ désigne la durée de l'intervalle de temps considéré et n, m respectivement le nombre de symboles dans les séquences S_1, S_2 (voir chapitre 6).

Table 3.5 – Résumé synthétique des mesures entre séquences étudiées – 2

Chapitre 4
Analyse, extraction et découverte de connaissances

4.1 Découverte de comportements

Dans le chapitre précédent, nous avons abordé en section 3.2 la notion bourdieusienne d'*habitus* pour introduire une notion relative au comportement du point de vue social. Dans cette section, nous allons plus loin et définissions dans un premier temps la notion de comportement selon le cadre de pensée de la psychologie comportementale *behavioriste*. La seconde section développe les approches pragmatiques pour d'extraction de comportements de mobilité telles que la fouille de motifs séquentiels fréquents ou le clustering.

4.1.1 Approche comportementale de la mobilité

Le terme *comportement* est usuellement défini en anthropologie et en psychologie sociale comme un ensemble de manifestations observables d'un individu – humain ou autre – en réaction à des situations dans son environnement [49]. Le courant de pensée Behavioriste, fondé en partie par le psychologue Frederick Skinner, décrit un schéma en trois phases ABC dans le processus de mise en place du comportement [217] :

$$A \rightarrow B \rightarrow C$$

où A désigne l'*antécédent*, en substance il s'agit de l'environnement avec lequel l'individu interagit et qui le stimule. En réponse à ces stimulis, l'individu produit un comportement (*behavior*) B suivit de *conséquences* C sur l'individu. Ce schéma a particulièrement été étudié dans la mise en place du conditionnement pavlovien mais aussi la mise en place de stratégies de marketing [233, 177, 195]. Un exemple simple illustrant cette notion est celui du centre commercial où l'environnement A a pour objectif de produire de nombreux stimuli destinés à susciter le désir et déclencher un comportement d'achat B. La conséquence C pour l'individu est une satisfaction "ponctuelle" de son désir.

On comprend alors mieux la citation de Miller énoncée en section 2.2.2 : la Time-Geography ne constitue pas, en elle-même, un modèle d'explication du comportement des individus mais elle forme le cadre, l'environnement et les contraintes dans lesquels les individus sont plongés et qui les conduisent à se comporter [143]. Cette vision, avant tout spinoziste et déterministe, met en avant le poids des structures et de l'environnement sur le conditionnement des comportements des individus.

Un autre exemple de l'influence de l'environnement sur l'individu est fourni par la psychologie environnementale qui cherche à étudier les interrelations entre l'individu et son environnement physique et social. C'est alors l'analyse des perceptions et émotions de l'individu, ses représentations mentales de l'environnement réelles ou fantasmées ainsi que sa façon de l'investir qui permettent de donner une signification aux comportements. La psychologie environnementale est très étudiée en urbanisme où elle démontre, par exemple, le bénéfice des espaces verts sur le moral et la réduction du stress des usagers [119], en architecture combinée à la Gestalt Theory [21] mais aussi dans le domaine de la publicité, en particulier celui du marketing sensoriel [27, 124]. Par ces quelques éléments, on comprend pourquoi et comment l'environnement agit sur nos comportements.

En informatique, l'analyse orientée vers le comportement a une portée très importante dans l'exploration de données [239] pour des applications typiques telles que la détection de comportements anormaux [75], l'analyse du comportement pour la fidélisation et la recommandation client [6] et les modèles de comportement d'exploration – basés sur des logs pour le Web ou les bases de données [185, 110]. Néanmoins, comme le note Cao dans [35], nous nous rendons compte que la soi-disant analyse comportementale n'est pas basée sur de véritables éléments comportementaux mais plutôt sur des données transactionnelles simples. Or, dans ce type de vision, le comportement est implicite et les propriétés comportementales sont cachées, dispersées dans les données. En conséquence, Cao propose une modélisation des comportements sous la forme d'un vecteur γ de 13 attributs :

- Le *sujet* (subject) : L'individu faisant l'action.
- L'*objet* (object) : L'entité vers laquelle le comportement est ciblé.
- Le *contexte* (context) : L'environnement où le comportement opère.
- Le *but* (goal) : Objectif / Intention à accomplir lors de l'action menée.
- La *croyance* (belief) : L'état informationnel et la connaissance du sujet sur le monde.
- L'*action* (action) : Action menée par le sujet.
- Le *plan* (plan) : Séquence d'actions à conduire pour atteindre un but.
- L'*impact* (impact) : Résultat conduit sur le sujet, l'objet ou l'environnement à l'issue de l'action.
- Les *contraintes* (constraints) : Ensemble de contraintes liées à l'environnement qui contraignent ou influencent le comportement du sujet.

- Le *temps* (time) : Instant où survient le comportement.
- Le *lieu* (place) : Lieu où survient le comportement.
- Le *statut* (status) : Étape du comportement parmi : passif (comportement déclenché en cours), actif (comportement initié en cours), fini.
- Les *associés* (associates) : Ensemble d'instances de comportements potentiellement associées.

La modélisation de séquences comportementales $\langle \gamma_1, ..., \gamma_n \rangle$ est également abordée par l'auteur. L'intérêt particulier de ce modèle est qu'il reprend de nombreux attributs évoqués au sein des modèles de la psychologie cognitive (e.g., contexte, impact, statut) mais aussi de la Time-Geography (e.g., contraintes, temps et lieux) et des modèles de Parent et al. [182] et Bogorny et al. [28] (e.g., but, action, associés) abordés section 2.3.

Néanmoins, bien qu'il arrive à capturer l'ensemble des attributs comportementaux principaux, le modèle de Cao cristallise du même coup le problème majeur dans la découverte de comportements : *la disponibilité de l'information*. Un fait unanime au sein des différents modèles est que la compréhension du comportement passe par un enrichissement du contexte dans lequel évolue l'individu [58]. Toutefois, et comme nous l'avons noté section 2.3, ces données ne sont pas toujours disponibles, intelligibles ou computationnables. Le second point fait référence aux attributs *but* et *croyance*. Outre le fait que ces attributs dénotant l'intention sont des notions privées et subjectives qui peuvent être difficilement renseignées autrement que par l'individu lui-même, si l'on revient à la définition du comportement, celui-ci désigne un ensemble de manifestations observables. Ainsi, en ajoutant ces notions intérieures et personnelles de l'individu, on déforme la définition même du comportement et on commet une sorte de diallèle, un raisonnement circulaire qui tente d'expliquer ici le processus du comportement par le comportement lui-même. Dès lors, nous pensons que l'analyse des comportements, d'un point de vue opérationnel, doit se limiter aux faits observés et, si possible, le contexte dans lequel ils surviennent.

Ainsi, nous admettrons dans la suite de cette dissertation que le comportement de mobilité d'un individu est exprimé (en partie) par la séquence de ses localisations / activités observées au cours du temps. Nous décrivons dans la section suivante les approches et techniques utilisées pour l'extraction et la découverte de comportements de mobilité à partir d'un ensemble de séquences de mobilité sémantique.

4.1.2 Fouille de comportements de mobilité

Comme nous avons pu le voir en section 3.3.1, la découverte de comportements dans les séquences de mobilité est étroitement liée à la recherche de motifs dans les séquences. Initialement introduit dans [3], le problème de la fouille de motifs séquentiels fréquents (Frequent Sequence Mining FSM) est défini sur une base de données de séquences $\mathcal{D}$, où chaque élément est une séquence d'éléments appelée

itemset. Le problème de la fouille de motifs séquentiels fréquents consiste alors à trouver toutes les séquences qui sont fréquentes dans $\mathcal{D}$, c'est-à-dire qui apparaissent comme sous-séquence d'un grand pourcentage de séquences de $\mathcal{D}$. On cherche ainsi à découvrir un ensemble de sous-séquences respectant un certain seuil d'apparition (support) minimal σ. Considérons deux séquences $\alpha = \langle \alpha_1, ..., \alpha_n \rangle$ et $\beta = \langle \beta_1, ..., \beta_p \rangle$ avec $n \leqslant p$. On dit que α est une sous-séquence de β (notée $\alpha \preceq \beta$) si et seulement si $\exists 1 \leqslant i_1 < ... < i_n \leqslant p$, n entiers tels que $\forall k \in [\![1, n]\!], \alpha_k \subseteq \beta_{i_k}$. Ainsi, on définit le support $supp(S)$ d'une séquence S comme le pourcentage d'itemsets $T \in \mathcal{D}$ tel que $S \preceq T$ et l'on dit que S est fréquent par rapport à σ si $supp(S) \geqslant \sigma$. Pour de plus amples détails sur les algorithmes utilisés, on renvoie à [82] qui fournit une revue de la littérature sur ce sujet. Cependant, ces techniques ne tiennent pas compte de la durée des éléments dans les séquences.

Dans [84], Gianotti et al. proposent l'extraction de *T-pattern* au sein d'un ensemble de trajectoires GPS. Un T-pattern est, en substance, une séquence de régions d'intérêt (RoI) densément fréquentées sur une période de temps donnée. Pour se faire, les auteurs incorporent la dimension temporelle au problème de la fouille de motifs séquentiels fréquents par la notion de τ-containment ($\preceq_\tau$) qui ajoute une contrainte temporelle de durée τ sur l'enchaînement de visites des RoI dans les séquences. La contrainte spatiale ($\preceq_N$) est quant à elle gérée à l'aide d'une fonction de voisinage $N : \mathbb{R}^2 \rightarrow \mathcal{P}(\mathbb{R}^2)$ qui permet de relaxer l'appartenance stricte à une région d'intérêt. Enfin, la contrainte de densité au sein des RoI est contrôlée à l'aide d'un seuil δ.

L'aspect de contrainte sémantique est rajouté par Zhang et al. dans [254] où l'algorithme SPLITTER est présenté. Selon les auteurs, un motif ne peut refléter la régularité des mouvements que lorsque les contraintes suivantes sont réunies : (i) compacité spatiale qui reflète la densité de fréquentation des RoI, (ii) consistance sémantique qui permet d'assurer une explication claire quant à la symbolique de la RoI, (iii) continuité temporelle qui garantit que le motif conserve une cohérence temporelle.

Néanmoins, les méthodes précédentes ne couvrent pas le cas des trajectoires sémantiques multi-dimensionnelles. Dans ce cadre, nous noterons l'algorithme MASTERMOVELETS [77] qui consiste à sélectionner et combiner automatiquement les dimensions des données afin de découvrir les sous-trajectoires qui permettent de mieux discriminer leurs caractéristiques. Toutefois, une lacune majeure de la méthode est sa complexité de calcul, exponentielle selon le nombre de dimensions. Un second algorithme, M^3SP, proposé par Plantevit et al. dans [188] se base sur une approche par FSM et tient compte à la fois du caractère multi-dimensionel mais aussi de la similarité des éléments organisés à l'aide d'ontologies hiérarchiques (e.g., taxonomies) ; la dimension temporelle n'est toutefois pas considérée par la méthode.

Ainsi, les méthodes FSM sont très efficaces pour représenter une abstraction agrégée de nombreuses trajectoires individuelles partageant la propriété de visiter la même séquence de lieux avec des durées similaires. Elles sont simples à mettre en place et ont également l'avantage incontestable d'être parfaitement interprétables en termes de

	Approche FSM	**Approche Clustering**
Simplicité	✓	×
Souplesse	×	✓
Explicabilité	✓	×
Globale	✓	✓
Individuelle	×	✓
Robustesse	×	✓

Table 4.1 – Table comparative synthétique des avantages et inconvénients des approches FSM et clustering

comportement. Pour autant, si ces approches apportent un éclairage macroscopique précieux sur l'ensemble du jeu de données, elles ne permettent en rien de qualifier un individu précis quant à son comportement de mobilité. De plus, comme le notent Shan et al. dans [210], ces approches souffrent souvent de trois types de problèmes : (i) *Semantic Absence*, qui les rendent peu résistantes aux données manquantes. (ii) *Semantic Bias*, les données de mobilité – particulièrement celles provenant de réseaux sociaux – souffrent souvent d'un déséquilibre lié à la sélectivité des sujets. De fait, les données récoltées peuvent ne pas être représentatives de la population et de la réalité urbaine. Par exemple, il est connu que les populations jeunes utilisent davantage les réseaux sociaux que leurs aînés ; de plus, ces utilisateurs ont une tendance à partager davantage des activités de loisir ou de sociabilité ce qui peut corrompre le résultat d'approches basées sur la fréquence. (iii) *Semantic Complexity*, lorsqu'il existe une forte diversité de labels sémantiques au sein des séquences et que le jeu de données est trop restreint, il est parfois difficile de voir émerger des motifs fréquents. L'approche par FSM reste sensible aux seuils et peine à rendre compte de la similarité entre concepts. Enfin, elle ne permet pas une comparaison inter-individu des comportements de mobilité.

Dès lors, si l'on souhaite disposer d'une vision à la fois globale et individuelle quant aux comportements à extraire, une approche orientée clustering semble plus adéquate. Le problème du *clustering* [109] (ou partionnement de données) consiste, étant donné une dissimilarité $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^+$ sur un ensemble de données $\mathcal{D}$, à déterminer une fonction d'attribution $C : \mathcal{D} \to [\![1, K]\!]$ qui affecte une classe à chaque élément de $\mathcal{D}$ sous la contrainte de minimiser une certaine fonction de coût, par exemple la variabilité intra-classe. Par exemple dans notre cas une classe représente un type de comportement de mobilité, on cherche alors à constituer des groupes d'individus et/ou de trajectoires sémantiques les plus similaires au sens de d. On dispose ainsi du même coup d'un ensemble de comportements vus de manière globale mais aussi individuelle (i.e., par instance) car chaque élément est caractérisé par sa classe affectée. À noter également que certaines méthodes de clustering permettent de mettre en lumière même les comportements plus marginaux (outliers) qui seraient indétectables par une approche FSM.

La table 4.1 fournit une comparaison synthétique entre l'approche FSM et Clustering. Concernant la *simplicité* de mise en place, si l'approche FSM peut demander quelques réglages empiriques concernant les seuils de support, l'approche par clustering est quant à elle confrontée à l'épineux problème du choix de la dissimilarité utilisée pour comparer ces éléments (d'autant plus difficile quand ceux-ci sont des séquences d'éléments complexes). On pourra préférer une dissimilarité à une autre selon la nature des séquences et sur la base de critères / exigences à respecter, par exemple pour DTW les séquences sont similaires à une déformation près. Ainsi, l'approche par clustering permet une plus grande *flexibilité* que l'approche FSM. Modifier la dissimilarité ou la technique de clustering utilisée (voir section 4.2) permet d'obtenir un partitionnement des données parfois très différent. A contrario, la marge de manoeuvre pour la découverte de motifs séquentiels est très faible. La qualité d'*explicabilité* est d'ailleurs une conséquence directe des deux précédentes. L'approche FSM plus rigide dans sa conception mais aussi plus simple permet une explication très claire quant au modèle qu'elle fournit. À l'inverse, l'approche par clustering nécessite une explication *a posteriori* des classes qu'elle a constituées car si elle permet d'affirmer que deux éléments sont dans une même classe, elle n'explique pas ce que représente la classe (e.g., le comportement) de façon intrinsèque. Concernant l'extraction de connaissance, comme souligné plus haut, les deux approches permettent une vision *globale* du jeu de données. Cependant, seule l'approche par clustering permet de caractériser les données de façon *individuelle* dans le sens où l'on affecte une classe à l'individu / instance mais aussi que l'on peut comparer les données entre elles. Enfin, l'approche par clustering permet de mieux gérer les données manquantes ou biaisées qui peuvent être gênantes dans une approche basée sur la fréquence comme FSM.

Sur la base de ces caractéristiques, nous adoptons, dans la suite de cette dissertation, le parti pris d'axer notre processus de découverte de comportements de mobilité sur une approche par clustering. La section suivante est dédiée aux différents processus de clustering adaptés pour les séquences sémantiques. Des exemples d'applications issus de la littérature dédiés et à la fouille de comportements dans les séquences de mobilité sémantique sont également abordés.

4.2 Clustering de séquences sémantiques

Comme abordé dans la section précédente, l'extraction de comportements depuis un ensemble de séquences sémantiques est un processus qui peut être réalisé à l'aide de techniques issues de l'apprentissage automatique (*machine learning*) tel que le clustering. La figure 4.1, partiellement inspirée de [203], montre les 4 grandes étapes dans un processus de clustering complet. Ces étapes sont étroitement liées les unes aux autres et affectent les clusters dérivés.

Dans le cas des séquences sémantiques, les dissimilarités usuellement employées sont détaillées section 3.3.1. Une fois la dissimilarité d sélectionnée, on applique celle-ci pour toute paire de séquences de la base de données $\mathcal{D}$. On obtient ainsi une

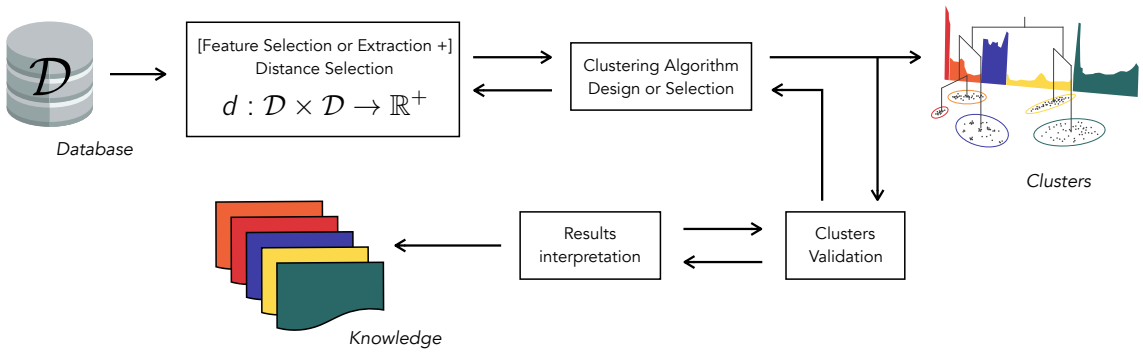


Figure 4.1 – Processus de clustering et d'extraction de connaissance : L'analyse typique des clusters se compose de quatre étapes avec une voie de rétroaction [203]

matrice de distance $D = \{d_{ij}\}$ où $d_{ij} = d(S_i, S_j)$ avec $S_i, S_j \in \mathcal{D}$ qui constitue l'entrée du processus de clustering. Cependant, les espaces topologiques créés par ces mesures sont très difficiles à appréhender, en particulier pour les méthodes d'OM (Edit Distance, DTW), car ceux-ci ne sont généralement ni euclidiens ni métriques et ne peuvent pas être visualisés.

Au meilleur de nos connaissances, il existe relativement peu d'algorithmes de clustering capables de traiter des matrices de distances arbitraires (pas seulement des métriques). Nous détaillons sommairement ci-dessous les algorithmes et techniques utilisés au cours de la thèse. La table 4.2 en reprend les propriétés importantes. Également, la figure 4.2 montre les résultats de processus de clustering selon les techniques étudiées (en colonnes) pour plusieurs des jeux de données d'exemples (en lignes) dans le plan $\mathbb{R}^2$ ce qui permet d'apprécier intuitivement leur fonctionnement. Enfin, pour de plus amples détails sur les aspects formels, nous renvoyons aux études [203, 204] et pour une vue plus technique, à la librairie Sklearn [1] de Python.

k-means et k-medoids

L'algorithme des k-moyennes (k-*means*) de MacQueen [144] consiste à initialiser k centres de gravité $\mu_{i \in [\![1,k]\!]}$ de manière aléatoire et assigne à la classe C_i, à chaque itération, tous les objets $x \in \mathcal{D}$ tel que μ_i est le plus proche. Les μ_i sont ensuite affinés en recalculant leurs coordonnées pour les $x \in C_i$. Ainsi, l'objectif de k-means est de minimiser la fonction :

$$f(\mathcal{D}, \mu) = \sum_{i=1}^{n} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4.1}$$

La convergence de l'algorithme est assurée car le centre de gravité est l'objet qui minimise l'inertie intra-cluster. Par le théorème de Huygens, cela revient aussi à maximiser l'inertie inter-classe.

1. `https://scikit-learn.org/stable/modules/clustering.html`

L'algorithme des *k-medoids* peut être vu comme une version discrète de *k*-means. Les données sont modélisées de manière très similaire, mais en utilisant k objets représentatifs réels issus du jeu de données et appelés médoïdes m_i définis tels que :

$$m_i = \text{argmin}_{x \in C_i} \left\{ \sum_{y \in C_i} d(x, y) \right\} \tag{4.2}$$

Les m_i, en tant qu'éléments non fictifs, servent de *prototypes* pour le cluster au lieu du centre de gravité et forment une base d'explication, de prototype pour l'ensemble du cluster. Cette définition permet l'utilisation de distances arbitraires et, de fait, les données n'appartiennent pas obligatoirement à $\mathbb{R}^d$. La fonction objectif devient :

$$g(\mathcal{D}, M) = \sum_{i=1}^{n} \sum_{x \in C_i} d(x, m_i) \tag{4.3}$$

On remarque que si $d(x, m) = \|x - m\|^2$, alors on retrouve la fonction objectif de l'équation 4.1 à l'exception que *k*-means est libre de choisir $\mu_i \in \mathbb{R}^d$ tandis que dans *k*-medoids $m_i \in C_i$.

Il existe plusieurs algorithmes permettant le calcul des *k*-medoids, le plus connu étant PAM (*Partitioning Around Medoids*) [120] dont de récentes optimisations par Schubert et Rousseeuw lui garantissent une complexité en $O(N^2)$ [208].

L'inconvénient de *k*-medoids est son impossibilité à traiter les clusters non-convexes et le fait qu'il faille lui spécifier le nombre de clusters en entrée. Ce défaut est visible notamment sur la figure 4.2 colonne 1, ligne 1 où les deux cercles concentriques ne sont pas retrouvés par l'algorithme.

Enfin, notons qu'il existe également des variantes graduelles (floues, possibilistes, évidentielles) des algorithmes *k*-means et *k*-medoids permettant de rendre compte d'une appartenance imparfaite à une classe donnée. Citons par exemple l'algorithme fuzzy *c*-means [63, 25] qui nécessite cependant un espace $\mathbb{R}^d$ euclidien pour être pleinement applicable et interprétable. On notera toutefois l'initiative de Hathaway et Bezdek [98] permettant de généraliser fuzzy *c*-means à toute dissimilarité (symétrique).

Spectral

Popularisé par Shi et al. et Ng et al. [212, 172], le principe du *clustering spectral* est de transformer les données $x_i \in \mathcal{D}$ en un nouvel espace de points $y_i \in \mathbb{R}^k$ via un graphe de similarité, sa matrice laplacienne associée et les k premiers vecteurs propres de cette matrice. En outre, les propriétés des matrices laplaciennes font que ce nouvel ensemble de points y_i est facilement classifiable en k groupes.

Le clustering spectral utilise donc une représentation des données sous la forme d'un graphe pondéré non orienté tel que tout $x \in \mathcal{D}$ est un noeud du graphe et où un l'arc (x_i, x_j) est valué par la similarité entre ceux-ci.

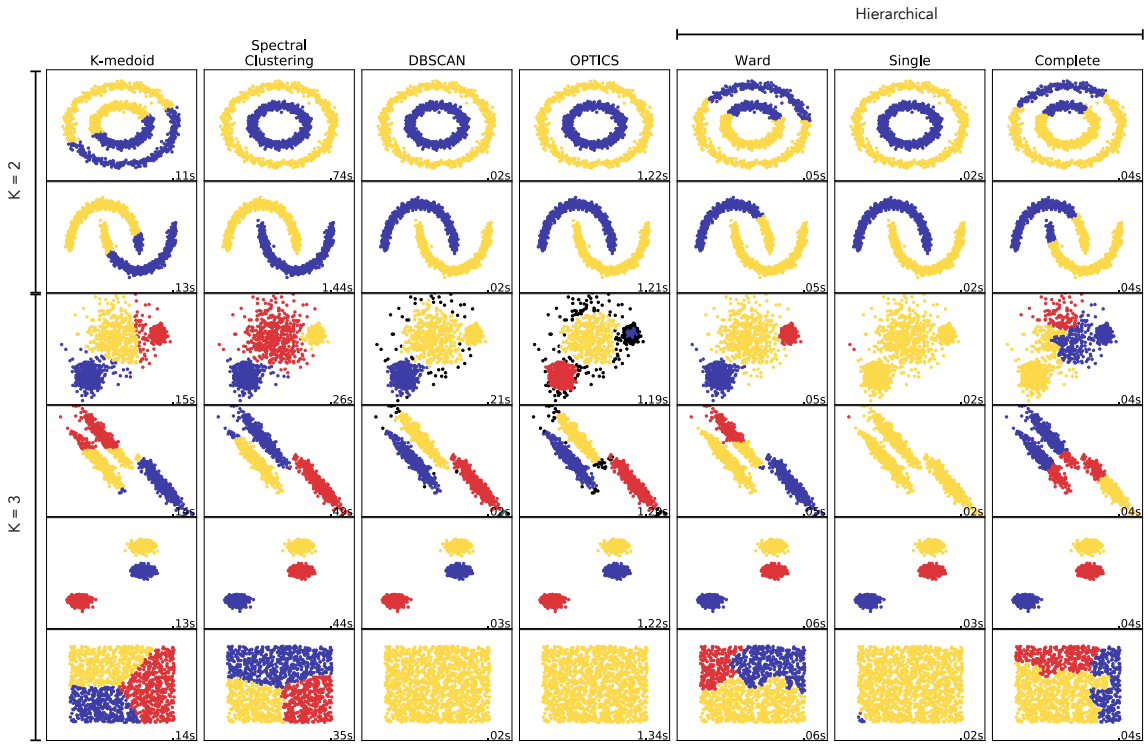


Figure 4.2 – Comparaison des différents algorithmes de clustering testés

Plus formellement, soit le graphe $G = (\mathcal{D}, A)$ des données avec $A = \{(x_i, x_j, s_{ij}) | x_i, x_j \in \mathcal{D} \text{ et } s_{ij} > \varepsilon\}$, l'ensemble des arcs dont la similarité entre x_i et x_j est supérieure à un seuil $\varepsilon \in [0, 1[$ [2]. La similarité s_{ij} est donnée par une matrice de similarité S calculée à partir de la matrice de distance D. Il existe plusieurs méthodes pour transformer D en S, une construction connue est d'avoir recours à un noyau gaussien tel que :

$$S = \{s_{ij}\} = \left\{ \exp\left(-\frac{d(x_i, x_j)^2}{2\sigma^2} \right) \right\}$$

On construit ensuite la matrice laplacienne (ou laplacien) $\mathcal{L}$ de G telle que :

$$\mathcal{L} = X - S \tag{4.4}$$

Où $X_{ii} = \sum_j S_{ij}$ est une matrice diagonale qui, pour chaque élément $x \in \mathcal{D}$, effectue la somme de ses similarités.

Notons qu'il existe de nombreuses façons différentes de définir un laplacien qui ont chacune des interprétations mathématiques différentes ; le clustering aura en conséquence des interprétations différentes. Ng et al., dans [172], fournissent plusieurs approches quant à la définition de $\mathcal{L}$. En règle générale, on définit un laplacien normalisé $\mathcal{L}_{norm}$ tel que :

2. Si $\varepsilon = 0$, dans ce cas le graphe est complet.

$$\mathcal{L}_{norm} = I - X^{-1/2} S X^{-1/2} \tag{4.5}$$

L'étape suivante consiste à extraire les vecteurs propres $\mathbf{v}_i$ (correspondant aux k plus petites valeurs propres) de $\mathcal{L}_{norm}$. La matrice Λ est alors construite en stockant ces vecteurs propres en colonnes telle que $\Lambda = [\mathbf{v}_1, ..., \mathbf{v}_k]$. Enfin, un algorithme de clustering – souvent k-means – est appliqué sur Λ afin d'extraire les individus semblables.

L'intuition du clustering spectral repose sur une bonne interprétation algébrique du spectre du Laplacien. Celui-ci est particulièrement étudié en physique statistique pour analyser la diffusion sur un graphe d'origine, par exemple la chaleur [48]. Ici, métaphoriquement, plus des données sont similaires, plus l'information se diffuse facilement au sein des zones denses du graphe. Enfin, l'analyse du spectre revient à l'extraction des k premiers vecteurs propres ce qui est très similaire à une analyse par ACP (Analyse en Composante Principale). Dans ce nouvel espace, les individus initiaux (noeuds du graphe) sont mieux séparés et c'est pourquoi un algorithme de clustering tel que k-means donne de bons résultats. Notamment, nous constatons sur la figure 4.2 que spectral est apte à retrouver à la fois des clusters convexes, non-convexes et de densités différentes.

L'inconvénient majeur de spectral est sa complexité de calcul en $O(N^3)$ qui le rend peu performant sur des bases de données volumineuses et la diversité des normalisations possibles entre la matrice de distance et le laplacien.

DBSCAN et OPTICS

Contrairement aux algorithmes précédents, DBSCAN (et OPTICS) sont basés sur la densité. Proposé par Ester et al. [71], l'algorithme DBSCAN utilise deux paramètres : un rayon ε et un nombre minimum de points $minPts$. L'idée de l'algorithme est, pour un objet donné $x \in \mathcal{D}$, déterminer son ε-voisinage $V_\varepsilon(x)$ défini tel que :

$$V_\varepsilon(x) = \{y | y \in \mathcal{D}, d(x, y) \leqslant \varepsilon\} \tag{4.6}$$

Le voisinage correspond à la boule de rayon ε autour de l'objet x. On dit alors que le voisinage de x est dense si :

$$|V_\varepsilon(x)| \geqslant minPts \tag{4.7}$$

Ainsi, pour un objet x donné dont le voisinage est dense, on calcule ensuite le voisinage $V_\varepsilon(y)$ pour tout $y \in V_\varepsilon(x)$, on continue récursivement et de proche en proche pour tout voisinage dense afin de trouver l'ensemble des objets du cluster.

L'avantage de cette méthode, en plus de sa simplicité, est sa capacité à détecter des clusters non-convexes comme sur l'exemple de la figure 4.2. De plus, DBSCAN est

capable de détecter de lui-même le nombre de clusters ainsi que les individus aberrants (outliers), en noir sur la figure précédente.

Cependant, les hyperparamètres ε et $minPts$ de l'algorithme sont souvent difficiles à estimer et il n'est pas capable de générer les clusters de densités différentes comme on le remarque sur la figure 4.2 au niveau de la 3ème ligne. L'algorithme OPTICS [12] vise à pallier ce défaut en inspectant les ruptures de densité dans le diagramme des distances entre points les plus proches (*reachability-plot*).

Clustering Hiérarchique

Précisons en préambule de cette section que nous aborderons uniquement la *classification ascendante hiérarchique* (ou CAH) qui est la technique de clustering hiérarchique la plus communément utilisée. Pour plus de détails, nous renvoyons à [170].

La CAH organise les données dans un système de classes emboîtées dont l'hétérogéneité augmente avec la taille des classes. Ce système représente une hiérarchie de parties indicées que l'on visualise graphiquement sous la forme d'un arbre hiérarchique appelé *dendrogramme*. Le dendrogramme fournit une description et visualisation précieuses et très informatives des structures potentielles de regroupement des données.

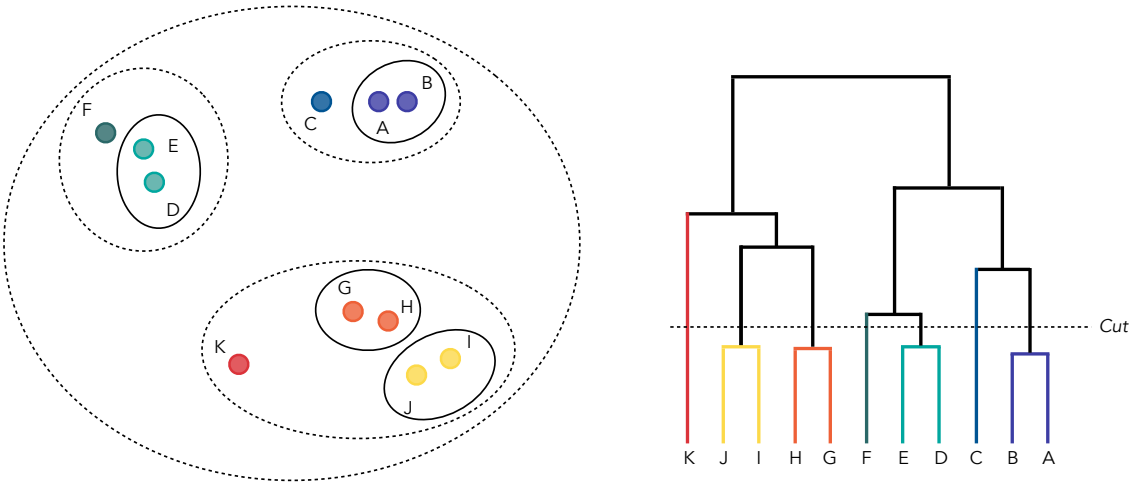


Figure 4.3 – Exemple de regroupement hiérarchique et de dendrogramme

Formellement, une hiérarchie de parties $H_{\mathcal{D}}$ d'un ensemble de données $\mathcal{D}$ est un sous-ensemble de $\mathcal{P}(\mathcal{D})$ ayant les propriétés suivantes :

- $\mathcal{D} \in H_{\mathcal{D}}$, l'ensemble complet des données est dans la hiérarchie (en haut de l'arbre).
- $\forall x \in \mathcal{D}, \{x\} \in H_{\mathcal{D}}$, toute donnée individuelle est dans la hiérarchie (en bas de l'arbre).
- $\forall h, h' \in H_{\mathcal{D}}, h \cap h' \in \{\varnothing, h, h'\}$, autrement dit, pour deux parties quelconques de $H_{\mathcal{D}}$, celles-ci sont soit disjointes, soit incluses l'une dans l'autre.

La hiérarchie de parties $H_{\mathcal{D}}$ est dite indicée lorsque l'on peut affecter, pour tout $h \in H_{\mathcal{D}}$, un nombre $v(h) \geqslant 0$ tel que si $h \subset h'$, alors $v(h) < v(h')$.

Une propriété remarquable est qu'il est équivalent de définir une hiérarchie de parties indicées sur $\mathcal{D}$ ou de munir $\mathcal{D}$ d'une ultramétrique [170].

Pour rappel, une distance ultramétrique d_u (ou simplement ultramétrique), est une distance où l'inégalité triangulaire est substituée par la propriété :

$$d_u(x, y) \leqslant \max\{d_u(x, z), d_u(y, z)\} \tag{4.8}$$

Cette propriété, plus restrictive que l'inégalité triangulaire, induit que toute ultramétrique est également une distance.

Ainsi, il est possible de construire un arbre hiérarchique à partir d'une matrice de distance (i.e., d'une dissimilarité d) pour rendre celle-ci ultramétrique.

La construction de l'arbre hiérarchique est assurée par l'algorithme Agglomerative Nesting (AGNES) dont la complexité peut-être ramenée en $\Theta(N^2)$ [168] pour une matrice de distance donnée en entrée. Celui-ci peut être résumé comme suit :

Data : Matrice de distance D sur $\mathcal{D}$.
Dissimilarité $\delta : \mathcal{P}(\mathcal{D}) \times \mathcal{P}(\mathcal{D}) \to \mathbb{R}^+$ entre les parties de $\mathcal{D}$.
Result : Classification hiérarchique $H_{\mathcal{D}}$
do
  $(h_i, h_j) \leftarrow \text{argmin}_{h,h' \in \mathcal{P}(\mathcal{D})}\{\delta(h, h')\}$ ▷ On agrège les éléments h_i et h_j.
  Supprimer la ligne/colonne h_j de D.
  Remplacer la ligne/colonne h_i par $h_i \cup h_j$.
  for $k : 1 \to |D|$ **do**
    ▷ Mise à jour de la matrice de distance.
    $D_{i,k} \leftarrow \delta(h_i, h_k)$
    $D_{k,i} \leftarrow D_{i,k}$
  end
while *Continuer jusqu'à agréger tous les éléments de D*;

La première ligne de l'algorithme ci-dessus (celle avec le argmin) montre qu'il suffit d'établir un ordonnancement, c'est-à-dire un classement, des couples d'éléments de $\mathcal{D}$ dans l'ordre croissant de leur dissimilarité pour créer un arbre hiérarchique [117]. On parle de pré-ordonnance dans le cas où certaines distances sont égales.

Ainsi, la définition de la dissimilarité δ constitue une des clés du fonctionnement de la méthode. En général, elle repose sur le choix d'un *critère d'agrégation*. Il existe de nombreux critères d'agrégation [120] dont Lance et Williams fournissent un cadre unificateur dans [128]. Dans notre cas, nous décrivons ici les trois critères les plus couramment employés : Ward, Single linkage et Complete linkage. La figure 4.2 offre un rendu visuel de leur fonctionnement respectivement aux colonnes 5, 6 et 7.

Critère de Ward Imaginé par Joe H. Ward [238], ce critère permet, de façon similaire à la fonction objective de k-means, de minimiser l'inertie intra-classe à chaque étape de la procédure de construction de l'arbre. On retrouve d'ailleurs un résultat de clustering très similaire entre K-medoid et Ward sur la figure 4.2. L'équation du critère d'agrégation est telle que :

$$\delta_{ward}(X, Y) = \frac{|X||Y|}{|X| + |Y|} \times \|\mu_X - \mu_Y\|^2 \tag{4.9}$$

Où μ_X, μ_Y sont respectivement les centroids des nuages de points X et Y. Cette formule fait ainsi l'hypothèse de disposer de données décrites dans un espace euclidien. Cependant, et en pratique, le calcul est assuré par la méthode de Lance et Williams [128] pour toute matrice de dissimilarité.

Critère du saut minimal Le critère du saut minimal (*Single linkage*) agrège en priorité les paires d'ensembles en considérant la plus petite distance entre les éléments de ceux-ci.

$$\delta_{\min}(X, Y) = \min_{x \in X, y \in Y} \{d(x, y)\} \tag{4.10}$$

Ce critère peut induire un effet de chaîne c'est-à-dire que deux objets très éloignés (au sens de d) mais reliés l'un à l'autre par une suite d'objets proches les uns des autres seront rassemblés dans la même classe.

Critère du diamètre Le critère du diamètre (*Complete linkage*) agrège en priorité les ensembles qui minimisent la distance maximale entre les éléments de ceux-ci.

$$\delta_{\max}(X, Y) = \max_{x \in X, y \in Y} \{d(x, y)\} \tag{4.11}$$

Le critère du diamètre favorise la compacité des classes et leur interprétation.

Coupure et obtention des classes La partition finale en clusters est fournie par une coupure de l'arbre selon une droite horizontale. La zone de coupure s'effectue en général selon le critère de maximisation du saut d'inertie : on coupe en cherchant les deux étages successifs où la distance dans l'arbre est la plus grande. On peut également chercher le nombre de clusters qui maximise l'indice silhouette [199], notamment dans le cas où les clusters sont susceptibles d'être convexes comme dans le cas d'utilisation du critère de Ward, ou bien s'en remettre à des critères métiers. Au demeurant, rappelons que la définition du nombre optimal de clusters est une question difficile et nous renvoyons à [170] pour de plus amples détails et techniques à ce sujet.

Algorithmes	**Complexité temporelle**	**Visualisation**	**Basé sur**	**Clusters non-convexes**	**Outliers**
K-Medoid	$O(N^2)$	×	Minimise la somme des distances aux medoids	×	×
Spectral	$O(N^3)$	×	Matrice des vecteurs propres du Laplacien de la matrice de similarité	✓	×
DBSCAN, OPTICS	$O(N\log(N))$	×	Densité	✓	✓
Hierarchical Single	$O(N^2)$	✓	Minimise la distance minimale entre éléments	×	×
Hierarchical Complete			Minimise la distance maximale entre éléments	✓	×
Hierarchical Ward			Minimise l'inertie intra-classe	×	×

Table 4.2 – Résumé synthétique des méthodes de clustering pour les séquences sémantiques

Réduction de dimensionnalité

Comme abordé au début de la section, un des problèmes majeurs lié au clustering de séquences sémantiques est le manque d'intuition quant à la topologie formée par la dissimilarité retenue et l'ensemble des séquences. Par exemple, classer des points dans l'espace $\mathbb{R}^2$ muni d'une distance ℓ_1 ou ℓ_2, comme sur la figure 4.2 est très intuitif pour quiconque car *on visualise* les objets.

Afin de briser la barrière d'abstraction formée par la topologie étrange sur laquelle nous officions et retransférer les données dans l'espace $\mathbb{R}^2$, il est possible d'avoir recours à une techniques de réduction de dimensionnalité sur la matrice de distance D. Il existe plusieurs techniques de réduction de dimensionnalité dont la figure 4.4 donne un aperçu comparatif sur quatre jeux de données classiques de la littérature. Notamment, on voit sur la figure que la méthode *UMAP* (Uniform Manifold Approximation and Projection) proposée par Mc.Innes et al. dans [148] sépare mieux les différentes classes pour la plupart des jeux de données. En ce sens, UMAP surpasse les précédentes méthodes utilisées pour cette tâche, en particulier t-SNE qui était l'ancienne méthode privilégiée. En outre, UMAP est plus rapide que t-SNE et est capable de traiter des données jeux de données de haute dimensionnalité. De plus, elle réclame peu d'hyperparamètres : $n\_neighbors$ un nombre de voisins à prendre en compte dans la reproduction de la distribution des distances et min_dist qui permet de contrôler la rigueur avec laquelle UMAP est autorisé à regrouper les points dans l'espace d'arrivée.

Sans détailler les outils mathématiques utilisés par cette technique (qui dépassent largement le cadre de cette thèse), UMAP est inspirée des techniques de réduction non-linéaire [226], en particulier LLE [201] qui cherche à conserver les relations topologiques locales en priorité. Au final les dimensions ne sont pas des combinaisons linéaires des distances de départ, comme cela serait le cas pour une ACP classique, mais une reproduction de la distribution normale des distances dans l'espace de départ par une distribution de Student à l'arrivée.

Ainsi, UMAP fournit un mode de visualisation efficace pour apprécier les interactions topologiques dans le cadre d'objets et distances complexes. De plus, une fois les données transformées dans l'espace $\mathbb{R}^2$, il est possible de se ré-approprier la plupart des techniques de clustering, notamment celles liées au clustering flou qui permettraient de représenter non plus une appartenance unique et stricte à un comportement dans notre cas, mais plutôt un degré d'appartenance à de multiples comportements détectés.

Clustering : application à la mobilité sémantique

Au meilleur de nos connaissances, il existe plutôt peu d'articles scientifiques abordant la problématique du clustering de séquences de mobilité sémantique. Les raisons sont multi-factorielles et concernent des points déjà évoqués dans les précédentes sections comme : la disponibilité des données, les questions éthiques relatives à la vie privée des utilisateurs ou encore de méthodes, le clustering étant une approche minoritaire

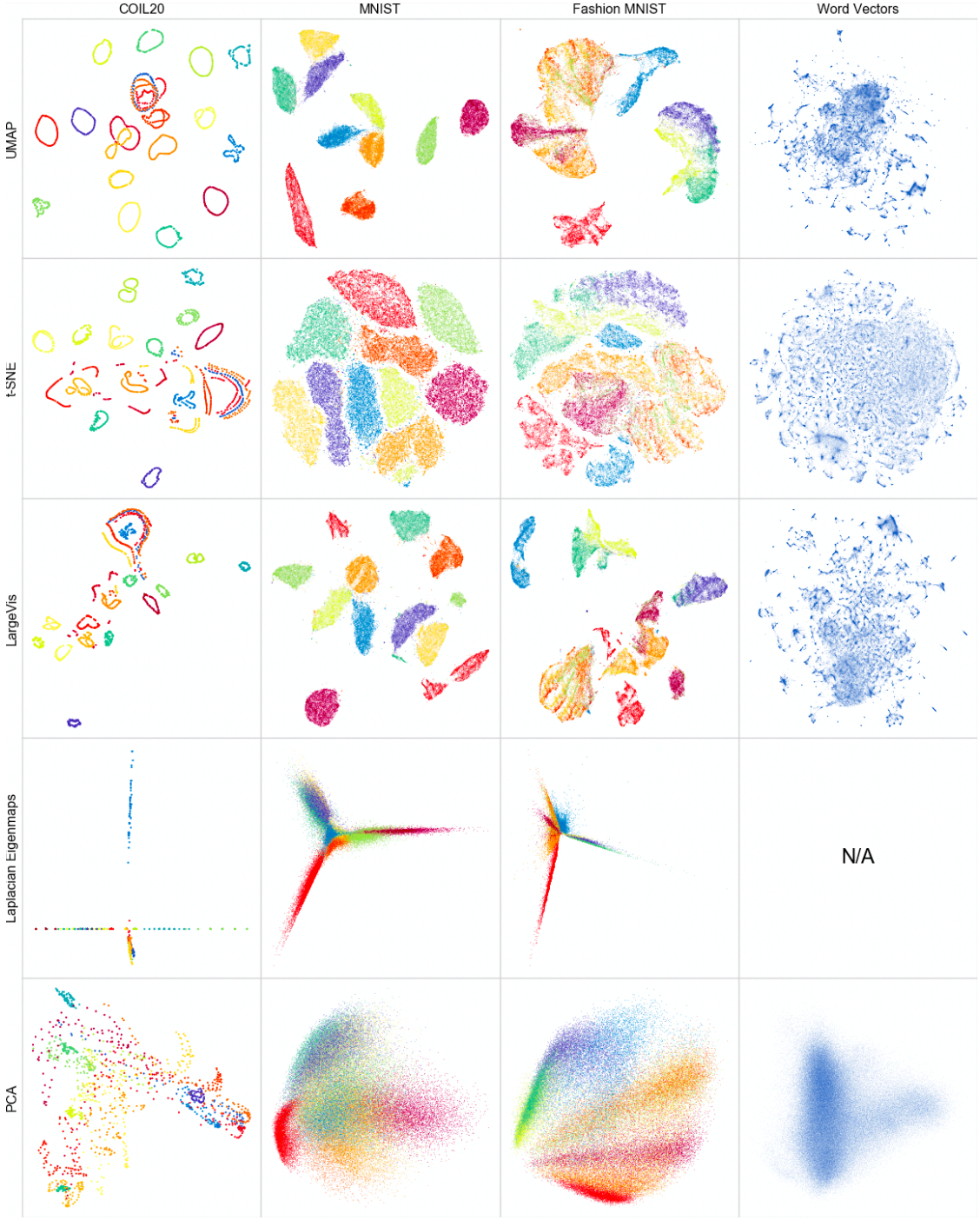


Figure 4.4 – Réduction de dimensionnalité sur des jeux de données de la littérature [148]

par rapport à l'approche FSM. Nous décrivons ici quelques articles ayant inspiré notre démarche et cadre de travail quant à la méthodologie d'analyse pour la découverte de comportements de mobilité.

L'article de Jiang et al. [113] propose une approche de clustering pour l'extraction de comportements de mobilité. L'étude est conduite sur un Travel and Activity Survey au sein de la ville de Chicago en 2008. Les activités recensées sont organisées de façon taxonomique en 9 grandes catégories $A_{i\in\{1\ldots,9\}}$ (e,g., home, work, school, transport, etc.)

Pour représenter une séquence de mobilité au cours d'une journée, les auteurs proposent un modèle binaire où le temps est segmenté en intervalles $I_{k\in\{1,\ldots,288\}}$ d'une durée de 5 minutes et où chaque intervalle est associé à un vecteur binaire $v_k \in \{0,1\}^9$ tel que $v_{i,k} = 1$ si et seulement si une activité $a \in A_i$ est réalisée pendant l'intervalle I_k. Les auteurs imposent que $\|v_k\| = 1$, autrement dit, une seule activité est réalisée par intervalle. Il en résulte une séquence S représentant une journée telle que $S \in \{0,1\}^{2592}$. Afin de réduire la taille de ces séquences et extraire les dimensions importantes, les auteurs conduisent une ACP sur l'ensemble des séquences. Ils appliquent ensuite une distance euclidienne sur les séquences en sortie du processus d'ACP puis un algorithme des k-moyennes afin d'extraire les différents comportements de mobilité. Une explication des clusters est donnée en termes d'analyse de la distribution des activités au cours de la journée. Une analyse basée sur des variables explicatives sociaux démographiques est également conduite permettant d'étayer le discours des auteurs en termes de comportements.

Néanmoins, nous pouvons relever plusieurs écueils quant à la méthode mise en place par les auteurs. Premièrement la représentation des séquences sous une forme discrétisée par pas de 5 min avec un codage binaire des activités conduit à une représentation stricte du temps et de la sémantique. Par exemple, des individus ayant effectué des activités similaires à des instants proches mais néanmoins différents risquent d'être jugés peu similaires. En outre cette représentation ne permet pas de refléter la notion floue, continue et intuitive du temps comme de la sémantique où le principe de taxonomie n'est pas exploitée. Un autre point est que la vision des auteurs est orientée selon une approche *structurelle* du temps, à l'opposé d'une approche *compositionnelle* [1, 150]. Dans la première, c'est le temps qui structure les journées et oriente l'analyse ; on s'attache alors à l'importance du budget temps associé à chaque activité. La seconde approche tient davantage compte des activités en elle-même ainsi que de leur fréquence et de leur relation de précédence pour comprendre comment la journée se compose. Ainsi, si la première approche incarne davantage la réalité, elle élude en grande partie la valeur des activités intrinsèquement courtes qui peuvent pourtant avoir une importance métier, par exemple les activités de déplacements. Il doit être noté que la prise en compte d'autres indicateurs comme la fréquence d'activités ou les motifs topologiques de déplacement (daily pattern) sont proposés par les auteurs dans [115, 116, 114].

Une méthodologie similaire aux travaux de Jiang et al. est proposée par Su et al. dans [224]. Dans ce dernier, les auteurs utilisent un jeu de données issu d'un Travel Survey Californien de 2017. Un processus de clustering basé sur la Dynamic Hamming Distance (DHD) telle que décrite équation 3.28, suivi d'une CAH est appliqué à l'ensemble des séquences. L'analyse reprend les indicateurs proposés par Jiang et al. mais est étendue par des indicateurs issus du package du TraMineR comme l'entropie ou la complexité des séquences [80] et détaille avec un soin particulier les motifs topologiques découverts.

Enfin, nous citerons le travail récent de Damiani et al. [53] qui se ré-approprie des techniques issues du traitement automatique du langage tel que le plongement lexical (notamment le word embedding : word2vec [155]) afin de transposer un ensemble de trajectoires symboliques dans un espace latent. En outre, les auteurs proposent un framework appelé *mob2vec* basé sur l'apprentissage d'un espace latent des trajectoires à l'aide de la méthode Sqn2Vec [173] et d'une réduction de dimensionnalité via UMAP pour mieux appréhender la topologie créée. Cette méthodologie très originale peut former un cadre théorique intéressant pour la représentation des trajectoires symboliques permettant de conserver de bonnes propriétés mathématiques de l'espace et l'utilisation des méthodes de clustering classiques. Néanmoins, les auteurs mettent peu en avant l'interprétation et l'analyse des comportements issus de l'espace créé. Nous pensons que ces tâches sont les plus susceptibles d'intérêt pour l'utilisateur final.

La fin de ce chapitre sera consacrée à la caractérisation des clusters découverts en termes de comportements. Nous abordons dans la section suivante les différents indicateurs classiques issus de la littérature permettant de résumer de façon complémentaire les différentes propriétés d'un ensemble de séquences sémantiques.

4.3 Analyse et explicabilité des comportements

Cette section développe différentes méthodes et techniques permettant l'analyse et l'explicabilité d'ensembles de séquences sémantiques en termes de comportements intelligibles. La première sous-section est dédiée à la description d'indicateurs statistiques permettant une description globale des séquences sémantiques et sur les techniques permettant de synthétiser, résumer ou expliquer un ensemble de données complexes. La seconde sous-section généralise ce dernier point et aborde les problématiques récentes de l'eXplainable Artificial Intelligence (XAI) qui vise la production de méthodes d'apprentissage et de fouille de données éthiques et centrées sur l'intelligibilité des modèles et la compréhension de l'utilisateur.

4.3.1 Analyses statistiques de la mobilité et des séquences

Comme abordé en section 3.2, de nombreux auteurs et travaux se sont employés à découvrir les grandes lois statistiques et propriétés qui gouvernent la mobilité dans

le but d'en comprendre les principes fondateurs et d'en extraire la substance. Globalement, nous avons vu plus tôt que l'immense majorité des individus sont caractérisés par des déplacements de faible distance [85], répètent les mêmes activités (e.g., fréquentent les mêmes lieux) [219] ce qui forme une périodicité des observations [219, 206, 43, 181], que le temps est entre-coupé de quelques activités très longues mais majoritairement de nombreuses activités courtes [16] et que l'apparente complexité de la mobilité humaine masque un haut degré de prédictibilité de celle-ci [220, 8].

Ces études se sont très majoritairement concentrées jusqu'à présent sur le caractère spatial de la mobilité. Or, la plupart de celles-ci peuvent être facilement transposées dans un cadre sémantique afin de fournir un schéma d'analyse formel pour la compréhension des caractéristiques des séquences de mobilité sémantique et leur traduction en comportements intelligibles d'un point de vue thématique. Ces ensembles de séquences étant des objets complexes, il est nécessaire de décomposer leur analyse en différentes sous-observations afin de saisir pleinement leur nature et la rendre assimilable pour l'utilisateur. En outre, nous proposons une décomposition selon 6 axes d'analyses permettant de renseigner de façon complémentaire les séquences et clusters :

1. *Distribution des fréquences* : Renseigne le nombre de symboles au total et par séquences.
2. *Transitions* : Étudie les transitions et la cyclicité des séquences.
3. *Désordre* : Quantifie la prédictibilité et la complexité des séquences.
4. *Lien statistiques* : Renseigne les liens et caractéristiques saillantes entre symboles, séquences et clusters.
5. *Dispersion* : Mesure la compacité, le pourcentage d'éléments aberrants et la ressemblance des séquences au sein d'un même cluster.
6. *Résumé explicatif* : Fournit un résumé synthétique permettant de caractériser un élément ou un groupe d'éléments de façon minimale et concise.

Nous détaillons dans la suite de cette section ces différentes familles d'analyse.

Distribution des fréquences

La *distribution des fréquences* intervient au niveau de deux caractéristiques des séquences : le nombre de symboles qui la composent (*distribution des longueurs*) et la fréquence d'apparition des symboles au sein de l'ensemble des séquences (*distribution des états*). La première analyse permet de donner une indication sur la mobilité et/ou le caractère actif l'individu. Plus une séquence est longue (i.e., composée de beaucoup de d'activités), plus cela traduit, de manière indirecte, que la personne est mobile / active. En effet, en accord, avec la perspective de la Time-Geography, nous n'ignorons pas que l'Espace et la Sémantique ont des dimensions corrélées, en particulier dans les milieux urbains où les zones ont souvent une sémantique associée (quartier

d'affaires, zone résidentielle, zone commerciale, etc). La seconde analyse reprend les concepts développés par Song et al. [219] où les auteurs quantifient la fréquence de visite d'un lieu par un individu. Nous proposons de quantifier la fréquence associée à la sémantique des lieux. En outre, cette analyse permet de mettre en lumière les éléments constitutifs de la mobilité (i.e., les plus redondants). Une analyse similaire selon une vision temporelle est également possible en comptant, non plus le nombre d'occurrences des symboles, mais le budget temps qui leur est associé.

Transitions

L'étude des *transitions* entre les symboles (i.e., activités) fournit une compréhension importante quant à la nature même des séquences. Prenons en exemple la linguistique : l'analyse fréquentielle des bigrammes (paires de lettres dans les mots) est capable de caractériser une langue avec un degré de certitude supérieur à 90 % [111]. De la même façon, l'analyse des transitions entre activités / lieux consécutifs, appelée *matrice Origine-Destination* (OD), permet de mettre en évidence la structure interne générale qui règle les séquences de mobilité, comme s'il s'agissait des mots d'une langue. En outre, Li et Lee filent l'analogie et montrent qu'il est possible de représenter la majorité des séquences de mobilité sémantique à l'aide de grammaires non-contextuelles stochastiques [140].

Une notion proche du concept de transition est celle des *daily patterns* [206] (ou motifs topologiques) qui cherchent à capter la topologie générale des séquences à l'aide de graphes. Il est ainsi plus facile de mettre en évidence les structures de redondance qui se matérialisent par un cycle dans les graphes résultants. Nous proposons l'algorithme suivant pour le calcul des daily patterns :

Data : Base de données des séquences $\mathcal{D}$.
Result : Dictionnaire $\mathcal{G}$ des graphes non-isomorphes et de leur fréquence.
$\mathcal{G} \leftarrow \varnothing$ ▷ Dictionnaire $\mathcal{G}$ où chaque clé est un graphe et la valeur associée un entier.
▷ Pour chaque séquence $S \in \mathcal{D}$, on construit le graphe – daily pattern – associé.
for $S \in \mathcal{D}$ **do**
  $V_S \leftarrow \{x | x \in S\}$ ▷ Ensemble des noeuds.
  $E_S \leftarrow \{(x_i, x_{i+1}) | i \in [\![1, |S| - 1]\!]\}$ ▷ Ensemble des arcs.
  $G_S \leftarrow (V_S, E_S)$
  if $\exists G \in \mathcal{G}.keys() | G \simeq G_S$ **then**
    ▷ S'il existe déjà un graphe G isomorphe à G_S dans $\mathcal{G}$.
    $\mathcal{G}[G] \leftarrow \mathcal{G}[G] + 1$ ▷ On incrémente la fréquence du graphe G.
  else
    $\mathcal{G}[G_S] \leftarrow 1$ ▷ On ajoute le nouveau graphe $\mathcal{G}$.
  end
end

Notons que la vérification d'isomorphisme entre deux graphes G et G' peut être réalisée par l'algorithme Nauty de McKay et al. [149].

Désordre

La question de la prédictibilité du comportement humain est fondamentale dans toute une série d'applications allant de l'épidémiologie à la planification urbaine et la gestion des ressources. Dans les séquences de mobilité sémantique, cette notion est étroitement liée à celle de *désordre*, elle-même rattachée à la notion d'*entropie*. Initialement introduite par Shannon [211], sa formalisation diffère légèrement dans le cadre particulier des séquences où plusieurs définitions cohabitent [220, 80]. L'idée générale étant toujours rattachée à l'estimation d'un degré de désordre et/ou quantité d'information contenue dans la séquence.

On donne ici la définition de l'entropie d'une séquence S telle que proposée par Song et al. dans [220]. Déjà, plusieurs entropies peuvent être définies.

- L'entropie aléatoire $H^{rand} = \log_2 \delta(S)$, où $\delta(S)$ est le nombre de symboles distincts dans la séquence S.
- L'entropie temporelle non-corrélée $H^{unc} = -\sum_{i=1}^{|S|} p(x_i) \log_2 p(x_i)$, où $p(x_i)$ est la probabilité d'apparition de l'activité x_i dans S.
- L'entropie réelle H qui dépend à la fois de la probabilité d'apparition de x_i mais aussi de son ordre d'apparition dans la séquence S.

$$H(S) = -\sum_{S' \subset S} p(S') \log_2 p(S') \tag{4.12}$$

  où $p(S')$ est la probabilité de trouver la sous-séquence S' dans S.

En pratique, la complexité de calcul l'équation 4.12 rend H incalculable pour des séquences longues. Toutefois, il est possible de calculer un estimateur H^{est} de H tel que proposé par Kontoyiannis et al. dans [123] :

$$H^{est}(S) = \left(\frac{1}{|S|} \sum_i \lambda_i \right)^{-1} \log_2 |S| \tag{4.13}$$

où $\lambda_i = \text{argmin}_{k \geqslant 1} \{x_i ... x_k \notin x_1 ... x_{i-1}\}$ est la taille de la plus petite sous-séquence débutant à i et qui n'est pas contenue dans la sous-séquence $\langle x_1, ..., x_{i-1} \rangle$. En outre, les auteurs ont démontré que $\lim_{|S| \to \infty} H^{est}(S) = H(S)$.

De plus, Song et al. dévoilent le lien entre entropie et *prédictibilité*. Les auteurs définissent la notion de prédictibilité Π comme étant la probabilité qu'un algorithme approprié (i.e., basé sur l'entropie de S) puisse prédire correctement la destination / activité future de l'individu. Grâce à l'inégalité de Fano [74], il est possible d'obtenir une borne supérieure $\Pi^{\max}$ de Π [220]. $\Pi^{\max}$ est obtenu par la résolution approchée de l'équation :

$$H(S) = \mathcal{H}(\Pi^{max}) + (1 - \Pi^{max}) \log_2(|S| - 1) \tag{4.14}$$

où $\mathcal{H}(x) = -x \log_2 x - (1 - x) \log_2(1 - x)$ est la fonction d'entropie binaire.

Une mesure de *complexité* de la séquence est également proposée dans [80] tenant compte du nombre de transitions dans la séquence. Enfin, le nombre de *symboles distincts* dans une séquence permet aussi d'exprimer la notion de désordre. Déjà analysé dans [219] où les auteurs montrent que celui-ci est indépendant de la distance parcourue, il est repris par Teixeira et al. [225] où les auteurs mettent en avant son lien étroit avec l'entropie telle que formulée par Song et al. dans [220] avec l'avantage d'être très simple à comprendre et universel.

Lien statistique

L'étude de *liens statistiques* entre variables qualitatives est une tâche qui n'est pas si triviale et dont certains concepts sont parfois confondus. Dans la littérature statistique qui étudie le lien entre deux variables qualitatives X et Y, on distingue deux catégories d'indicateurs : (i) Des indicateurs qui s'intéressent à l'apport d'information de X par rapport à Y et (ii) des indicateurs qui s'intéressent à l'indépendance entre X et Y, en ce sens que connaissance de la réalisation de l'une des variables n'a aucune incidence sur la probabilité de réalisation de l'autre variable.

Dans la première catégorie, les indicateurs se basent principalement sur une distribution conditionnelle. Par exemple, dans le domaine de la recherche de *règles d'association* [5], la mesure d'intérêt du $lift(X \rightarrow Y) = \frac{supp(X \cap Y)}{supp(X) \times supp(Y)}$ permet de mesurer l'amélioration apportée par la règle d'association entre X et Y par rapport à un jeu de transactions aléatoire. Si le lift est supérieur à 1, alors on peut conclure à un lien d'association entre X et Y [3]. Ainsi, si on s'intéresse aux activités pratiquées par un individu au cours d'une journée, sans considérer de contraintes temporelles, alors cette approche peut être utile notamment pour détecter des ensembles d'activités ayant un lien dans leur réalisation. Par exemple, on pourra aisément admettre que les activités "Déposer quelqu'un" et "Aller chercher quelqu'un" sont liées et souvent réalisées au cours de la même journée.

La seconde catégorie consiste à vérifier l'hypothèse nulle H_0 selon laquelle X et Y sont deux variables indépendantes à l'aide d'un test d'indépendance du χ^2 et l'obtention d'une p-valeur. Par exemple, X pourra être la liste des clusters de séquences de mobilité sémantique et Y les activités réalisées, on souhaite alors vérifier que la réalisation des activités à un effet sur l'attribution de nos clusters. Supposons maintenant que le test du χ^2 a montré une dépendance entre les variables X et Y (i.e., $p < 0.05$). Le problème qui se pose désormais est de déterminer quelles sont les combinaisons (modalités ligne-colonne) pour lesquelles l'association – attraction ou

3. Cette notion ne doit pas être confondue avec celle de *corrélation* qui mesure l'existence d'une fonction linéaire entre les valeurs des deux variables, ce qui n'a pas de sens dans le cadre de variables qualitatives.

répulsion – est significative. Pour se faire, on peut effectuer une analyse factorielle des correspondances [22] ou calculer les *résidus de pearson* [91] qui mesurent l'écart des modalités entre la valeur observée et la valeur théorique. Nous détaillons ici la seconde méthode.

Ainsi, considérons un échantillon de données de taille N distribuées selon deux variables aléatoires A et B de modalités respectives $a_1, ..., a_p$ et $b_1, ..., b_q$. On note $(n_{ij}), 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant q$, le nombre respectif où les modalités a_i et b_j sont observées conjointement. Une telle représentation des données peut être modélisée par une *table de contingence*. On note $(n^*_{ij}) = \frac{n_{i+} \times n_{+j}}{N}$ la valeur théorique correspondant à l'observation de a_i et b_j. Le résidu de Pearson r_{ij} correspondant est défini tel que :

$$r_{ij} = \frac{n_{ij} - n^*_{ij}}{\sqrt{n^*_{ij}}} \tag{4.15}$$

Notons ici que la statistique du χ^2 est alors $\chi^2 = \sum_i \sum_j r_{ij}$.

Le coefficient r_{ij} représente la force et la direction du lien qui unit les modalités a_i et b_j. La force est définie par la valeur absolue du résidu $|r_{ij}|$ et sa direction par son signe. L'unité est en écart-type, ainsi un résidu tel que $|r_{ij}| > 2$ montre un écart significatif à la normale et indique un lien statistique entre a_i et b_j avec un degré de confiance supérieur à 95% (p-valeur < 0.05).

Dispersion

Une fois les clusters établis, il est intéressant d'essayer d'analyser la *dispersion* des séquences dans les clusters et à travers l'espace topologique. Deux premiers indicateurs pour appréhender la compacité des clusters sont le *rayon et le diamètre* du cluster qui conservent une interprétation géométrique élémentaire et forment un point de départ assez palpable. D'ailleurs, de nombreux indices jugeant la qualité des partitions formées par le clustering se basent sur des critères proches, comme l'indice de Dunn [63] qui mesure la distance maximale qui sépare deux objets classés ensemble (diamètre) et la distance minimale qui sépare deux objets classés séparément. L'indice de Davies-Bouldin [54] quant à lui s'inspire du rayon et représente la moyenne du rapport maximal entre la distance d'un objet au centre de sa classe et la distance entre deux centres de classes différentes. Néanmoins, ces indicateurs s'accommodent mal de clusters non convexes ou formés par effet de chaîne. Pour ces raisons, on ferra généralement l'hypothèse de clusters convexes. L'indice *Silhouette* de Rousseeuw et al. [199] quantifie la similarité d'un objet avec son propre groupe (cohésion) par rapport aux autres groupes (séparation). Ce critère est très approprié pour valider la qualité d'un clustering effectué selon une méthode comme k-means ou hiérarchique selon le critère de Ward car il est basé sur les mêmes hypothèses de construction (i.e., minimisation de l'inertie intra-classe).

Enfin, comme abordé dans la section précédente, il reste très commode de pouvoir

visualiser les objets dans un espace 2D. UMAP permet la conservation des topologies locales et ainsi apprécier les interactions d'objets proches les uns des autres.

Résumé

Bien que les méthodes précédentes apportent chacune un éclairage différent et particulier sur les données, il est souvent utile pour des questions de concision du discours de disposer d'un *résumé* synthétique des clusters formés. L'idée générale est alors de capturer, au sein d'un mode unique de visualisation simple et intuitif pour l'utilisateur, l'ensemble des caractéristiques saillantes du comportement ou une description minimale.

Les *arbres de décision* sont de bons candidats pour ce type de tâche de par leur lisibilité et leur capacité à extraire des règles de décision. Par exemple, l'algorithme C4.5 de Quinlan [191] se base sur le gain d'information (une forme d'entropie) afin d'extraire les variables et modalités qui discriminent le mieux chaque classe dans le but de les rendre les plus homogènes possible. Il peut donc être pertinent, une fois les clusters établis, de construire un arbre de décision qui résume par un ensemble de règles caractéristiques les comportements dans les clusters. Néanmoins, sa mise en place n'est pas toujours simple, en particulier quant à la définition du critère de segmentation, de la profondeur de l'arbre et stratégie d'élagage.

Un autre moyen utilisé pour résumer un ensemble d'objets complexes peut être d'exhiber un *élément prototypique* dudit ensemble. Par exemple, le medoid défini équation 4.2 ou le mode (i.e., objet le plus fréquent) sont des éléments aptes à capter la nature d'un ensemble de données. Cependant, le mode reste un indicateur sensible et n'est pas toujours unique ; quant au medoid, celui-ci est peu pertinent lorsque la topologie de l'ensemble ne forme pas une hypersphère. Pour pallier ces manques, Lesot et Kruse soulèvent le problème de la typicalité qui consiste à déterminer une mesure d'agrégation entre similarités avec les éléments d'un cluster et dissimilarités avec les éléments des autres clusters et proposent des techniques inspirées des modèles flous pour déterminer des éléments prototypiques dans les données [136, 137].

Une autre approche pour appréhender le processus de machine learning qui a reçu quantité d'intérêt récemment est celle des *explications contre-factuelles* introduite dans [183] par Judea Pearl. En philosophie, un raisonnement contre-factuel est une proposition de la forme "*Si ... alors ...*" qui imagine l'issu d'une situation en modifiant ses causes. Par exemple, la proposition "Si Napoléon avait gagné la bataille de Waterloo, alors la gare Centrale de Londres ne s'appellerait pas la gare de Waterloo." est un contre-factuel. Dans le cadre de l'explicabilité, on cherche alors à modifier le moins possible une donnée x afin d'obtenir une classe différente pour celle-ci, l'explication est donc locale. Dans notre cas, étant donnée une séquence S identifiée comme comportement b, on cherche à déterminer les changements minimaux à apporter à S pour qu'elle soit identifiée en comportement $b' \neq b$. Formellement, étant donné un classifieur (ou quelconque méthode de machine learning) C et une donnée x, on

cherche à construire l'explication e telle que $C(e) \neq C(x)$ tout en minimisant l'effort de changement, soit :

$$e^* = \text{argmin}_{e \in \mathcal{X}} \{\gamma_x(e)\} \tag{4.16}$$

où γ_x est une fonction de coût, par exemple la distance à x, et $\mathcal{X}$ l'espace de recherche. Notons que la définition de γ_x, de l'espace de recherche $\mathcal{X}$ ainsi que son exploration sont des problèmes clés de ce type de méthode [14]. Dans notre cas, on peut imaginer une solution économe telle que proposée par [89] qui se restreint aux données disponibles :

$$\tilde{e} = \text{argmin}_{e \in \sigma_\varphi(\mathcal{D})} \{d(x, e)\} \tag{4.17}$$

où d est la distance utilisée dans le processus de clustering et $\sigma_\varphi(\mathcal{D})$ correspond à un filtrage de l'ensemble des données $\mathcal{D}$ selon la condition logique φ. Ce procédé peut être très utile d'un point de vue pédagogique afin d'éclairer l'utilisateur (mais aussi le concepteur) sur le "Pourquoi" telle séquence a été identifiée comme tel comportement (i.e., cluster). De plus, la possibilité de filtrer les données permet notamment d'ignorer les variables sur lesquelles l'individu ne peut pas avoir d'impact pour changer son comportement (par exemple son âge) ou qui sont considérées comme non pertinentes. Cependant, cette solution fait une hypothèse très forte quant à la quantité de données (i.e., complétude) qui doit être importante pour être véritablement exploitable.

Enfin, une dernière approche consiste à résumer les variables et modalités saillantes d'un cluster à l'aide d'un *nuage de mots*. Ce type de résumé a l'avantage fort d'être *a priori* compréhensible par tous et de permettre un éclairage macroscopique préalable des données avant une inspection de graphiques plus détaillés. Ainsi, l'usage de termes et concepts abscons doit être proscrit ici pour laisser place à la simplicité. Néanmoins, les nuages de mots posent de nombreuses questions quant à l'objectif communicationnel qu'ils mettent en place, notamment en termes de visualisation. Notamment, le choix des informations à retenir est un problème épineux qui peut être envisagé selon différentes approches. L'approche la plus commune est l'approche fréquentiste qui consiste à apparaître les termes selon leur fréquence dans le jeu de données, ce qui pose l'inconvénient de masquer les termes minoritaires. Une alternative serait alors de faire apparaître les termes qui caractérisent le mieux l'ensemble de données (par l'utilisation de critères statistique comme des tests χ^2 ou de résidus de Pearson). Néanmoins, une question se pose sur la restitution de l'information quand celle-ci est absente, c'est-à-dire lorsque le jeu de données se caractérise par l'absence d'un terme. Par exemple, comment représenter à l'aide d'un nuage de mots un ensemble de données où la caractéristique principale est l'absence d'utilisation de la voiture ? Ce type d'interrogation vient s'ajouter à des problématiques plus prosaïques comme la taille des mots, leur disposition au sein du nuage ou encore leur couleur qui sont autant d'interrogations qui rendent l'usage des nuages de mots périlleux sans une étude ergonomique préalable.

Conclusion sur les indicateurs

La table 4.3 fournit un résumé de l'ensemble des méthodes et indicateurs permettant de caractériser un ensemble de séquences sémantiques de mobilité. Cette table reprend les 6 grandes catégories complémentaires que nous avons détaillées ci-dessus et dont chacune vient expliquer une des faces complexes du prisme de la mobilité. La colonne "use" indique si l'indicateur a été utilisé dans la contribution [165]. Ces indicateurs ont notamment été sélectionnés pour leur simplicité et le fait de pouvoir disposer de méthodes de visualisation efficaces, sobres et synthétiques à destination de l'utilisateur final.

Ainsi, par cet ensemble d'indicateurs, notre objectif est la mise en évidence des informations caractéristiques et essentielles des données et clusters établis. Aussi, ces informations doivent maintenir un équilibre entre exhaustivité et concision afin de prévenir la surcharge cognitive et veiller à ce qu'elles soient correctement assimilables et intelligibles. Pour ce faire, une ultime phase d'*explication* est nécessaire afin de finaliser le processus de compréhension auprès de l'utilisateur. Cette phase peut se concrétiser notamment à l'aide de représentations graphiques qui, combinées à la pratique du *data storytelling*, permettent de raconter une histoire avec les données en les présentant de façon pédagogique, visuelle et attrayante aux yeux de l'utilisateur qui va en extraire un *sens métier*.

Techniques	Description	Utilisé
	Distribution des fréquences	
Distribution des longueurs	Histogramme du nombre de symboles dans chaque séquence du jeu de données.	×
Distribution des états	Fréquence d'apparition de chaque symbole dans toute séquence du jeu de données.	×
	Transitions	
Matrice O-D	Nombre de transition depuis un état (i.e., symbole) x_i vers x_j.	×
Daily pattern	Représentation topologique sous forme de graphe des déplacements dans la séquence.	×
	Désordre	
Entropie & prédictibilité	Degré d'information, de surprise ou d'incertitude inhérent au prochaine état dans la séquence. Plusieurs définitions de l'entropie co-existent.	×
Complexité	Mesure composite entre l'entropie longitudinale et le nombre de transitions dans une séquence.	
Symboles distincts	Nombre de symboles distincts dans la séquence.	×
	Lien statistique	
Règles d'association	Relation, basée sur des mesures d'intérêt, quantifiant les liens d'apparition des symboles dans une séquence.	
Résidus de Pearson	Mesure du degré d'indépendance entre deux variables. Variante de la p-valeur.	×
	Dispersion	
Rayon et diamètre	Interprétations géométriques des distances entre éléments au sein d'un cluster.	×
Silhouette	Mesure la similarité d'un objet avec son propre groupe (cohésion) par rapport aux autres groupes (séparation).	×
	Résumé	
Arbre de décision	Ensemble de règles conditionnelles permettant de résumer un cluster.	
Éléments prototypiques	Séquence modèle permettant de représenter un comportement d'un cluster.	×
Nuage de mots	Ensemble de mots issus des variables caractérisant un cluster	×
Contre-factuel	Explication minimisant l'effort de changement d'une donnée pour qu'elle change de cluster.	

Table 4.3 – Résumé des indicateurs possibles pour l'analyse d'ensembles de séquences sémantiques de mobilité

4.3.2 Explicabilité de modèles

Au-delà de la question de la performance des méthodes d'apprentissage automatique, l'importance liée à la lisibilité et la compréhension humaine des résultats obtenus, notamment à des fins pratiques et éthiques, est croissante. Ces problématiques liées à la transparence des systèmes intelligents d'aide à la décision sont regroupées sous le terme "eXplainable Artificial Intelligence" (*XAI*).

Cette collaboration entre humain et machine est reconnue comme la condition *sine qua non* pour que l'IA puisse continuer à progresser de façon sereine et avec la confiance des utilisateurs et du grand public. Le besoin de résultats et de modèles clairs et interprétables est de plus en plus important, en particulier pour les algorithmes de type "boîte noire", lorsque les données sont énormes et complexes, ou pour les méthodes dotées de nombreux paramètres. Par exemple pour les entreprises qui souhaitent améliorer la gestion et la compréhension des besoins de ces clients [81] mais aussi améliorer l'ouverture de la découverte scientifique et le progrès de la recherche. L'interprétabilité est cruciale pour tester, observer et comprendre les différences entre les modèles. En outre, la compréhension des données améliore aussi le processus d'apprentissage et d'exploration en termes de validité.

Dans [2], Adadi et Berrada proposent une étude exhaustive des opportunités et techniques relatives à l'XAI en soulignant la prolifération des algorithmes dans notre société actuelle, notamment pour l'aide à la prise de décision. Aussi, lorsque ces systèmes mettent en jeu des vies humaines (décisions médicales) ou des coûts financiers considérables, il est d'une indéniable importance de connaître les raisons qui sous-tendent de telles recommandations. Cette nécessité de rendre les systèmes intelligents plus transparents, responsables, et "human-friendly" est motivée par les auteurs selon 4 quatre grands axes :

- Besoin de justifier (*explain to justify*). Ces dernières années ont été marquées par de nombreuses controverses concernant des systèmes basés sur l'IA et le Machine Learning (ML) produisant des résultats biaisés et/ou discriminatoires [131, 103]. Au regard de ces erreurs passées, la RGPD impose depuis Mai 2018 un "Droit à l'explication"[4] qui réglemente toute prise de décision à l'aide d'algorithme et autorise tout individu à s'opposer sur le fait d'être l'objet d'une décision fondée exclusivement sur un traitement automatisé [209]. En conséquence, il devient nécessaire aujourd'hui de fournir des éléments qui détaillent et justifient la prise de décision afin de démontrer ces processus comme étant justes et éthiques et afin instaurer la confiance du grand public.
- Besoin de contrôler (*explain to control*). Une meilleure compréhension du comportement du système offre une plus grande visibilité sur les vulnérabilités et les failles inconnues, et aide à identifier et corriger rapidement les erreurs.
- Besoin d'améliorer (*explain to improve*). Dans la même lignée que le point

4. Voir articles 22 et 13–15

précédent, les auteurs défendent qu'un modèle qui peut être expliqué et compris est un modèle qui peut être plus facilement amélioré.

- Besoin d'apprendre (*explain to discover*). Dans le cadre de modèles boîtes noires, il serait souhaitable de comprendre le mécanisme d'apprentissage des algorithmes. Par exemple, dans le cadre du jeu de Go, l' algorithme AlphaGo Zero [215] surpasse les joueurs humains ; il serait souhaitable que la machine puisse nous expliquer ses stratégies apprises afin d'en faire bénéficier les connaissances humaines.

Dès lors, nous distinguons deux approches en termes d'explication : (i) l'explication des données, relative à la signification des résultats eux-mêmes. Dans notre cas il s'agirait par exemple d'expliquer la nature et les caractéristiques des comportements découverts au sein des clusters. La seconde approche concerne (ii) l'explication des modèles, c'est-à-dire des méthodes et algorithmes – pourquoi retournent-ils ces résultats ? Ainsi, le premier type d'explication s'adresse avant tout aux thématiciens et experts métier tandis que la seconde est davantage au service du concepteur du processus de fouille dans un but de supervision et d'amélioration.

Par ailleurs, précisions ici que le domaine est en plein bouillonnement et manque encore de définitions nettes. Nous ne présenterons ici uniquement les approches d'explication des données en nous concentrons sur les méthodes basées sur la visualisation et les résumés, et nous renvoyons à l'exposé de Marie-Jeanne Lesot pour plus de références et d'explications sur les techniques actuelles [5] et aux surveys d'Adadi et Berrada [2], Guidotti et al. [89] et Arrieta et al. [13] qui englobent une bonne partie des débats, techniques et opportunités récentes quant au domaine de l'XAI.

Lorsque les processus de clustering mettent en oeuvre des objets complexes (e.g., des séquences sémantiques), une dissimilarité non euclidienne aux propriétés particulières voire même un réglage hyperparamètres, alors on bascule – en termes de compréhension du système – dans une réalité non-appréhendable pour l'humain qui s'apparente, en termes d'opacité, aux modèles *boîtes noires*. Dans notre cas, l'utilité de l'XAI reprend les trois premiers points avancés par Adadi et Berrada, elle sert à la fois à justifier, expliquer et comprendre les résultats, contrôler et valider le processus et possiblement l'améliorer en optimisant certains paramètres. Dans cet objectif, une stratégie d'explicabilité primaire consiste à effectuer une synthèse des données en extrayant les caractéristiques saillantes et les dépendances entre données et clusters pour en faciliter l'intégration et la compréhension par l'utilisateur. En ces termes, cette approche s'apparente à une forme de data storytelling [104, 66] au sens où la finalité consiste à expliquer les données à l'utilisateur et/ou en simplifier l'analyse exploratoire par le biais de méthodes de visualisation graphiques ou de résumés linguistiques de la façon la plus naturelle possible [118, 130]. De tels résumés permettent ainsi de guider la compréhension de l'utilisateur concernant les résultats qui lui sont rendus. L'intérêt principal de ces approches étant qu'elles sont adaptatives

5. https://www.youtube.com/watch?v=hOHrt80HKeM&t=3423s&ab_channel=AssociationEGC

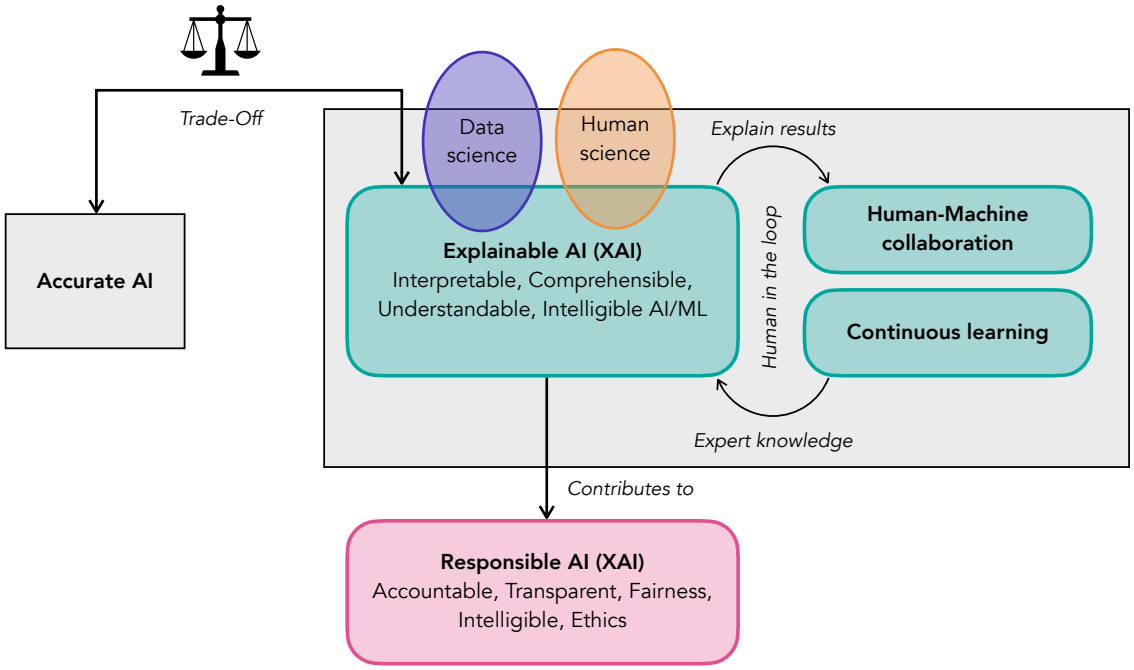


Figure 4.5 – Vue schématique des concepts liés à la l'eXplainable AI inspirée de [2]

selon le besoin de l'utilisateur et ont pour but d'apporter plus de clarté, de confort et de réflexion pour une meilleure prise de décision. Néanmoins, ces résumés sont tenus de rester un outil d'aide à la découverte pour l'utilisateur et ne doivent pas devenir un symbole d'autorité, un substitut complet à sa propre analyse ou pensée. En outre, se pose la question de l'adéquation entre ce qui est compris par l'utilisateur et ce qui est implémenté, voire même du sens réel des données.

En conséquence la question de la présentation des données est centrale dans le processus d'explicabilité car c'est elle qui ici conditionne la réussite de l'approche *human in the loop*. Nous pensons que des approches centrées sur l'humain, relatives aux domaines de la psychologie, de la communication et de l'interaction homme-machine peuvent guider cette problématique. Pourtant, l'examen de la littérature a identifié le manque de travaux se concentrant sur l'impact du facteur humain au sein de l'XAI. À ce sujet, le travail de Miller [160] est, au meilleur de nos connaissances, la tentative la plus significative d'articuler les sciences humaines et l'XAI. En se basant sur un corpus d'articles issus de la philosophie, psychologie et des sciences cognitives, l'auteur souligne trois résultats majeurs pour qu'une explication soit correctement assimilée par l'humain : (i) Les explications doivent être contrastives : les gens ne demandent pas pourquoi l'événement E s'est produit, mais plutôt pourquoi l'événement E s'est produit au lieu d'un certain événement F. (ii) Les explications doivent être sélectives et se concentrent sur une ou deux causes possibles et non sur toutes les causes du résultat. (iii) Les explications doivent s'appuyer sur des exemples prototypiques du monde réel [121]. Ces propositions viennent partiellement appuyer nos précédentes

propositions en matière d'indicateurs pour l'analyse des ensembles de séquences (voir table 4.3).

Ainsi, l'XAI cherche aujourd'hui à construire un pont, un intermédiaire entre les algorithmes efficaces et les connaissances humaines. Notamment, elle vise à construire des systèmes intelligents plus responsables, plus éthiques en incluant l'humain dans la boucle (approche *human in the loop*) et en lui expliquant les résultats fournis, données et modèles construits par les algorithmes. À son tour, l'expert humain doit venir aider la machine en apportant un éclairage contextuel des données et résultats. En outre cette approche peut former un terreau fertile pour la sérendipité et la découverte de motifs métiers intéressants. Néanmoins, cette interaction entre humain et machine pose notamment la question de la présentation et l'assimilation des résultats par l'humaine et nous pensons que la solution à cette problématique doit être trouvée en concours et avec l'expertise de disciplines à la fois de sciences humaines (psychologie, sciences cognitives) et liées à la donnée (data sciences). La figure 4.5 reprend les concepts et idées développés ci-dessus en une vue schématique.

Conclusion sur l'état de l'art

Pour conclure cette partie, nous revenons sur les différents points essentiels soulevés :

1. Nous avons vu que l'identification de comportements de mobilité demande une modélisation préliminaire des activités au cours du temps. Dans ce contexte, la Time-Geography fournit un cadre formel efficace, centré sur l'individu, afin de représenter ses interactions avec l'environnement. Par extension, l'approche dite « activité-centrée » et des trajectoires symboliques visent à étudier les comportements de l'individu en fonction de l'enchaînement temporel (i.e., séquence) de ses actions et/ou lieux visités.

   Nous proposons dans la suite de cette thèse de ré-utiliser la modélisation proposée de Güting et al. des trajectoires symboliques [90], notamment pour sa généricité. De plus, afin de standardiser les connaissances entre experts et non-experts tout en conservant souplesse et adaptabilité, nous proposons d'enrichir cette modélisation en y incorporant une structuration préalable des symboles des séquences (lieux, activités, etc.) en ontologies. Ces ontologies répondent également à la problématique de comparaison des symboles dans les séquences et permettent la gestion de différentes granularités (niveau de détail) sémantique.

2. Sur la base de telles séquences, nous cherchons ensuite à constituer des groupes de séquences similaires. L'élaboration d'une mesure pour quantifier la similarité entre ces séquences suppose de tenir compte à la fois de la sémantique des données dans les séquences (e.g., lieux, actions), de leur caractère temporel (précédence et durée) et de caractéristiques structurelles (cyclicité de symboles similaires, déformations temporelles).

   Les mesures classiques étant incapables de rendre compte de tels besoins, nous situons une partie de nos contributions dans la création de telles mesures efficaces pour la comparaison de séquences d'activités humaines, notamment dans le cadre de la mobilité.

3. L'extraction des comportements prend alors place dans un processus de clustering issu de la mesure précédente. Une fois les clusters (i.e., groupes de séquences) constitués, une phase d'explication doit s'ensuivre afin de les traduire en comportements compréhensibles. Cette explication prend notamment la forme d'un ensemble d'indicateurs visuels en charge de décrire les différentes caractéristiques des clusters de séquences et d'expliquer les résultats à l'utilisateur.

   Dans une volonté de partage et de collaboration entre experts et non-experts, sciences humaines et formelles, nous proposons un cadre méthodologique de travail pour l'analyse et la découverte de comportements dans les séquences de mobilité. Ce cadre de travail orienté vers une approche *human in the loop* s'appuie notamment sur les connaissances établies dans le précédent état de l'art avec nous objectif une intelligibilité maximale des résultats pour l'utilisateur. Il

est concrétisé par une application web, SIMBA, à destination des experts métier pour l'analyse de séquences de mobilité sémantique.

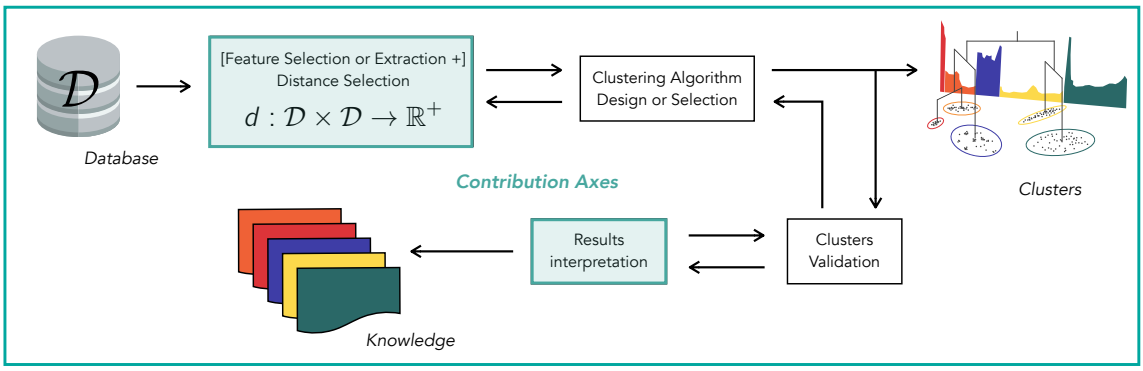


Figure 4.6 – Axes de contribution de la thèse au sein du processus d'extraction de connaissance : La méthodologie d'analyse, les distances et l'interprétation

La partie suivante aborde les contributions de la thèse en matière de découverte de comportements depuis un ensemble de séquences de mobilité sémantique. La figure 4.6 place nos apports relativement à la figure 4.1 dans le processus de clustering et d'extraction de connaissances. Ceux-ci se situent en amont et en aval du processus et portent sur la création de mesures pour caractériser les séquences d'activités humaines et d'un cadre méthodologique pour la découverte de comportements intelligibles. Au-delà de ces contributions, nous proposons également un cadre unifié et global ainsi qu'une adaptation des méthodes classiques de clustering pour la fouille de séquences de mobilité sémantique.

Deuxième partie

Contributions

Chapitre 5
CED : Une mesure pour la comparaison de séquences sémantiques

5.1 Définition du problème

Dans la section 2.3, nous avons vu plusieurs façons de représenter les trajectoires sémantiques. Les modèles présentés tiennent compte des dimensions temporelle et sémantique, voire spatiale. Or, nous avons vu en section 3.3.1 que les mesures dédiées à la comparaison de telles données séquentielles édulcorent en partie la dimension temporelle et ne tiennent compte que de la relation de précédence entre éléments.

Dans le contexte de la mobilité et des activités, cette omission partielle du temps peut néanmoins être intéressante à deux niveaux : (i) D'une part elle permet de donner une vision *compositionnelle* de la mobilité en s'attachant davantage à *ce que font les individus* et à la chronologie des activités plutôt qu'au budget temps associé. (ii) Ensuite, elle permet de mettre au même niveau toutes les activités, là où les approches classiques octroient un poids nettement plus important aux activités intrinsèquement longues (e.g., travail, être à la maison). Or, nous défendons le fait que des activités courtes puissent être prises en compte avec la même importance que des activités plus longues.

Ainsi, nous représentons dans un premier temps les séquences sémantiques comme une suite de symboles ordonnés temporellement. Nous incorporons plus loin la notion de durée dans le chapitre 6. La définition suivante explicite la formalisation des séquences sémantiques.

Définition 1 (Séquence sémantique)**.** *Soit Σ, un ensemble de symboles (lieux, activités, etc.) tel que l'on dispose d'une mesure de similarité $sim : \Sigma \times \Sigma \rightarrow [0, 1]$ (voir section 3.1.2). Une séquence sémantique $S_i \in \Sigma^n$ est une suite ordonnée de symboles telle que $S_i = \langle x_{i1}, ..., x_{in} \rangle$ où $\forall k \in [\![1, n]\!], x_{ik} \in \Sigma$ et $j < k$ indique que x_{ij} s'est déroulé avant x_{ik}. De plus, on considère ici que les symboles ne sont pas répétés de façon consécutive (i.e., $\forall k \in [\![1, n-1]\!], x_{ik} \neq x_{i(k+1)}$).*

Appliquée à la mobilité, une telle séquence traduit que l'on observe d'abord l'activité x_{i1} menée, puis x_{i2}, ..., puis finalement x_{in}.

Nous avons également soulevé dans l'état de l'art que les mesures classiques entre séquences de symboles prennent mal en compte les spécificités des activités humaines. Afin de répondre à cette problématique, nous commençons par définir trois notions de proximités entre activités (i.e., symboles) dans les séquences qu'il faut distinguer.

Définition 2 (Proximité sémantique)**.** *Soit une séquence sémantique* $S_i = \langle x_{i1}, ..., x_{in} \rangle$*. Deux activités* $x_{ij}, x_{ik} \in \Sigma$ *de* S_i *sont proches sémantiquement si elles sont perçues comme semblables. D'un point de vue ontologique, cette proximité* $ProxSem$ *se traduit par l'existence de relation entre ces activités qui peut être mesurée à l'aide des mesures de similarité sémantique présentées section 3.1.2.*

Définition 3 (Proximité temporelle)**.** *Soit une séquence sémantique* $S_i = \langle x_{i1}, ..., x_{in} \rangle$*. Deux activités* $x_{ij}, x_{ik} \in \Sigma$ *de* S_i *sont proches temporellement si elles sont réalisées à des périodes proches. Cette proximité* $ProxTemp$ *peut se mesurer dans une approche continue par une durée* δ *séparant les deux activités si l'on dispose des durées. Dans le cas où la représentation de la dimension temporelle est réduite à un ordonnancement des activités, comme pour les séquences sémantiques, elle se base sur la différence* $|k - j|$ *entre les indices des deux activités.*

Définition 4 (Proximité contextuelle)**.** *Soit une séquence sémantique* $S_i = \langle x_{i1}, ..., x_{in} \rangle$*. Deux activités* $x_{ij}, x_{ik} \in \Sigma$ *de* S *sont proches contextuellement si elles ont à la fois une proximité sémantique et une proximité temporelle forte. Mathématiquement, cette conjonction peut se représenter à l'aide d'une* $\top$*-norme, on a alors :*

$$ProxContext(x_{ij}, x_{ik}) = \top (ProxSem(x_{ij}, x_{ik}), ProxTemp(x_{ij}, x_{ik}))$$

On considère alors 3 séquences sémantiques $S_1 = \langle x_{1.1}, ..., x_{1n} \rangle$, $S_2 = \langle x_{2.1}, ..., x_{2m} \rangle$ et $S_3 = \langle x_{3.1}, ..., x_{3p} \rangle$, et une dissimilarité d sur les séquences sémantiques. Afin de définir de nouvelles mesures qui intègrent les propriétés universelles de la mobilité humaine abordées section 3.2, nous avons traduit ces propriétés en un ensemble de *6 spécificités* $(\mathcal{S}_i)_{i \in \{1...6\}}$ suivantes que d doit tendre à respecter :

($\mathcal{S}_1$). *Homogénéité sémantique* : Deux séquences sémantiques regroupant des activités proches sémantiquement devraient être plus similaires que deux séquences d'activités regroupant des activités sémantiquement différentes. Formellement, on cherche à vérifier une assertion de la forme :

$$\sum_{k=1}^{n} \max_{j \in [\![1,m]\!]} \{ProxSem(x_{1k}, x_{2j})\} \geqslant \sum_{k=1}^{n} \max_{j \in [\![1,p]\!]} \{ProxSem(x_{1k}, x_{3j})\}$$
$$\Rightarrow d(S_1, S_2) \leqslant d(S_1, S_3)$$

($\mathcal{S}_2$). *Temporalité d'activités* : Deux séquences sémantiques ayant les mêmes activités se déroulant à des temporalités proches devraient être plus similaires que deux séquences d'activités ayant les mêmes activités, mais se déroulant à des

temporalités éloignées. Formellement, on cherche à vérifier une assertion de la forme :

$$\sum_{k=1}^{n} ProxTemp(x_{1k}, x_{2j}) \geqslant \sum_{k=1}^{n} ProxTemp(x_{1k}, x_{3k}) \Rightarrow d(S_1, S_2) \leqslant d(S_1, S_3)$$

Par hypothèse qu'ici les séquences sont composées des mêmes activités, on suppose par commodité que $n = m = p$.

($\mathcal{S}_3$). *Décalage temporel* : Deux séquences sémantiques ayant les mêmes activités à un décalage temporel Δt près (en termes d'ordre ou de durée) devraient avoir une dissimilarité d'autant plus faible que le décalage temporel est faible lui-même. Formellement, la dissimilarité entre les deux séquences S_1 et S_2 est proportionnelle au décalage Δt, soit :

$$d(S_1, S_2) \propto \Delta t$$

Cette spécificité est un corollaire de la spécificité précédente ($\mathcal{S}_2$).

($\mathcal{S}_4$). *Permutation d'activités* : Deux séquences sémantiques ayant les mêmes activités à une permutation près, devraient avoir une dissimilarité faible. Cette dissimilarité est d'autant plus faible que les activités permutées sont proches temporellement. Cette spécificité est un corollaire des spécificités ($\mathcal{S}_1$) et ($\mathcal{S}_2$).

($\mathcal{S}_5$). *Redondance d'activités* : Deux séquences sémantiques ayant les mêmes activités à des redondances (i.e., répétitions) près, devraient avoir une dissimilarité faible. Cette spécificité est un corollaire de la spécificité ($\mathcal{S}_1$).

Au meilleur de nos connaissances, aucune des distances abordées en section 3.1.2 ne semble combler les spécificités précédentes. Pour autant, certaines peuvent servir à l'élaboration de distances plus sophistiquées qui pourraient tenir compte des propriétés recherchées. De fait, la distance d'édition est une bonne candidate pour une telle entreprise car elle officie sur des séquences de tailles différentes, est basée sur de l'Optimal Matching, peut tenir compte d'une similarité entre symboles et possède la capacité d'effectuer des permutations (ce qui n'est pas le cas de LCS et DTW de façon structurelle). Remarquons également que les spécificités ($\mathcal{S}_i$) sont définies à l'aide de qualificatifs imprécis comme "proche", "faible", etc. Une approche s'appuyant sur la logique floue semble alors pertinente pour traiter de tels quantificateurs.

Dans la suite de cette section, nous présentons la Contextual Edit Distance, qui généralise la distance d'édition grâce à une nouvelle définition de la fonction de coût appliquée aux opérateurs d'édition. Cette nouvelle mesure repose sur l'utilisation de la logique floue et permet, par construction, une prise en compte globale des spécificités ($\mathcal{S}_i$) décrites.

5.2 La Contextual Edit Distance [1]

Publication

C. Moreau, T. Devogele, L. Etienne, V. Peralta, *A contextual Edit Distance for Semantic Trajectories*, ACM SAC (2020) – Poster

Nous présentons dans cette section les définitions de notre proposition nommée la Contextual Edit Distance (CED) permettant de tenir compte des spécificités énoncées lors de la section précédente. Le formalisme s'appuie sur celui construit par Wagner & Fisher dans [237] pour la distance d'édition. Une seconde section apporte un exemple pilote avec quelques séquences sémantiques afin de mettre à l'épreuve CED et d'expliciter ses propriétés.

5.2.1 Formalisation et définitions

Sur la base des notations établies par Wagner et Fischer dans [237] et détaillées en section 3.3.1, nous proposons une nouvelle définition des opérations d'édition qui tient compte à la fois de la position et de la séquence où l'opération est effectuée.

Définition 5 (Opération d'édition contextuelle)**.** *Une opération d'édition contextuelle e est définie comme un 4-uplet tel que :*

$$e = (op, \mathbf{x}, k_{edit}, S_i) \in \{\texttt{add}, \texttt{del}, \texttt{mod}\} \times \Sigma \times \mathbb{N}^* \times \Sigma^n$$

Où $S_i = \langle x_{i1}, ..., x_{in} \rangle$*. L'intuition de l'opération d'édition contextuelle est que l'on applique l'opération d'édition classique op à l'indice d'édition* k_{edit} *dans la séquence S. Le symbole édité par op est le symbole* $\mathbf{x}$*. Les mots-clés* `add`, `del`, `mod`*, réfèrent aux opérations d'édition classiques telles que :*

- $\texttt{add}(\mathbf{x}, k_{edit}, S_i) \mapsto \langle x_{i1}, ..., x_{i(k_{edit}-1)}, \mathbf{x}, x_{ik_{edit}}, ..., x_{in} \rangle$
- $\texttt{del}(\varepsilon, k_{edit}, S_i) \mapsto \langle x_{i1}, ..., x_{i(k_{edit}-1)}, x_{i(k_{edit}+1)} ..., x_{in} \rangle$
- $\texttt{mod}(\mathbf{x}, k_{edit}, S_i) \mapsto \langle x_{i1}, ..., x_{i(k_{edit}-1)}, \mathbf{x}, x_{i(k_{edit}+1)} ..., x_n \rangle$

On note E l'ensemble des opérations d'édition contextuelle.

Dans le but d'illustrer nos séquences d'activités, nous utilisons des émojis structurés au sein d'une ontologie d'activités représentée figure 5.1.

Exemple 1. *Considérons la séquence* $S_1 = \langle$🎬, 🚃, 🚶, 🛍️, 🚃, ⚽$\rangle$ *et l'opération d'édition contextuelle* $e_1 = (\texttt{mod}, $🚶$, 5, S_1)$*. La séquence résultante de l'application de e est alors* $S_1 = \langle$🎬, 🚃, 🚶, 🛍️, 🚶, ⚽$\rangle$

1. **Avertissement :** *La section suivante, issue de la publication ci-dessus ([164]), vient enrichir certains aspects de l'article originel.*

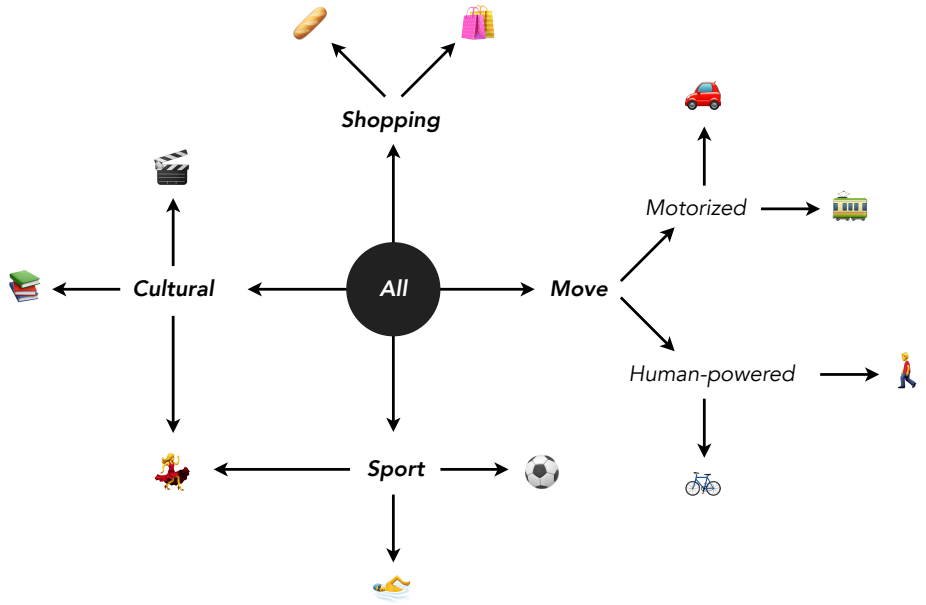


Figure 5.1 – Exemple d'ontologie d'activités

La proximité temporelle est ici basée sur l'écart en nombre de symboles par rapport à l'indice d'édition k_{edit}. Par exemple, si l'on reprend la séquence S_1, pour $k_{edit} = 2$, le symbole 🚃 est relativement proche temporellement du symbole 🛍️, il y a deux symboles d'écart. Plus généralement, cette notion de proximité temporelle peut être étendue afin de traduire le degré de prise en compte du symbole à l'indice i dans l'évaluation de la similarité entre activités. Pour retranscrire cette intuition, nous utilisons la notion de vecteur temporel basée sur des concepts issus de la logique floue.

Définition 6 (Vecteur temporel)**.** *Étant donné une opération d'édition contextuelle* $e = (op, x, k_{edit}, S_i)$*, avec* $|S_i| = n$*, on associe un vecteur temporel* $\nu_e \in [0,1]^n$ *tel que, pour un indice* $k \in [\![1, n]\!]$ *si* $\nu_{e,k} \to 1$*, alors le symbole à l'indice* i *est fortement prix en compte. À l'inverse, si* $\nu_{e,k} \to 0$*, l'indice* k *est d'autant négligé.*

Afin de simplifier la définition du vecteur, il est possible de subordonner son encodage par l'intermédiaire d'une fonction d'appartenance floue $\mu : \mathbb{R} \to [0,1]$*, centrée sur* 0 *et respectant les deux axiomes suivants :*

1. *Le noyau (core) de* μ*,* $C(\mu) = \{t | t \in \mathbb{R}, \mu(t) = 1\}$ *est réduit au seul élément* 0 *(i.e.,* $C(\mu) = \{0\}$*). Intuitivement, l'élément* 0 *sur l'axe des abscisses correspond à la position dans la séquence où est effectuée l'opération d'édition.*
2. *La fonction* μ *est monotone décroissante (au sens large) de part et d'autre de* 0*. Formellement,* $\forall t_1, t_2 \in \mathbb{R}^*, |t_1| \leqslant |t_2| \Rightarrow \mu(t_1) \geqslant \mu(t_2)$*.*

En termes de vocabulaire issu de la logique flou, la fonction μ *s'apparente ici au concept du nombre flou*[2] $\tilde{0}$*.*

2. Un nombre flou $\tilde{n}$ permet de transcrire l'incertitude autour d'une valeur n donnée : intuitivement, plus on s'éloigne de la valeur n, plus l'appartenance à l'ensemble flou correspondant diminue [30].

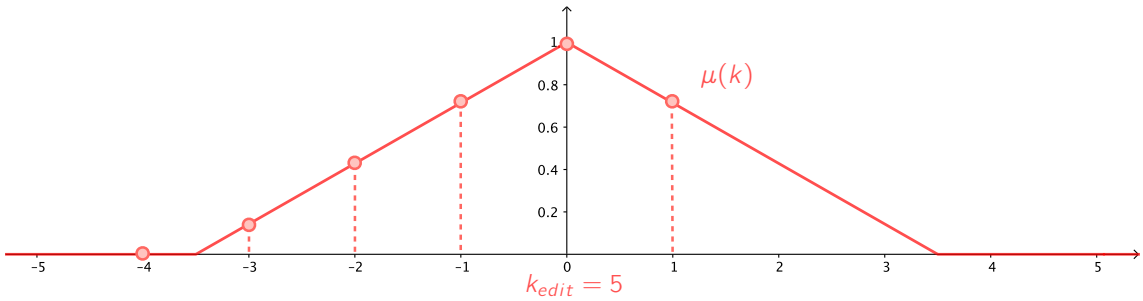


Figure 5.2 – Exemple de fonction floue μ pour l'encodage de l'opérateur `mod`

On note $\Delta k = k - k_{edit}$, la variable représentant l'écart en nombre de symboles entre la position d'édition k_{edit} et l'indice i dans la séquence. Ainsi, en fonction de l'opération d'édition op, le vecteur temporel ν_e est défini comme suit :

- *Si op =* `mod` *:*

$$\nu_{e,k} = \mu(\Delta k)$$

La figure 5.2 montre un exemple de fonction pour l'encodage du vecteur temporel lors d'une opération d'édition de modification. On constate que le symbole à l'indice $k_{edit} = 5$ est pleinement pris en compte (i.e., $\nu_{e,5} = 1$). La prise des symboles x_{ik} diminue d'autant que la quantité $|k - k_{edit}|$ augmente.

- *Si op =* `add` *:*

$$\nu_{e,k} = \begin{cases} \mu(-1) & Si\ \Delta k = 0 \\ \mu(\Delta k + 1) & Si\ \Delta k \geqslant 1 \\ \mu(\Delta k) & Si\ \Delta k \leqslant -1 \end{cases}$$

L'opération d'ajout à l'indice k_{edit} s'effectue entre les symboles aux indices $k_{edit} - 1$ et k_{edit}. Afin de respecter l'axiome 2. de la fonction μ, nous proposons ici de tenir compte de ces deux symboles avec le même degré d'appartenance. En tant qu'opération d'ajout, nous défendons le fait que la prise en compte d'aucun des symboles ne doit être de 100%[3] *; en conséquence nous fixons ici que $\nu_{e,k_{edit}} = \mu(-1)$. Dès lors, l'indexation des éléments aux indices k tels que Δk positif doit être décalée de $+1$. L'indexation des éléments à Δk négatif reste inchangée.*

- *Si op =* `del` *:*

$$\nu_{e,k} = \begin{cases} 0 & Si\ \Delta k = 0 \\ \mu(\Delta k) & Sinon \end{cases}$$

Le cas $\nu_{e,k_{edit}} = 0$ traduit ici le fait que, dans le cas d'une opération de suppression, le symbole à l'indice k_{edit} est virtuellement effacé de la séquence. Ainsi,

3. Ce choix est également motivé par le fait qu'une valeur à 1 dans le vecteur temporel pour les opérations `add` ou `del` serait susceptible de violer l'axiome de séparabilité de la mesure CED présentée définition 9.

son degré de pris en compte est mis à 0. Les autres symboles sont pris en compte de façon classique comme dans le cas d'une opération de modification.

Ainsi, par l'usage d'une telle fonction, on arrive à tenir compte des spécificités ($\mathcal{S}_2$) et ($\mathcal{S}_3$) décrite dans la section précédente.

Exemple 2. *Soit l'opération d'édition contextuelle définie exemple 1. On peut associer le vecteur temporel ν_e par la fonction floue de la figure 5.2 tel que $\nu_e = (0, 0.14, 0.43, 0.71, 1, 0.71)$.*

Grâce aux définitions précédentes nous proposons une nouvelle définition de la fonction de coût d'opération γ de la distance d'édition afin de tenir compte des exigences formulées section 5.1.

Définition 7 (Fonction de coût γ)**.** *Soit une opération d'édition contextuelle $e = (op, x, k_{edit}, S_i)$, avec $|S_i| = n$, la fonction de coût $\gamma : E \rightarrow [0, 1]$ est définie telle que :*

$$\gamma(e) = 1 - \max_{k \in [\![1,n]\!]} \{sim(x_{ik}, \mathbf{x}) \times \nu_{e,k}\} \tag{5.1}$$

L'usage du max dans l'équation 5.1 permet de rendre compte de la spécificité ($\mathcal{S}_1$). Par cet effet, si l'on détecte une activité similaire dans la séquence, le coût est minimisé, proportionnellement à sa distance à l'indice d'édition grâce au vecteur temporel ν_e. Ainsi, le contexte est considéré comme similaire si $\exists k \in [\![1, n]\!]$ tel que $sim(x_{ik}, \mathbf{x}) \times \nu_{e,k} \approx 1$, c'est-à-dire que le symbole édité $\mathbf{x}$ est "temporellement" proche (en nombre de symboles) d'un symbole sémantiquement similaire x_{ik} dans S_i. Dans ce cas, le coût d'édition est faible.

Exemple 3. *On considère l'opération d'édition contextuelle $e_1 = (\texttt{mod}, 🧍, 5, S_1)$ définie dans l'exemple 1. Pour rappel, $S_1 = \langle 🎬, 🚃, 🧍, 👜, 🚃, ⚽ \rangle$. On considère également la similarité de Wu-Palmer définie équation 3.4 et l'ontologie dressée figure 5.1 afin de calculer la similarité entre activités.*

Sur la base de ces éléments et du vecteur temporel défini exemple 2, on peut calculer le coût de l'opération selon l'équation 5.1 :

$$\gamma(e_1) = 1 - \max \left\{ \begin{array}{c} sim(🎬, 🧍) \times 0 \\ sim(🚃, 🧍) \times 0.14 \\ sim(🧍, 🧍) \times 0.43 \\ sim(👜, 🧍) \times 0.71 \\ sim(🚃, 🧍) \times 1 \\ sim(⚽, 🧍) \times 0.71 \end{array} \right\} = 1 - \max \left\{ \begin{array}{c} 0 \times 0 \\ 1/3 \times 0.14 \\ 1 \times 0.43 \\ 0 \times 0.71 \\ 1/3 \times 1 \\ 0 \times 0.71 \end{array} \right\}$$

$$= 1 - 0.43 = 0.57$$

Basé sur l'équation 3.19 de la distance d'édition, il est possible de définir une dissimilarité entre séquences sémantiques.

Définition 8 (Contextual Edit Distance unilatérale)**.** *Soient deux séquences sémantiques* $S_1 = \langle x_{1,1}, ..., x_{1,n} \rangle$ *et* $S_2 = \langle x_{2,1}, ..., x_{2,p} \rangle$*, la Contextual Edit Distance unilatérale* $CED_{S_1 \to S_2} : \Sigma^n \times \Sigma^p \to \mathbb{R}^+$ *est définie telle que :*

$$CED_{S_1 \to S_2} = \min_{(e_1,...,e_N) \in \mathscr{P}(S_1,S_2)} \left\{ \sum_{k=1}^{N} \gamma(e_k) \right\} \tag{5.2}$$

Où γ *est telle que définie équation 5.1.*

Cette équation traduit le coût global pour transformer la séquence S_1 en S_2. Néanmoins, il doit être noté que cette définition de $CED_{S_1 \to S_2}$ sacrifie l'axiome de symétrie.

Lemme 1. $CED_{S_1 \to S_2}$ *respecte l'axiome de séparabilité :* $\forall S_1, S_2 \in \Sigma^*, CED_{S_1 \to S_2} = 0 \Leftrightarrow S_1 = S_2$

Démonstration. On a $CED_{S_1 \to S_2} = 0 \Leftrightarrow S_1 = S_2$ si et seulement si $CED_{S_1 \to S_2} = 0 \Rightarrow S_1 = S_2$ et $CED_{S_1 \to S_2} = 0 \Leftarrow S_1 = S_2$. Démontrons ces deux implications :

- Cas $CED_{S_1 \to S_2} = 0 \Leftarrow S_1 = S_2$.
  Les deux séquences S_1 et S_2 sont identiques. Dès lors, aucune opération n'est nécessaire pour transformer une séquence en l'autre[4]. Il vient trivialement que $CED_{S_1 \to S_2}(S_1, S_2) = 0$.
- Cas $CED_{S_1 \to S_2} = 0 \Rightarrow S_1 = S_2$.
  On raisonne ici par l'absurde. Supposons que $\exists S_1, S_2 \in \Sigma^*$ telles que $S_1 \neq S_2$ et $CED_{S_1 \to S_2} = 0$.
  D'après l'équation 5.2, $CED_{S_1 \to S_2} = 0$ si et seulement si tous les coûts du chemin d'édition minimal $(e_1, ..., e_N)$ sont à 0 (i.e., $\forall k \in [\![1, N]\!], \gamma(e_k) = 0$).
  Or, on a $\gamma(e_k) = 0 \Leftrightarrow \exists j \in [\![1, n]\!]$ tel que $sim(x_{2,j}, x_{1,k_{edit}}) = 1$ et $\nu_{e_k,j} = 1$.
  Or par l'axiome 2 de la fonction d'encodage μ du vecteur temporel de la définition 6, cela ne peut pas arriver dans le cadre d'une opération de modification à l'indice $j = k_{edit}$. Dès lors, on a $sim(x_{2,k_{edit}}, x_{1,k_{edit}}) = 1$ si et seulement si les symboles $x_{2,k_{edit}}$ et $x_{1,k_{edit}}$ sont identiques. De plus, on constate qu'ils sont au même indice k_{edit} dans chacune des séquences. On conclut qu'il n'y a aucune de modification à effectuer.
  Ce raisonnement est valable pour toute opération d'édition contextuelle, e_k, on aboutit au fait que les deux séquences sémantiques S_1 et S_2 sont identiques. L'hypothèse de départ était absurde.

□

4. Plus précisément, on n'effectue que des modifications de coût égal à 0.

Enfin, dans le but de retrouver la symétrie, nous appliquons une $\top$-conorme entre les Contextual Edit Distance unilatérales. Pareillement à la distance de Hausdorff [106], nous proposons une agrégation par l'opérateur max qui fait l'hypothèse la plus pessimiste quant à la similarité des séquences.

Définition 9 (Contextual Edit Distance)**.** *Soient deux séquences sémantiques* $S_1 = \langle x_{1.1}, ..., x_{1n} \rangle$ *et* $S_2 = \langle x_{2.1}, ..., x_{2p} \rangle$*, la Contextual Edit Distance* $CED(S_1, S_2) : \Sigma^n \times \Sigma^p \rightarrow \mathbb{R}^+$ *est définie telle que :*

$$CED(S_1, S_2) = \max\{CED_{S_1 \rightarrow S_2}, CED_{S_2 \rightarrow S_1}\} \tag{5.3}$$

Théorème 1. *(CED,* Σ^**) est un espace semi-métrique.*

Démonstration. Par construction, l'équation 5.3 assure la symétrie. De plus, grâce au lemme 1, on sait que $CED_{S_1 \rightarrow S_2}$ respecte la séparabilité. Il vient alors immédiatement que la séparabilité tient également pour CED ce qui conclut la preuve. □

La définition 7 de la fonction de coût γ permet de résoudre l'ensemble des contraintes que nous avions posées comme la possibilité d'effectuer des répétitions et des permutations à moindre coût ainsi que favoriser l'homogénéité sémantique des séquences. Nous donnons une ébauche des démonstrations pour les propriétés de permutations et répétitions de symboles, l'homogénéité sémantique étant une conséquence directe de l'équation 5.1.

Théorème 2. *Le coût de répétition d'un symbole situé à p symboles d'écart est majoré par* $1 - \mu(p + 1)$*.*

Démonstration. Soient deux séquences sémantiques telles que :

- $S_1 = \langle a \underbrace{*...*}_{p} \rangle$
- $S_2 = \langle a \underbrace{*...*}_{p} a \rangle$

On considère, dans le pire cas, que $sim(a, *) = 0$. On a de façon triviale $CED_{S_1 \rightarrow S_2} = \gamma(e_1)$ et $CED_{S_2 \rightarrow S_1} = \gamma(e_2)$. Ainsi, considérons les opérations d'édition contextuelle suivantes :

- $e_1 = (\texttt{add}, a, p + 2, S_1)$
- $e_2 = (\texttt{del}, \varepsilon, p + 2, S_2)$

Il est clair que le seul indice k susceptible de maximiser la quantité $sim(x_{ik}, a) \times \nu_{e,k}$ de l'équation 5.1 est pour $k = 1$. En effet, $\forall k \neq 1, sim(a, x_{ik}) = 0$, dès lors, on a :

- $\gamma(e_1) = 1 - sim(a, a) \times \nu_{e_1,1}$
- $\gamma(e_2) = 1 - sim(a, a) \times \nu_{e_2,1}$

Dans les deux cas, on a $\Delta k = 1 - (p + 2)$. Il vient que :

- $\gamma(e_1) = 1 - sim(a, a) \times \mu(1 - (p + 2))$
- $\gamma(e_2) = 1 - sim(a, a) \times \mu(1 - (p + 2))$

Or, par l'axiome 2 de la fonction μ d'encodage de la définition 6 et comme $|1 - (p + 2)| = |p + 1|$, on a $\mu(p + 1) = \mu(1 - (p + 2))$ et donc $\gamma(e_1) = \gamma(e_2) = 1 - \mu(p + 1)$.

Comme nous utilisons la mesure de similarité triviale, le résultat fournit une majoration du coût d'édition, il vient que $CED(S_1, S_2) \leqslant 1 - \mu(p + 1)$. □

Théorème 3. *Le coût de permutation de deux éléments situés à p symboles d'écart est majoré par* $2(1 - \mu(p + 1))$.

Démonstration. Soient deux séquences sémantiques telles que :

- $S_1 = \langle a \underbrace{*\ldots*}_{p} b \rangle$
- $S_2 = \langle b \underbrace{*\ldots*}_{p} a \rangle$

On considère en tant que pire cas la mesure de similarité triviale entre symboles telle $sim(x, y) = \begin{cases} 1 & \text{Si } x = y \\ 0 & \text{Sinon} \end{cases}$.

Il est clair ici que l'on cherche à permuter les éléments a et b, les opérations d'édition contextuelle requises pour transformer S_1 en S_2 (resp. S_2 en S_1) sont donc :

- $e_1 = (\texttt{mod}, a, 1, S_2)$ (resp. $e_1 = (\texttt{mod}, b, 1, S_1)$)
- $e_2 = (\texttt{mod}, b, p + 2, S_2)$ (resp. $e_2 = (\texttt{mod}, a, p + 2, S_1)$)

Il vient que $CED(S_1, S_2) = \gamma(e_1) + \gamma(e_2)$. Nous détaillons le premier cas pour transformer S_1 en S_2.

On commence par calculer $\gamma(e_1)$. Par définition de sim et de S_2, l'unique indice k capable de maximiser la quantité $sim(x_k, a) \times \nu_{e_1,i}$ de l'équation 5.1 est $k = p + 2$. Dès lors :

$$\begin{aligned} \gamma(e_1) &= 1 - sim(a, a) \times \nu_{e_1,p+2} \\ &= 1 - 1 \times \mu(p + 2 - 1) \\ &= 1 - \mu(p + 1) \end{aligned}$$

Le raisonnement est analogue pour $\gamma(e_2)$, on a :

$$\begin{aligned} \gamma(e_2) &= 1 - sim(b, b) \times \nu_{e_2,1} \\ &= 1 - 1 \times \mu(1 - (p + 2)) \\ &= 1 - \mu(p + 1) \end{aligned}$$

Dès lors, $CED(S_1, S_2) \leqslant 2(1 - \mu(p + 1))$. □

Les théorèmes 2 et 3 rendent compte de la dualité entre les spécificités $(\mathcal{S}_1)$ et $(\mathcal{S}_2)$. En effet, l'application des deux corollaires (permutation et répétition) de la spécificité d'homogénéité sémantique $(\mathcal{S}_1)$ est conditionnée par la fonction μ qui incarne la spécificité de temporalité $(\mathcal{S}_2)$.

L'équation 5.1 de la fonction de coût γ de CED représente un compromis entre ces deux spécificités, difficilement conciliables. Ainsi, selon la forme de μ, la préférence est donnée à l'une ou l'autre d'après les deux cas suivants :

- Favoriser $(\mathcal{S}_1)$.
  Pour favoriser l'homogénéité sémantique des séquences, alors μ doit être très étalée sur l'ensemble des indices de la séquence et proche de la valeur 1. Le cas théorique extrême [5] est celui où $\forall t, \mu(t) = 1$, dans ce cas toute permutation ou répétition peut être pratiquée avec un coût nul de 0.
- Favoriser $(\mathcal{S}_2)$.
  Pour favoriser la temporalité d'activité, alors μ doit être restreinte autour de 0 afin de tenir compte uniquement d'activité proche temporellement. Le cas extrême est celui où $\mu(t) = \begin{cases} 1 & \text{Si } t = 0 \\ 0 & \text{sinon} \end{cases}$, dans ce cas on peut démontrer que la mesure de CED est équivalente à la distance de Levenshtein [6].

Théorème 4. *Soient deux séquences $S_1 \in \Sigma^n$ et $S_2 \in \Sigma^p$. La complexité temporelle du calcul de $CED(S_1, S_2)$ est en $O(n \times p \times \max\{n, p\})$.*

Démonstration. L'équation 5.2 peut être calculée par l'algorithme de Wagner-Fischer [237] en $O(n \times p)$. Cependant, il faut prendre ici en compte le calcul de la fonction de coût γ. On suppose que le vecteur temporel et la fonction d'encodage μ sont calculés en temps contant $O(1)$. En outre, le calcul de l'équation 5.1 est assuré en la taille de la séquence S qui subit l'opération d'édition contextuelle. Dès lors, il vient que $CED_{S_1 \to S_2}$ a une complexité en $O(n^2 \times p)$, $CED_{S_2 \to S_1}$ en $O(n \times p^2)$ et donc $CED(S_1, S_2)$ a une complexité en $O(n \times p \times \max\{n, p\})$ □

Dans la section suivante, nous appliquons CED sur un jeu de données de quelques séquences afin de montrer l'applicabilité de notre méthode et vérifier expérimentalement dans un environnement contrôlé les spécificités réclamées lors de la construction de la mesure.

5.2.2 Exemple pilote

On considère l'ensemble des 8 séquences sémantiques suivantes :

5. Ce cas est normalement interdit par la définition 2 car il viole l'axiome de séparabilité pour CED, mais il reste théoriquement plausible.

6. On remarquera que la définition de μ correspond à la définition non flou du nombre 0.

- $S_1 = \langle$🚲$\rangle$
- $S_2 = \langle$🚲, 🥖, 🚲$\rangle$
- $S_3 = \langle$🚶, 👜, 🚶$\rangle$
- $S_4 = \langle$⚽, 🚶, 🚃, 🎬$\rangle$
- $S_5 = \langle$🚃, 🎬, 🚶, ⚽$\rangle$
- $S_6 = \langle$🚶, 🏊, 🚃, 🎬$\rangle$
- $S_7 = \langle$🚶, 📚, 🚶, 🎬$\rangle$
- $S_8 = \langle$🚶, 🐓, 📚$\rangle$

Les couleurs font référence aux classes expertes devant être retrouvées en fin de processus de clustering. Les classes se justifient de façon suivante :

- Les séquences de la classe orange sont constituées d'activités d'achat (🥖 et , 👜) et/ou transport humain (🚶 et 🚲). De plus, par les spécificités d'homogénéité sémantique ($\mathcal{S}_1$) et de répétition ($\mathcal{S}_5$), nous pensons que la séquence S_2 est plus proche de la séquence S_1 que S_3 (i.e., $d(S_1, S_2) < d(S_3, S_2)$).
- Les séquences de la classe bleu sont composées d'activités de transport (motorisé (🚃) et humain (🚶)), sportives (🏊 et ⚽) et culturelles (🎬). Les séquences S_4 et S_5 sont composées des mêmes activités mais dans un ordre différent. Nous mettons à l'épreuve ici les spécificités ($\mathcal{S}_2$) de temporalité d'activités et de permutation ($\mathcal{S}_4$). En outre, les séquences S_4 et S_6 ont pour différence d'avoir une transposition sur l'activité sportive et la marche à pied. De mêmes, les séquences S_5 et S_6 sont semblables par leurs activités et par la spécificité de décalage ($\mathcal{S}_3$) sur les activités 🚃, 🎬. Ainsi, par homogénéité sémantique et temporalité d'activité, on doit avoir $d(S_4, S_6) \leqslant d(S_4, S_5)$.
- La paire rouge teste la robustesse d'homogénéité semantique ($\mathcal{S}_1$) vis-à-vis des autres séquences. En particulier, ces deux séquences doivent être proches car composées toutes deux de deux activités artistiques, lire et danser, (📚 et 🐓) et de l'activité marche à pied. On teste alors l'homogénéité sémantique et la capacité à répéter ($\mathcal{S}_5$) l'activité de marche à pied à faible coût de CED.

Sur la base de ces séquences sémantiques, la figure 5.3 représente les dendrogrammes construits à l'aide d'un clustering hiérarchique pour les mesures de CED et de l'Edit distance (ED). La fonction d'encodage du vecteur temporel de CED utilisée est une fonction floue triangulaire semblable à la figure 5.2 mais de pente $1 - \frac{|k_{edit} - i|}{4}$. Les dendrogrammes ont été calculés selon l'algorithme Linkage et le critère d'agrégation de Ward de la bibliothèque Scipy (1.4.1) de Python 3. L'expérimentation peut être retrouvée sur GitHub [7] et est directement opérationnelle depuis le fichier Google colab [8] s'y trouvant.

7. hhttps://github.com/Clement-Moreau-Info/CED
8. https://github.com/Clement-Moreau-Info/CED/blob/main/contextual_edit_distance.ipynb

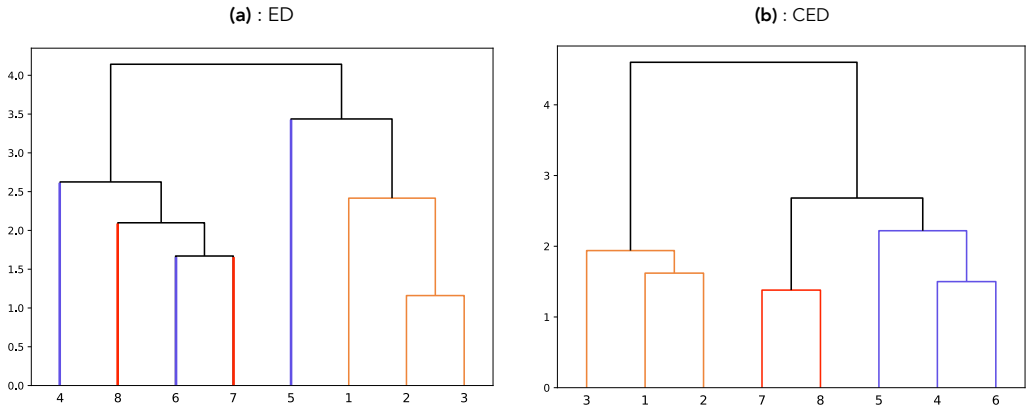


Figure 5.3 – Dendrogrammes des séquences sémantiques pour les mesures (a) Edit Distance (b) Contextual Edit Distance

Nous remarquons que CED arrive à retrouver toutes les classes de séquences avec l'ordre hiérarchique désiré ce qui corrobore nos hypothèses quant à la prise en compte des spécificités voulues. À l'inverse, l'Edit Distance échoue à correctement grouper les classes bleue et rouge. De même, si la classe orange est correctement découverte, l'ordre de groupement n'est pas celui défendu précédemment.

Dans le but de démontrer à la fois l'applicabilité et la généricité de CED pour la découverte de groupes de séquences d'actions humaines semblables, nous présentons dans la section suivante une application dans le cadre de l'extraction de comportements d'analyse de base de données. Les séquences sémantiques prennent alors la forme de séquences d'opérations SQL/OLAP (explorations) extraites depuis des fichiers de logs. Les séquences utilisées dans la section suivante sont labélisées, nous offrant l'occasion de disposer d'une vérité terrain. Ainsi, même si cette application ne relève pas de la mobilité (au sens classique), elle est particulièrement pertinente dans le cadre d'une validation théorique.

D'autres applications et études utilisateur sont décrites dans les chapitres suivants.

5.3 Application à l'exploration de base de données

Publication

C. Moreau, V. Peralta, P. Marcel, C. Chanson, T. Devogele, *Learning Analysis Patterns using a Contextual Edit Distance*, DOLAP Co-located with EDBT (2020)

Dans cette section nous présentons une application de CED pour la découverte de comportements d'exploration de base de données. Cette étude vise à démontrer la généricité de notre mesure. La première sous-section décrit la modélisation des explorations comme séquences sémantiques et les notions préliminaires. La seconde

sous-section est consacrée au clustering d'explorations, protocole et analyse des résultats obtenus.

Toutes les expérimentations, données et code sont disponibles sur notre GitHub [9]

5.3.1 Modélisation d'exploration comme séquence sémantique

L'analyse de la charge de travail (workload) des bases de données possède de nombreux intérêts comme l'optimisation des structures physiques d'accès, la génération automatique de requêtes sur-mesure orientées sur les besoins de l'utilisateur ou bien pour la détection de comportements frauduleux.

Si l'on concentre notre point vue sur les actions de l'utilisateur, il est possible de modéliser son exploration dans la base de données au cours du temps. Cette exploration peut être envisagée métaphoriquement comme une mobilité virtuelle, à travers les données, et donc peut être abstraite comme une séquence sémantique. Ces séquences peuvent être comparées à l'aide de la mesure CED afin de découvrir et comprendre les différents comportements d'exploration / analyse des utilisateurs au sein de la base de données.

Au meilleur de nos connaissances, il existe peu de travaux sur la caractérisation des comportements d'exploration de base de données, notamment car la tâche de caractérisation est complexe [216]. Pourtant, les bases de données décisionnelles (OLAP) sont un outil à la fois très présent et efficace pour l'analyse de données et la prise de décisions. Une base de données décisionnelle possède deux types d'attributs : (i) les dimensions qui permettent d'agréger et de filtrer les données (ex. totaux par mois pour le pays FRANCE) et (ii) les mesures qui forment un ensemble d'indicateurs numériques pour l'analyse (ex. Somme des ventes). Hélas, malgré l'intérêt indéniable que peut avoir l'amélioration des outils et la compréhension de la navigation au sein des bases de données, l'accès aux données d'explorations au sein des entreprises et institutions demeure un frein certain pour l'avancée du domaine. En conséquence, Rizzi et Galinucci proposent dans [197] un générateur d'explorations de base de données décisionnelle, CUBELOAD, implémentant 4 types de comportements d'analyse utilisateur :

- *Slice and Drill.* Un comportement classique suivi par les utilisateurs disposant d'une interface d'aide à la navigation dans les données consiste à explorer les hiérarchies en choisissant un niveau d'agrégation initial, puis en définissant des prédicats logiques de filtrage (sélection) et opérations d'agrégation (drill-down) successifs de plus en plus fins. Ce type de comportement est caractérisé par des opérations de filtrage et d'agrégation répétées.
- *Slice All.* Ce comportement génère des séquences d'opérations de filtrage/non-filtrage correspondant à des utilisateurs souhaitant naviguer dans une base de

9. https://github.com/Clement-Moreau-Info/DOLAP20

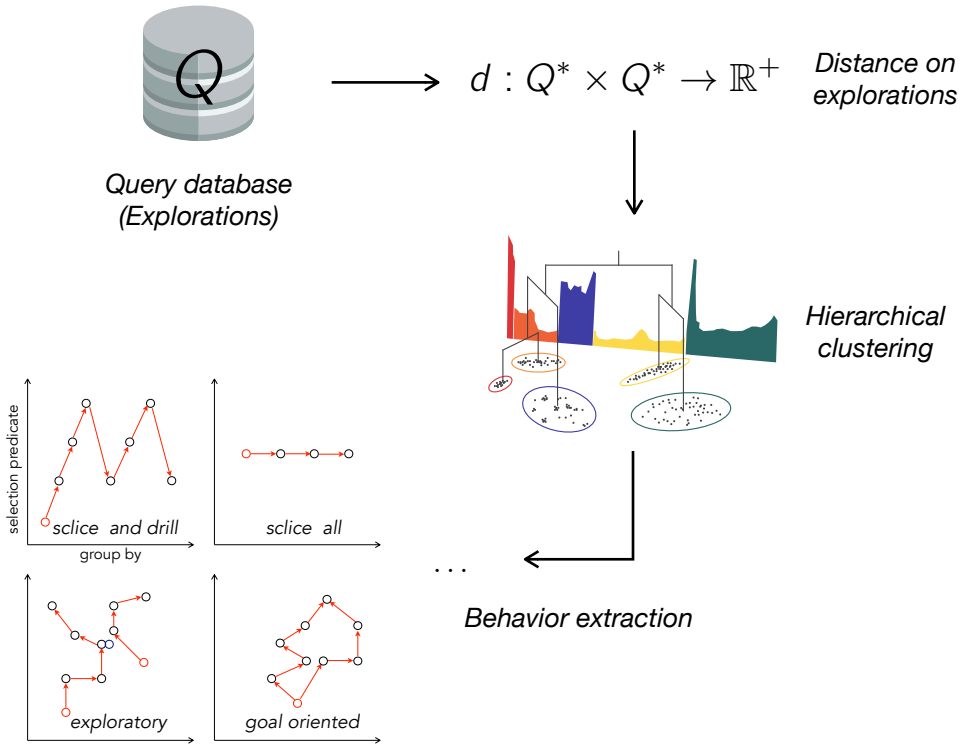


Figure 5.4 – Résumé de l'extraction de comportements dans les explorations. Les comportements extraits reprennent ceux présentés dans [197]

données par tranches, c'est-à-dire en exécutant plusieurs fois la même requête mais avec des conditions de filtrage différents.

- *Exploratory.* Ce comportement est motivé par l'hypothèse selon laquelle un utilisateur, tout en explorant la base de données à la recherche de corrélations significatives, sera "attiré" par une requête surprenante et évoluera ensuite de manière nonchalente à la cherche d'une explication au fait surprenant découvert. Ainsi, les explorations basées sur ce comportement contiennent des opérations aléatoires variées.
- *Goal Oriented.* Les explorations de ce type sont menées par des utilisateurs qui ont un objectif d'analyse spécifique mais dont les compétences d'analyse décisionnelle sont limitées de sorte qu'ils peuvent suivre un chemin complexe pour atteindre leur but. Les explorations liées à ce comportement contiennent des opérations variées, désordonnées mais convergent vers un point spécifique.

Afin de garder le formalisme simple, nous considérons des bases de données décisionnelles décrites par un schéma en étoile et où chaque dimension est décrite comme une hiérarchie unique sans branche $H = (h_1, ..., h_p)$ tel que $\forall i \in [\![1, p-1]\!], h_i \supset h_{i+1}$. Par exemple $H_{time} = (year, month, day)$.

Définition 10 (Requête décisionnelle)**.** *Une requête décisionnelle sur un cube de schéma S est un triplet $q = (G, P, M)$ tel que :*

Id	Feature	Description	Calcul
f_1	NAL	Nombre de niveaux d'agrégation ajoutés.	$\lvert G_k - G_{k-1}\rvert$
f_2	NDL	Nombre de niveaux d'agrégation supprimés.	$\lvert G_{k-1} - G_k\rvert$
f_3	NAF	Nombre de filtres ajoutés.	$\lvert P_k - P_{k-1}\rvert$
f_4	NDF	Nombre de filtres supprimés.	$\lvert P_{k-1} - P_k\rvert$
f_5	NAM	Nombre de mesures ajoutées.	$\lvert M_k - M_{k-1}\rvert$
f_6	NDM	Nombre de mesures supprimées.	$\lvert M_{k-1} - M_k\rvert$
f_7	Adepth	Profondeur d'aggrégation	$\sum_{g \in G_k} depth(g)$
f_8	Fdepth	Profondeur de filtrage.	$\lvert P_k\rvert$

Table 5.1 – Caractéristiques des requêtes décisionnelles

- *G est l'ensemble des niveaux d'agrégation de la requête. On précise que $G \subseteq H$.*
- *P l'ensemble des filtres de la requête.*
- *M l'ensemble des mesures de la requête.*

On note Q, l'ensemble des requêtes décisionnelles.

Ainsi, une exploration est une séquence de requêtes décisionnelles sur un même schéma S et conçue par un même utilisateur dans le but de répondre à un besoin d'information.

Définition 11 (Exploration)**.** *Soit une base de données de schéma S, une exploration est une séquence sémantique $e = \langle q_1, ..., q_n \rangle$ avec $q_k = (G_k, P_k, M_k) \in Q$.*

Cependant, notre approche vise à qualifier les comportements d'exploration sous-jacent des utilisateurs et doit donc s'abstraire du schéma de la base de données utilisée. En conséquence, nous proposons de comparer deux requêtes décisionnelles entre elles selon une représentation dans un espace de caractéristiques proposées par Djedaini et al. [60] et qui est indépendant de la base de données. Les requêtes sont alors traduites en un vecteur où chaque position correspond à une des caractéristiques décrites dans le tableau 5.1. Les 6 premières caractéristiques peuvent être interprétées deux-à-deux et décrites respectivement comme le nombre d'opérations (agrégation, filtres, mesures) ajoutées / supprimées. La caractéristique f_7 décrit la profondeur d'agrégation, c'est-à-dire la granularité d'analyse au sein de la hiérarchie de la dimension. La fonction $depth : H \rightarrow \mathbb{N}$ est définie telle que $depth(g) = \arg_{i \in [\![1,p]\!]}\{g = h_i\}$ et correspond au niveau de la hiérarchie auquel est effectuée l'agrégation. Enfin, la caractéristique f_8 correspond simplement au nombre de filtres utilisés dans la requête.

Définition 12 (Vecteur des caractéristiques)**.** *Soit une requête q_k d'une exploration e donnée. Le vecteur des caractéristiques de q_k noté $v(q_k) = (v_1, ..., v_8)$ est défini tel que $v_i = f_i(q_k, q_{k-1})$. On considère par défaut que $q_0 = (\varnothing, \varnothing, \varnothing)$ et que toute requête avec $k \geqslant 1$ possède un vecteur de caractéristiques.*

Exemple 4. *Soit une exploration e_1 composée des quatre requêtes suivantes :*

$q_1 = \langle \{year\}, \varnothing, \{qty\} \rangle$ – *Quantité des ventes par année ;*
$q_2 = \langle \{year\}, \{year = \text{“2019”}\}, \{qty\} \rangle$ – *Ajout de filtre ;*
$q_3 = \langle \{year, country\}, \varnothing, \{qty\} \rangle$ – *Suppression de filtre, Ajout niveau agrégation ;*
$q_4 = \langle \{year, city\}, \varnothing, \{qty, amount\} \rangle$ – *Ajout niveau agrégation, ajout de mesure ;*

Le vecteur des caractéristiques pour q_1 *est tel que* $v(q_1) = \langle 1, 0, 0, 0, 1, 0, 1, 0 \rangle$ *et indique l'ajout d'un niveau (year) et d'une mesure (qty) par rapport à la requête nulle* q_0*. La profondeur d'agrégation est de 1. Les vecteurs caractéristiques des requêtes* q_2, q_3 *et* q_4*, indiquant l'évolution par rapport à la requête précédente* q_{i-1} *sont tels que :* $v(q_2) = \langle 0, 0, 1, 0, 0, 0, 1, 1 \rangle$*,* $v(q_3) = \langle 1, 0, 0, 1, 0, 0, 2, 0 \rangle$*,* $v(q_4) = \langle 1, 0, 0, 0, 1, 0, 3, 0 \rangle$*.*

Cette représentation se concentre sur les opérations effectuées au long de l'exploration et est au coeur de notre proposition de calcul de la similarité entre requêtes. Ainsi, nous proposons d'utiliser la similarité du cosinus suivante pour le calcul de la similarité entre deux requêtes décisionnelles :

$$\cos(q, q') = \begin{cases} 1 & \text{Si } \|v(q)\| = 0 \text{ et } \|v(q')\| = 0 \\ 0 & \text{Si } \|v(q)\| = 0 \text{ ou } \|v(q')\| = 0 \\ \frac{v(q) \cdot v(q')}{\|v(q)\| \|v(q')\|} & \text{Sinon} \end{cases} \tag{5.4}$$

5.3.2 Clustering d'explorations

Notre objectif est de regrouper les explorations présentant un même comportement d'analyse. À cette fin, nous avons associé la mesure CED à un algorithme de clustering hiérarchique et testé notre approche sur plusieurs workloads concernant des utilisateurs aux compétences analytiques variées et aux interfaces utilisateur différentes, afin de découvrir différents types de comportements d'exploration. Dans un but de concision et de clarté nous présentons ici uniquement les expérimentations menées sur le jeu de données *Artificial* provenant du Star Schema Benchmark [179], utilisé conjointement avec des explorations générées par l'outil CUBELOAD nous fournissant ainsi une vérité terrain quant aux comportements à identifier. Le Star Schema Benchmark (SSB) est une variation de TPC-H, un benchmark populaire du Transaction Processing Performance Council (TPC). Le cube SSB consiste en une base de données relationnelle sous la forme d'un schéma en étoile, avec une table de faits et 4 tables de dimensions.

Pour plus de précisions et informations sur les autres jeux de données testés, nous enjoigons le lecteur attentif à se référer à la publication originelle [167].

Protocole Un algorithme de clustering hiérarchique est en charge de la partie segmentation des explorations. La matrice de distance entre explorations utilisée est basée sur le calcul de la mesure CED combinée à similarité du cosinus entre requêtes telle que définie équation 5.4. À titre de comparaison, nous avons calculé les deux

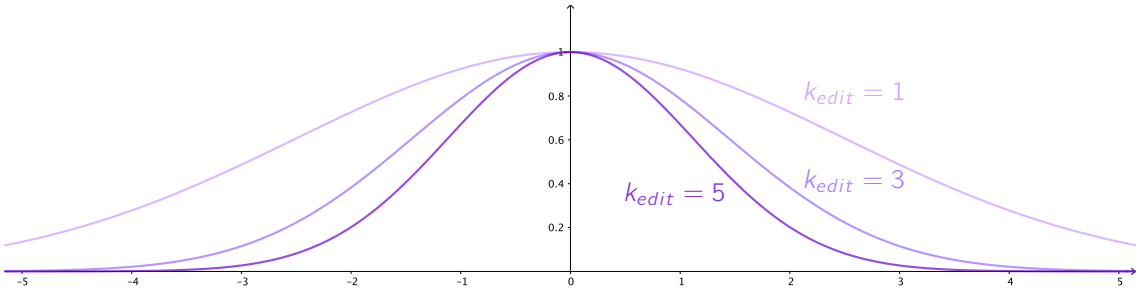


Figure 5.5 – Fonction d'encodage μ du vecteur temporel pour $k_{edit} = 1, 3, 5$ et $|e| = 5$

distances alternatives suivantes : (i) la distance d'édition classique (ED) pourvue de similarité du cosinus entre requêtes comme baseline, et (ii) la distance d'Aligon et al. [9] (AD), une mesure de référence pour le calcul de similarité entre explorations. Grâce au jeu de données *Artificial*, nous pouvons analyser les clusters obtenus en accord avec les modèles comportementaux de CUBELOAD utilisés pour la génération du workload. Cette expérience vise à montrer que notre approche est capable de mieux regrouper les explorations correspondant à un modèle donné que les autres mesures testées. Nous rapportons l'Ajusted Rand Index [193] (ARI) et la V-mesure [10] (moyenne harmonique entre l'homogénéité et la complétude des clusters), et nous comparons nos scores de clustering à ceux obtenus avec les méthodes ED et AD.

Implémentation et paramétrage Les méthodes et algorithmes utilisés sont implémentés en Python 3. Les bibliothèques utilisées pour le clustering hiérarchique et les mesures de qualité de clustering s'appuient sur l'utilisation de SCIPY [11] et SKLEARN [12]. Concernant CED, nous utilisons les paramètres suivants :

- La similarité entre requête est telle que définie équation 5.4.
- La fonction floue servant à l'encodage du vecteur temporel est définie telle que :

$$\mu(k) = \exp\left(-\frac{1}{2}\left(\frac{2k\sqrt{k_{edit}}}{|e|}\right)^2\right)$$

  La figure 5.5 montre le résultat pour quelques valeurs de k_{edit}. Elle a été déterminée empiriquement avec pour exigence de faire varier le coefficient d'applatissement (*kurtosis*) de la courbe autour de la position d'édition k_{edit}. En particulier, lorsque k_{edit} est petit au début de l'exploration, lorsque les intentions de l'utilisateur sont moins définies et que le comportement est plus exploratoire, la courbe de μ est aplatie ce qui permet d'inclure dans le contexte local certaines requêtes qui sont éloignées de l'indice k_{edit}. À l'inverse, quand $k_{edit} \to |e|$,

10. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.v_measure_score.html
11. https://docs.scipy.org/doc/scipy/reference/
12. https://scikit-learn.org/stable/

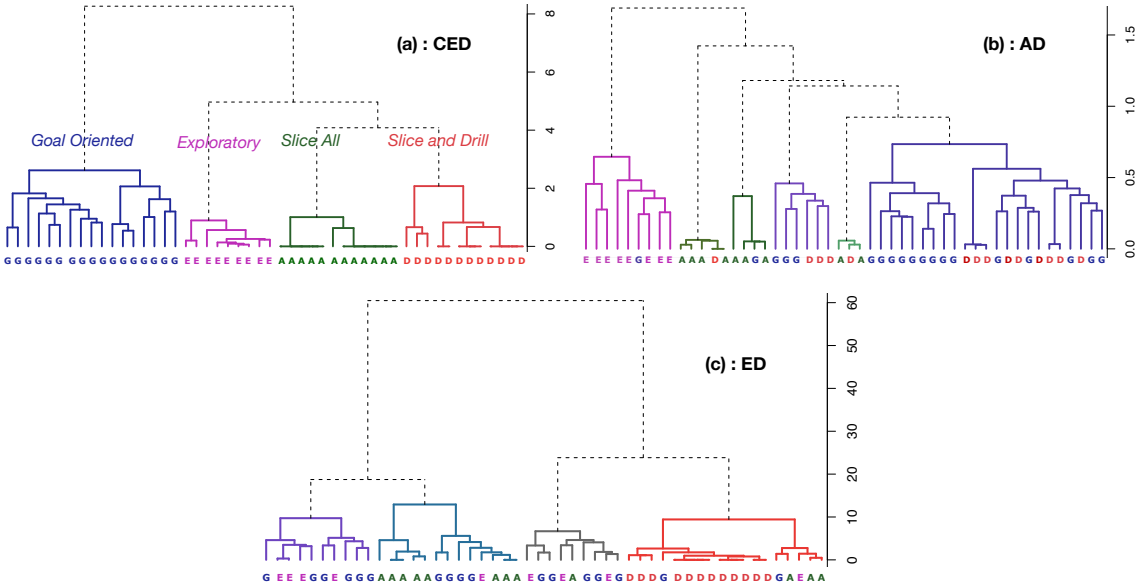


Figure 5.6 – Dendrogrammes sur le jeu de données Artificial selon les trois mesures (a) CED (b) Distance d'Aligon et (c) Distance d'édition

c'est-à-dire se rapproche de la fin de l'exploration, les requêtes convergent vers le but final de l'exploration et donc la courbe se resserre localement.

Analyses et résultats La figure 5.6 montrent les dendrogrammes obtenus à l'aide des trois mesures utilisées. Le dendrogramme de CED exhibe une correspondance parfaite entre les modèles de CUBELOAD (Slice and Drill, Slide All, Exploratory et Goal Oriented) qui constituent des groupes bien différenciés. Ainsi, sur cet exemple, la mesure de CED surpasse celles d'Aligon et al. et de l'Edit Distance. La table 5.2 fournit les scores de qualité pour les trois mesures, on remarque que la V-mesure et l'ARI sont à 1 pour CED, soit le meilleur score possible. Nous nous attendions à de bons résultats avec ce workload car les modèles de CUBELOAD sont bien différenciés. En outre, de nombreuses explorations du modèle Slice All (et certaines du modèle Slice and Drill) sont très similaires (distance proche de 0) car elles contiennent des séquences des mêmes opérations, même si la taille de l'exploration est variable. C'est l'une des caractéristiques qui fait de CED une distance bien adaptée à ce type de problème.

Measure	Nb clusters	ARI	V-measure
CED	4	1	1
AD	6	0.76	0.88
ED	4	0.26	0.36

Table 5.2 – Score de qualité de clustering sur le jeu de données *Artificial* pour les mesures CED, AD et ED

Discussion

Dans ce chapitre, nous avons présenté une des contributions principales de la thèse, la *Contextual Edit Distance* ou CED. CED se base sur la distance d'édition et propose une modification de la fonction de coût d'opération d'édition afin de tenir compte à la fois de la similarité entre symboles et du contenu interne des séquences. Cette nouvelle fonction de coût s'appuie sur la définition d'un vecteur temporel basé sur les principes de la logique floue permettant de retranscrire un degré de prise en compte de chaque symbole des séquences lors du processus de comparaison.

CED permet notamment le respect de caractéristiques déjà soulevées à de nombreuses reprises concernent la mobilité et les activités humaines au cours du temps. De plus, la mesure de CED a été appliquée à de nombreux jeux de données, notamment pour le clustering d'exploration de bases de données décisionnelles où elle a donné de meilleurs résultats que les mesures de référence pour ce type de problème.

Nous pouvons néanmoins souligner quelques limites et pistes d'amélioration quant à CED :

- La fonction floue d'appartenance servant à l'encodage du vecteur temporel est un point critique de CED. En outre, sa définition conditionne plusieurs caractéristiques de la mesure. Une piste de travaux futurs serait l'étude approfondie et la définition de fonctions stéréotypes permettant, selon les voeux de l'utilisateur, l'obtention de propriétés particulières.
- La complexité temporelle de CED en $O(n \times p \times \max\{n, p\})$ est assez limitative pour des calculs sur des volumes de données importants ou séquences longues. Néanmoins, chaque étape de calcul étant parallélisable, il est possible d'avoir recours à des techniques de calcul distribué comme Hadoop ou Spark [93] pour accélérer la construction de la matrice de distance.
- Enfin, CED ne tient compte de la dimension temporelle que par la notion de précédence des symboles. Une amélioration notable serait de pouvoir associer une durée aux symboles afin de traiter des séquences similaires au modèle des trajectoires symboliques de Güting (voir figure 2.5) tout en conservant les propriétés développées dans cette partie.

C'est pourquoi nous proposons dans le prochain chapitre une approche intégrant à la fois la notion de durée au sein des séquences sémantiques mais aussi pourvue d'un temps de calcul plus efficace. Pour ce faire, nous ré-utilisons les concepts développés dans ce chapitre mais dans un cadre de modélisation continue du temps. Enfin, afin de faire chuter la complexité de calcul, nous contraignons la taille des séquences à être fixée selon une durée totale $T_{\max}$ (par exemple 24h) ce qui permet la ré-appropriation de techniques computationnellement moins coûteuses comme la distance de Hamming.

Chapitre 6

FTH : Une mesure pour la comparaison de séquences sémantiques-temporelles

Publication

C. Moreau, T. Devogele, C. de Runz, V. Peralta, E. Moreau, L. Etienne, *A Fuzzy Generalisation of the Hamming Distance for Temporal Sequences*, Fuzz-IEEE (2021)
Best Student Paper Award

6.1 Le temps comme une durée continue

Cette première section décrit une modélisation des séquences sémantiques incorporant la notion de durée. Pour ce faire nous nous ré-approprions partiellement le modèle de Güting et al. [90] décrit en section 2.3. Enfin, nous développons les aspects liés à la complexité et les manques actuels pour la prise en compte du temps des mesures classiques (voir section 3.3.1) et les pistes d'amélioration.

6.1.1 Définitions préliminaires

Bien qu'étant un concept fondamental du temps, incontournable dans les modèles de trajectoires sémantiques (voir section 2.3), la notion de durée peine, au meilleur de nos connaissances, à être prise en compte de manière satisfaisante dans les processus de comparaison de séquences symboliques.

Afin d'illustrer les manques et problèmes actuels posés par l'incorporation de la durée dans les séquences sémantiques, nous proposons la formalisation suivante s'appuyant sur le paradigme développé par Güting et al. dans [90] et généralisant la Définition 1 des séquences sémantiques donnée au chapitre précédent. Dès lors, dans la suite de ce chapitre, nous considérons des séquences dont la durée globale (notée T_{max}) est identique.

Définition 13 (Séquence sémantique-temporelle)**.** *Soit Σ, un ensemble de symboles tel que l'on dispose d'une mesure de similarité $sim : \Sigma \times \Sigma \to [0, 1]$ et soit $I = [0, T_{\max}[$ avec $T_{\max} > 0$, l'intervalle temporel sur lequel les séquences sémantiques-temporelles sont décrites.*

Une séquence sémantique-temporelle S_i est une suite ordonnée de symboles chronométrés telle que :

$$S_i = \langle (x_{i1}, \delta_{i1}), ..., (x_{in}, \delta_{in}) \rangle$$

où $\forall k \in [\![1, n]\!], x_{ik} \in \Sigma$ et $\delta_{ik} > 0$ telle que δ_{ik} indique la durée de l'activité x_{ik} (selon une unité de temps convenue e.g., minute). De plus, S_i respecte les deux propriétés suivantes :

- *Il n'existe pas de symbole répété de façon consécutive i.e., $\forall k \in [\![1, n-1]\!], x_{ik} \neq x_{i(k+1)}$.*
- *La somme de toutes les durées est égale à $T_{\max}$ i.e., $\sum_{k=1}^{n} \delta_{ik} = T_{\max}$.*

On note $\mathbb{S}^n$ l'ensemble des séquences sémantiques-temporelles composées de n symboles.

Appliquée à la mobilité, une telle séquence traduit le fait que l'on observe d'abord l'activité x_{i1} menée pendant δ_{i1} unités de temps, puis x_{i2} pendant δ_{i2} unités, ..., puis finalement x_{in} pendant δ_{in} unités de temps.

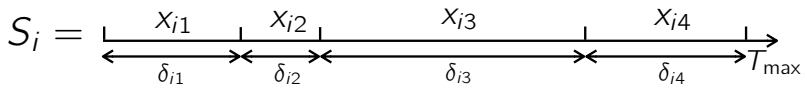


Figure 6.1 – Abstraction d'une séquence sémantique-temporelle

La figure 6.1 ci-dessus fournit une représentation d'une séquence sémantique-temporelle et des concepts clés en jeu dans la définition 13.

Afin d'illustrer nos différentes propositions, la figure 6.2 énumère 7 activités et leur similarité par paire. De plus, nous considérons une valeur de $T_{\max}$ représentant une journée, soit 1440 minutes.

Exemple 5. *On représente les activités quotidiennes d'Alice grâce au formalisme des séquences sémantiques-temporelles présenté précédemment :*

$S_1 = \langle$(🏠, 210), (🚋, 20), (🚶, 10), (💼, 250), (🚶, 15),
(🍽, 60), (🚌, 15), (🏠💼, 290), (🏠, 570)$\rangle$

En accord avec les emojis de la figure 6.2, la séquence S_1 représente la mobilité quotidienne suivante : "Alice est restée chez elle (🏠) durant 210min, puis a effectué un trajet en tramway (🚋) pendant 20min avant de marcher (🚶) pendant 10min. Elle a travaillé (💼) à son bureau durant 250min, marché pendant 15min avant d'aller

déjeuner au restaurant (🍽) pendant 60min. Alice est ensuite rentrée chez elle en effectuant un trajet en bus (🚌) de 15min puis a travaillé à son domicile (🏠💼) pendant 290min. Enfin, elle est restée à son domicile pour le reste de la journée, soit 570min."

Afin de modéliser les intervalles temporels des activités dans les séquences, nous considérons la définition suivante :

Définition 14 (Intervalle temporel)**.** *Soit une séquence sémantique-temporelle S_i, l'intervalle temporel $I(x_{ik})$ du symbole x_{ik} est défini tel que :*

$$I(x_{ik}) = [deb(x_{ik}), fin(x_{ik})[$$

où :

- $deb(x_{ik}) = \sum_{j=1}^{k-1} \delta_{ij}$
- $fin(x_{ik}) = \sum_{j=1}^{k} \delta_{ij} = deb(x_{ik}) + \delta_{ik}$

De plus, pour $k_1 \neq k_2$, on a $I(x_{ik_1}) \cap I(x_{ik_2}) = \varnothing$ et $\bigcup_{k=1}^{n} I(x_{ik}) = I$.

Dans la sous-section suivante, nous décrivons succinctement les inconvénients structurels et limites des mesures classiques pour traiter ce type de séquences sémantiques-temporelles.

6.1.2 Limites des mesures actuelles pour la comparaison de séquences sémantiques-temporelles

Nous avons déjà présenté dans les sections 3.3.1 et 3.3.2 quelques mesures permettant la comparaison des séquences symboliques ; les plus classiques dans le domaine de la mobilité étant la distance de Hamming, LCS, DTW et ED. La table 6.1 reprend la complexité en temps de calcul de celles-ci sur des séquences sémantiques et séquences sémantiques-temporelles. Notons que les séquences sémantiques-temporelles sont définies sur une taille fixe T_{max} et donc possèdent la même longueur.

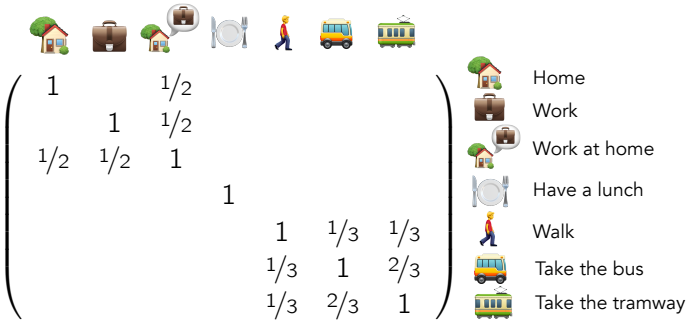


Figure 6.2 – Similarité entre des symboles sémantiques de Σ. Les cellules vides indiquent une similarité égale à 0.

Méthodes	Complexité temporelle	
	Seq. sémantiques	Seq. sémantiques-temporelles
Hamming	$O(n)$	$O(T_{\max})$
LCSS, DTW, ED	$O(n \times p)$	$O(T_{\max}^2)$
CED	$O(n \times p \times \max(n,p))$	$O(T_{\max}^3)$

Table 6.1 – Complexité temporelle des mesures classiques pour la comparaison de séquences sémantiques / sémantiques-temporelles

Conçues initialement pour des séquences d'éléments discrets, ces mesures peinent à s'accommoder de concepts intrinsèquement continus comme le temps et la durée. Ainsi, la solution la plus souvent envisagée pour retranscrire un semblant de durée dans les séquences consiste en la répétition consécutive des symboles. Par exemple, la séquence $\langle (X,3),(Y,1),(Z,2)\rangle$ sera discrétisée en $\langle X,X,X,Y,Z,Z\rangle$. Cet artifice nous semble cependant peu satisfaisant et pose plusieurs questions et problèmes en termes d'efficacité de représentation et de calcul, notamment du fait de la granularité. En effet, pour des séquences sémantiques-temporelles définies sur un grand intervalle de temps (par exemple, plusieurs jours) et une petite unité de temps (la seconde), les temps de calcul deviennent prohibitifs.

Ainsi, une mesure idéale sur les séquences sémantiques-temporelles devrait pouvoir reprendre les propriétés avantageuses développées par CED (en particulier le respect des spécificités abordées section 5.1), modéliser les durées d'activités de façon continue tout en réduisant la complexité temporelle de CED et maintenant un temps de calcul compétitif par rapport aux mesures existantes de l'état de l'art. Dans cet objectif, et comme les séquences sémantiques-temporelles sont définies sur un intervalle de temps fixe, nous proposons dans la suite de ce chapitre d'adapter la distance de Hamming dont la complexité de calcul est la plus faible afin de combler les lacunes précédemment soulevées, tout en respectant les différentes exigences présentées section 3.2.

6.2 Une approche floue de la distance de Hamming pour les séquences sémantiques-temporelles

Cette section présente une approche floue de la distance de Hamming afin de prendre en compte un voisinage temporel lors du processus de comparaison de séquences. En particulier, nous introduisons ici les concepts d'opération d'édition qui définit la transformation d'une partie d'une séquence, et celle de fonction de contexte floue qui permet de quantifier une période imprécise de temps autour de ladite opération d'édition. Enfin, une fonction de coût formant le coeur de notre proposition vient quantifier l'impact de l'opération d'édition sur la séquence. Finalement, la Fuzzy temporal Hamming distance est définie comme la somme maximale des coûts pour transformer une séquence sémantique-temporelle en une autre.

6.2.1 Formalisation des opérations d'édition

La distance de Hamming classique consiste à transformer une séquence donnée S_i en une autre séquence S_j par des modifications successives des parties (i.e., symboles) de S_i. La distance finale est alors définie, selon le modèle de l'équation 3.18, comme la somme des coûts de transformation.

Tout comme dans le chapitre précédent, nous redéfinissons le concept d'opération d'édition dans le cadre de la distance de Hamming afin de prendre en compte la notion temporelle de durée.

Définition 15. *Une opération d'édition e est définie comme un 4-uplet tel que :*

$$e = (\mathbf{x}, \delta, t_{edit}, S_i) \in \Sigma \times I \times I \times \mathbb{S}^n$$

Ainsi, on remplace par $\mathbf{x}$ *tous les symboles de* S_i *à partir de l'instant* t_{edit}*, ceci pour* δ *unités de temps.*

On note $\mathbb{E}$ *l'ensemble des opérations d'édition.*

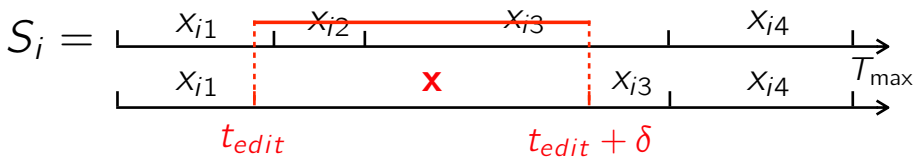


Figure 6.3 – Opération d'édition sur une séquence sémantique-temporelle. On remplace tous les symboles de t_{edit} à $t_{edit} + \delta$ dans S_i par $\mathbf{x}$

La figure 6.3 représente l'abstraction de l'opération d'édition donnée définition 15.

Définition 16 (Fonction temporelle)**.** *Soit une opération d'édition* $e = \langle \mathbf{x}, \delta, t_{edit}, S_i \rangle$*, la fonction temporelle* $\mu_e : I \rightarrow [0, 1]$ *est une fonction d'appartenance définie selon l'opération d'édition e telle que le noyau (core)* $C(\mu_e) = \{t | t \in I, \mu_e(t) = 1\} = [t_{edit}, t_{edit} + \delta[$ *et dotée d'une frontière floue (boundarie)* $B(\mu_e) = \{t | t \in I, 0 < \mu_e(t) < 1\} =]t_{edit} - \beta, t_{edit}[\cup [t_{edit} + \delta, t_{edit} + \delta + \beta[$ *où* $\beta \geqslant 0$*.*

Intuitivement, la fonction temporelle est utilisée pour quantifier l'emprise temporelle d'une opération d'édition sur la séquence S_i. Ainsi, l'emprise est égale à 1 entre t_{edit} et $t_{edit} + \delta$, puis décroît selon le paramètre β de part et d'autre de l'intervalle.

Définition 17 (Similarité sur S_i)**.** *Soit une opération d'édition* $e = \langle \mathbf{x}, \delta, t_{edit}, S_i \rangle$*, la similarité sur la séquence* S_i*, nommée* $sim_e : I \rightarrow [0, 1]$ *est définie telle que :*

$$sim_e(t) = \sum_{k=1}^{n} \mathbf{1}_{I(x_{ik})}(t) \times sim(x_{ik}, \mathbf{x}) \tag{6.1}$$

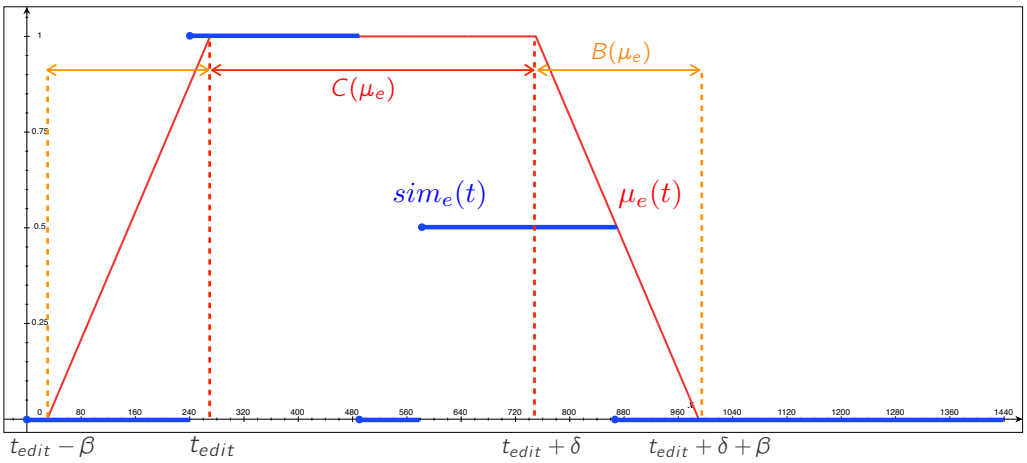


Figure 6.4 – Exemple de fonctions μ_e et sim_e

Où $\mathbf{1}_A$ *est la fonction indicatrice sur* I *pour le sous-ensemble* $A \subseteq I$ *et est définie telle que* $\mathbf{1}_A(t) = \begin{cases} 1 & si\ t \in A \\ 0 & sinon \end{cases}$.

Le concept clé de l'équation 6.1 est que, à l'instant t, nous calculons la similarité entre $\mathbf{x}$ et le symbole x_{ik} qui se produit à cet instant i.e., pour $t \in I(x_{ik})$. Il en résulte donc une fonction étagée représentant une combinaison linéaire de fonctions indicatrices pondérées par la similarité de ces symboles sur l'intervalle de temps I.

Exemple 6. *On considère l'opération d'édition* $e = (\mathbf{x} = $ 💼$, \delta = 480, t_{edit} = 270, S_i = S_1)$ *où* S_1 *est la séquence décrite dans l'exemple 5. La figure 6.4 montre les fonctions* μ_e *et* sim_e *pour l'opération d'édition* e.

La fonction rouge décrit la fonction temporelle μ_e *pour une frontière floue de deux heures (*$\beta = 240min$*).*

La fonction bleue représente quant à elle la similarité sur S_1 *de l'opération d'édition* e *avec* $\mathbf{x} =$ 💼*. En accord avec la matrice de similarité décrite figure 6.2, on remarque alors que* $sim_e(t)$ *est égale à 1 pour* $t \in [240, 490[$ *quand Alice est à son bureau (*💼*), à* $1/2$ *pour* $t \in [580, 870[$ *lorsqu'elle travaille à son domicile (*🏠💼*) et 0 sinon.*

6.2.2 Fonction de coût d'opération d'édition

Grâce aux définitions précédentes, il est possible, tout comme pour CED en section 5.2, de mettre au point une fonction qui quantifie le coût d'application d'une opération d'édition.

Définition 18 (Fonction de coût normalisée γ)**.** *Soit une opération d'édition* $e = \langle \mathbf{x}, \delta, t_{edit}, S_i \rangle$*, la fonction* $\gamma : \mathbb{E} \rightarrow [0, 1]$ *est la fonction de coût normalisée de l'application de l'opération d'édition* e*. Celle-ci est définie telle que :*

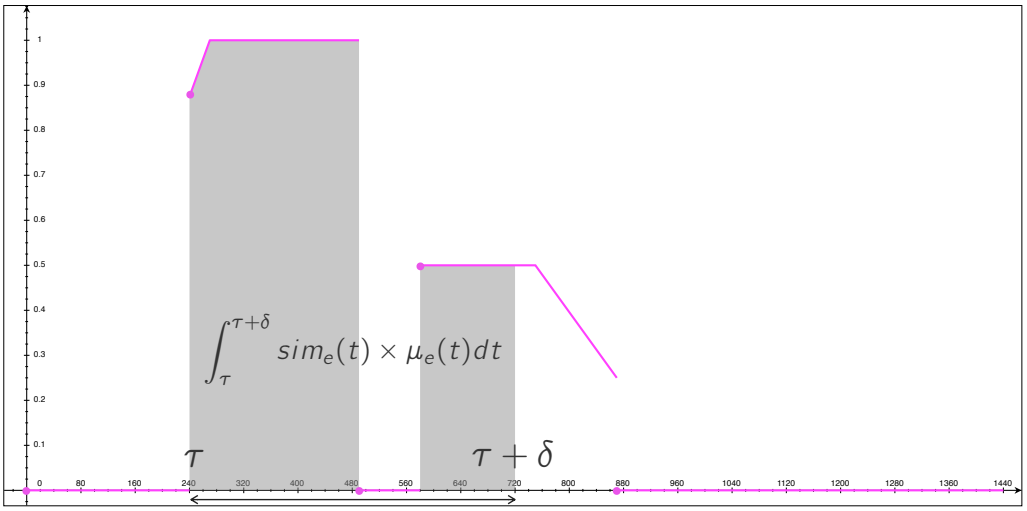


Figure 6.5 – Application de $\gamma(e)$

$$\gamma(e) = 1 - \sup_{\tau \in I} \left\{ \frac{1}{\delta} \int_{\tau}^{\tau+\delta} sim_e(t) \times \mu_e(t)dt \right\} \tag{6.2}$$

L'équation 6.2 est fortement inspirée de la fonction de coût d'édition définie équation 5.1 pour CED.

Derechef, le contexte est ici considéré comme similaire si l'activité éditée **x** est temporellement proche d'une activité similaire dans S_i i.e., $sim_e(t) \times \mu_e(t) \approx 1$. Par conséquent, étant donné une opération d'édition e, l'idée clé est de rechercher le segment temporel $[\tau, \tau + \delta[$ sur I qui maximise à la fois la similarité du symbole édité **x** et la fonction temporelle.

Du point de vue computationnel, l'équation 6.2 est équivalente au calcul du supremum du produit de convolution entre $sim_e(t) \times \mu_e(t)$ et la fonction $\mathbf{1}_{[0,\delta[}$ ce qui peut être calculé efficacement en $O(T_{\max} \log T_{\max})$ à l'aide d'algorithmes de transformation de Fourier rapide (Fast Fourier Transform) [122].

Exemple 7. *Soit l'opération d'édition présentée dans l'exemple 6. La figure 6.5 présente l'application de la fonction γ pour $e = (💼, 480, 270, S_1)$. La borne sup de l'intégrale équation 6.2 est atteinte pour $\tau = 240$ ainsi, on a $\int_{240}^{720} sim_e(t) \times \mu_e(t)dt = 318.21$. Au final, $\gamma(e) = 1 - \frac{318.21}{480} = 0.34$.*

Lemme 2. *Soit une opération d'édition $e = (\mathbf{x}, \delta, t_{edit}, S_i)$, on a $\gamma(e) = 0 \Leftrightarrow \exists k \in [\![1, n]\!], [t_{edit}, t_{edit} + \delta[\subseteq I(x_{ik})$ et tel que $x_{ik} = \mathbf{x}$*

Démonstration. On a :

$$\gamma(e) = 0 \Leftrightarrow \exists \tau \in I, \int_{\tau}^{\tau+\delta} sim_e(t) \times \mu_e(t)dt = \delta$$
$$\Leftrightarrow \exists \tau \in I, \forall t \in]\tau, \tau + \delta[, \mu_e(t) = 1, sim_e(t) = 1$$

Par la définition 16, on a $\mu_e(t) = 1 \Leftrightarrow t \in [t_{edit}, t_{edit}+\delta[$. Ainsi, on sait que $\tau = t_{edit}$. De façon similaire, $sim_e(t) = 1 \Leftrightarrow \exists k \in [\![1, n]\!], t \in I(x_{ik})$ et $x_{ik} = \mathbf{x}$ ce qui conclut la preuve. □

Lemme 3. *Soit une opération d'édition* $e = (\mathbf{x}, \delta, t_{edit}, S_i)$. *On a :* $\lim_{\delta \to 0} \gamma(e) = 1 - \sup_{\tau \in I}\{sim_e(\tau) \times \mu_e(\tau)\}$.

Démonstration. Afin de simplifier les notations, on note : $\Phi(t) = \int sim_e(t) \times \mu_e(t)dt$. Il vient immédiatement que la dérivée $\Phi'(t) = sim_e(t) \times \mu_e(t)$.

Ainsi, on peut ré-écrire $\gamma(e)$ telle que :

$$\gamma(e) = 1 - \sup_{\tau \in I} \left\{ \frac{\Phi(\tau + \delta) - \Phi(\tau)}{\delta} \right\}$$

Or, on rappelle ici que la dérivée d'une fonction à valeurs réelles f décrite sur un intervalle non vide peut être définie comme la limite du taux d'accroissement telle que :

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Dès lors, il vient que :

$$\begin{aligned}
\lim_{\delta \to 0} \gamma(e) &= \lim_{\delta \to 0} \left(1 - \sup_{\tau \in I} \left\{ \frac{\Phi(\tau + \delta) - \Phi(\tau)}{\delta} \right\} \right) \\
&= 1 - \sup_{\tau \in I} \left\{ \lim_{\delta \to 0} \frac{\Phi(\tau + \delta) - \Phi(\tau)}{\delta} \right\} \\
&= 1 - \sup_{\tau \in I} \{\Phi'(\tau)\} \\
&= 1 - \sup_{\tau \in I} \{sim_e(\tau) \times \mu_e(\tau)\}
\end{aligned}$$

□

La propriété soulignée dans le lemme 3, et qui concerne la fonction de coût γ, montre que même les symboles dont la durée est infinitésimale peuvent avoir un coût d'édition important dans le processus d'édition de la séquence sémantique-temporelle. En effet, selon certains besoins métiers, nous affirmons que les symboles de courte durée sont aussi importants que ceux de longue durée et donc que le coût doit être à la fois normalisé dans $[0, 1]$ et non-nécessairement proportionnel à la durée δ. Ce cas d'usage est particulièrement vrai lorsque la durée des symboles est fortement déséquilibrée, comme dans le contexte de la mobilité où peu d'activités concentrent une grande partie du temps au cours d'une journée (typiquement, être à la maison ou au travail).

Cependant, la plupart des dissimilarités discutées dans les sections 3.3 et 3.3.2 donnent un poids élevé aux activités de longue durée et, inversement, les activités courtes ont un poids négligeable par rapport à ces dernières. Conscients que cet usage dépend des besoins métiers, nous proposons également une fonction de coût proportionnelle à la durée du symbole édité (i.e., non normalisée dans $[0, 1]$).

Définition 19 (Fonction de coût pondérée par la durée Δ)**.** *Soit une opération d'édition $e = \langle \mathbf{x}, \delta, t_{edit}, S_i \rangle$, la fonction $\Delta : \mathbb{E} \rightarrow [0, 1]$ est la fonction de coût pondérée par la durée du symbole édité par e. Celle-ci est définie telle que :*

$$\Delta(e) = \delta \times \gamma(e) \tag{6.3}$$

Grâce à cette pondération la fonction de coût Δ respecte les propriétés suivantes.

Lemme 4. *Pour toute opération d'édition $e \in \mathbb{E}, 0 \leqslant \Delta(e) \leqslant \delta$.*

Démonstration. On sait que $\forall e \in \mathbb{E}, 0 \leqslant \gamma(e) \leqslant 1$. Dès lors, multiplier par $\delta > 0$ revient à dilater l'intervalle par ce facteur tel que : $0 \leqslant \delta \times \gamma(e) \leqslant \delta \Leftrightarrow 0 \leqslant \Delta(e) \leqslant \delta$. □

Théorème 5. *Δ est une fonction monotone croissante selon la durée δ i.e., soient deux opérations d'édition $e = \langle \mathbf{x}, \delta, t_{edit}, S_i \rangle$ et $e' = \langle \mathbf{x}, \delta', t_{edit}, S_i \rangle$ telles que $\delta < \delta'$, alors $\Delta(e) \leqslant \Delta(e')$.*

Démonstration. Afin de simplifier les notations, on note $\varphi(t) = sim_e(t) \times \mu_e(t)$. De plus, on précise que φ est Riemann-intégrable et on note $\Phi'(t) = \varphi(t)$.

On dérive la fonction $\Delta(e)$ par rapport à δ. On obtient :

$$\begin{aligned}
\frac{\partial \Delta}{\partial \delta}(e) &= \frac{\partial}{\partial \delta}(\delta \times \gamma(e)) \\
&= \frac{\partial}{\partial \delta}\left(\delta - \sup_{\tau \in I}\left\{\int_{\tau}^{\tau+\delta} \varphi(t)dt\right\}\right) \\
&= \frac{\partial}{\partial \delta}\left(\delta - \sup_{\tau \in I}\{\Phi(\tau + \delta) - \Phi(\tau)\}\right) \\
&= 1 - \sup_{\tau \in I}\{\varphi(\tau + \delta) - \varphi(\tau)\}
\end{aligned}$$

Par les définitions 16 et 17, on sait que $\forall \tau \in I, 0 \leqslant \varphi(\tau) \leqslant 1$. Dès lors, on a :

$$\begin{aligned}
-1 \leqslant \varphi(\tau) - \varphi(\tau + \delta) \leqslant 1 \\
-1 \leqslant \sup_{\tau \in I}\{\varphi(\tau) - \varphi(\tau + \delta)\} \leqslant 1 \\
1 \geqslant -\sup_{\tau \in I}\{\varphi(\tau) - \varphi(\tau + \delta)\} \geqslant -1 \\
2 \geqslant 1 - \sup_{\tau \in I}\{\varphi(\tau) - \varphi(\tau + \delta)\} \geqslant 0
\end{aligned}$$

Dès lors, on a $\frac{\partial \Delta}{\partial \delta}(e) \geqslant 0$. On en conclut que Δ est monotone croissante par rapport à δ. □

Le théorème 5 reflète le fait intuitif que, dans les mêmes conditions d'édition, plus une activité longue est éditée, plus le coût d'édition par la fonction Δ est élevé.

6.2.3 Fuzzy Temporal Hamming distance entre séquences sémantiques-temporelles

Grâce aux fonctions de coût précédentes, nous pouvons concevoir une dissimilarité entre deux séquences sémantiques-temporelles.

Définition 20 (Fuzzy Temporal Hamming unilatérale)**.** *Soient deux séquences sémantiques-temporelles* $S_1 = \langle (x_{1.1}, \delta_{1.1}, ..., (x_{1n}, \delta_{1n}) \rangle$ *et* $S_2 = \langle (x_{2.1}, \delta_{2.1}), ..., (x_{2p}, \delta_{2p}) \rangle$*. On considère également une fonction de coût* $f : \mathbb{E} \to \mathbb{R}^+$ *(par exemple* γ *ou* Δ*). La dissimilarité unilatérale Fuzzy Temporal Hamming depuis* S_1 *vers* S_2*,* $\text{FTH}_{S_1 \to S_2} : \mathbb{S}^n \times \mathbb{S}^p \to \mathbb{R}^+$ *est définie telle que :*

$$\text{FTH}_{S_1 \to S_2} = \sum_{k=1}^{n} f(e_k) \tag{6.4}$$

Où $e_k = (x_{1k}, \delta_{1k}, deb(x_{1k}), S_2)$*.*

L'équation 6.4 représente le coût total pour transformer S_1 en S_2. Ainsi, il doit être noté que $\text{FTH}_{S_1 \to S_2}$ est non symétrique.

Exemple 8. *On représente la séquence d'activités quotidiennes de Bob de la façon suivante :*

$$S_2 = \langle (🏠, 230), (🚶, 10), (🚃, 30), (💼, 480), (🚶, 60), (🏠, 630) \rangle$$

On considère également la séquence S_1 *d'Alice définie dans l'exemple 5. Pour* $f = \Delta$*, on a ainsi* $\text{FTH}_{S_1 \to S_2} = 252.31$ *et* $\text{FTH}_{S_2 \to S_1} = 280.22$*.*

Théorème 6. *Pour* $f = \Delta$*,* $\text{FTH}_{S_1 \to S_2}$ *est bornée par* $T_{\max}$*.*

Démonstration. Par le lemme 4, on sait que $\forall e \in \mathbb{E}, \Delta(e) \leqslant \delta$. Par conséquent, et par la définition 20, on sait que $\text{FTH}_{S_1 \to S_2} \leqslant \sum_{k=1}^{n} \delta_{1k}$. Or, par la définition 13, on a $\sum_{k=1}^{n} \delta_{1k} = T_{\max}$ ce qui conclut la preuve. □

Notons ici que le théorème 6 peut être utile afin de normaliser notre dissimilarité entre 0 et 1.

Lemme 5. *Pour* $f = \Delta$ *ou* γ*,* $\text{FTH}_{S_1 \to S_2}$ *respecte l'axiome de séparabilité :* $\forall S_1, S_2 \in \mathbb{S}, \text{FTH}_{S_1 \to S_2} = 0 \Leftrightarrow S_1 = S_2$*.*

Démonstration. On raisonne par l'absurde. Supposons que $\exists S_1, S_2 \in \mathbb{S}$ telles que $S_1 \neq S_2$ et $\text{FTH}_{S_1 \to S_2} = 0$. $S_1 \neq S_2$ signifie qu'il existe un intervalle de temps $[t, t+\varepsilon[$ avec $\varepsilon > 0$ tel que les activités dans S_1 et dans S_2 sont différentes. En accord avec l'équation 6.4, $\text{FTH}_{S_1 \to S_2} = 0$ est possible si et seulement si $\forall k \in [\![1, n]\!], \Delta(e_k) = 0$. Comme $\Delta(e_k) = \delta_{1k} \times \gamma(e_k)$ et $\forall k \in [\![1, n]\!], \delta_{1k} > 0$, on doit montrer que $\forall k \in [\![1, n]\!], \gamma(e_k) = 0$.

Par le lemme 2, ce résultat est possible si et seulement si $\forall i \in [\![1, n]\!], \exists k \in [\![1, p]\!], I(x_{1i}) \subseteq I(x_{2k})$ avec $x_{1i} = x_{2k}$. Or, comme $\bigcup_{i=1}^{n} I(x_{1i}) = [0, T_{\max}[$ et que tous les intervalles temporels sont disjoints (i.e., $\forall i, j \in [\![1, n]\!], i \neq j, I(x_{1k}) \cap I(x_{1j}) = \varnothing$), alors on en déduit que le seul moyen de satisfaire l'assertion précédente est que $I(x_{1i}) = I(x_{2k})$. De plus, il vient également par le lemme 2 que $\forall i \in [\![1, n]\!], x_{1i} = x_{2k}$. Dès lors, les intervalles entre les deux séquences sont identiques avec les mêmes activités à chaque instant. L'hypothèse de départ est absurde, on en déduit que $S_1 = S_2$. □

Finalement, afin d'assurer la symétrie au sein de notre mesure, nous appliquons une $\top$-conorme entre les deux dissimilarités unilatérales de la même façon que pour CED.

Définition 21. *Soient deux séquences sémantiques-temporelles $S_1 \in \mathbb{S}^n$ et $S_2 \in \mathbb{S}^p$, la Fuzzy temporal Hamming distance* $\text{FTH} : \mathbb{S}^n \times \mathbb{S}^p \to \mathbb{R}^+$ *entre S_1 et S_2 est définie telle que :*

$$\text{FTH}(S_1, S_2) = \max\{\text{FTH}_{S_1 \to S_2}, \text{FTH}_{S_2 \to S_1}\} \tag{6.5}$$

Théorème 7. *Pour $f = \Delta$ ou γ,* $(\text{FTH}, \mathbb{S})$ *forme un espace semi-métrique.*

Démonstration. Par construction, la définition 21 satisfait la symétrie. De plus, par le lemme 5 qui montre que $\text{FTH}_{S_1 \to S_2}$ respecte la séparabilité, alors la séparabilité tient immédiatement pour FTH. □

Théorème 8. *Si pour toute opération d'édition $e \in \mathbb{E}$, la frontière $B(\mu_e) = \varnothing$, alors FTH pour $f = \Delta$ est équivalente à la distance de Hamming.*

Démonstration. Soient $(S_1, S_2) \in \mathbb{S}^n \times \mathbb{S}^p$, on calcule $\text{FTH}_{S_1 \to S_2}$:

$$\begin{aligned} \text{FTH}_{S_1 \to S_2} &= \sum_{k=1}^{n} \Delta(e_k) \\ &= \sum_{k=1}^{n} \delta_{1k} - \sup_{\tau \in I} \left\{ \int_{\tau}^{\tau+\delta_{1k}} sim_{e_k}(t) \times \mu_{e_k}(t) dt \right\} \end{aligned}$$

De plus, on sait que $\forall e_k \in \mathbb{E}$ telle que $e_k = (x_{1k}, \delta_{1k}, \text{begin}(x_{1k}), S_2)$ et que $B(\mu_e) = \varnothing$, dès lors : $sim_{e_k}(t) \times \mu_{e_k}(t) = \begin{cases} sim_{e_k}(t) & \text{si } t \in I(x_{1k}) \\ 0 & \text{sinon} \end{cases}$.

Ainsi, on peut restreindre l'intégrale sur l'intervalle $I(x_{1k})$. On a :

$$\begin{aligned}\mathrm{FTH}_{S_1 \to S_2} &= \sum_{k=1}^{n} \delta_{1k} - \int_{I(x_{1k})} sim_{e_k}(t)dt \\ &= T_{\max} - \sum_{k=1}^{n} \int_{I(x_{1k})} \sum_{j=1}^{p} \mathbf{1}_{I(x_{2j})}(t) \times sim(x_{1k}, x_{2j})dt\end{aligned}$$

De façon similaire, on peut limiter la fonction indicatrice $\mathbf{1}_{I(x_{2j})}$ à $I(x_{1k})$ telle que :

$$\mathrm{FTH}_{S_1 \to S_2} = T_{\max} - \sum_{k=1}^{n} \int_{I(x_{1k})} \sum_{j=1}^{p} \mathbf{1}_{I(x_{1k}) \cap I(x_{2j})}(t) \times sim(x_{1k}, x_{2j})dt$$

Enfin, et comme $sim_{e_k}(t) = 0$ pour $t \notin I(x_{i1})$, on peut généraliser l'intégrale sur I :

$$\begin{aligned}\mathrm{FTH}_{S_1 \to S_2} &= T_{\max} - \int_{I} \sum_{k=1}^{n} \sum_{j=1}^{p} \mathbf{1}_{I(x_{1k}) \cap I(x_{2j})}(t) \times sim(x_{1k}, x_{2j})dt \\ &= T_{\max} - \sum_{k=1}^{n} \sum_{j=1}^{p} |I(x_{1k}) \cap I(x_{2j})| \times sim(x_{1k}, x_{2j})\end{aligned}$$

Cette dernière expression satisfait la symétrie et est équivalente à la distance de Hamming pour des séquences sémantiques-temporelles continues. □

Théorème 9. *Soient deux séquences sémantiques-temporelles $S_1 \in \mathbb{S}^n$ et $S_2 \in \mathbb{S}^p$, le temps de calcul de* $\mathrm{FTH}(S_1, S_2)$ *est en* $O(\max\{n, p\} T_{\max} \log T_{\max})$.

Démonstration. Nous avons vu précédemment que l'équation 6.2 peut être calculée en $O(T_{\max} \log T_{\max})$. Ainsi, pour $(S_1, S_2) \in \mathbb{S}^n \times \mathbb{S}^p, \mathrm{FTH}_{S_1 \to S_2}$ a une complexité de $O(nT_{\max} \log T_{\max})$. Ainsi, l'équation 6.5 de FTH a une complexité en $O((n + p)T_{\max} \log T_{\max}) = O(\max\{n, p\} T_{\max} \log T_{\max})$. □

Par conséquent, lorsque n et p sont bien inférieurs à $T_{\max}$ (ce qui est en pratique souvent le cas), le calcul de FTH est beaucoup moins gourmand que les autres mesures classiques d'Optimal Matching. Pour conclure cette section, la table 6.2 résume les avantages et les principales propriétés de chaque méthode étudiée. Nous notons que FTH vérifie, comme CED, l'ensemble des spécificités requises. De plus, FTH permet le calcul de la similarité selon une approche continue du temps qui tient compte des durées. Enfin, elle garantit un meilleur temps de calcul que la plupart des mesures de l'état de l'art, sous condition que les séquences aient la même taille ($T_{\max}$ fixe et aient peu de symboles i.e., n, p petits.

Mesures	Propriétés							
	Métrique	**Semi-métrique**	**Seq. taille fixe**	**Disto. temp.**	**Permut.**	**Sim.**	**Continue**	**Complex. temp.**
Hamming	×		×			×†		$O(T_{\max})$
LCS	×							$O(T_{\max}^2)$
DTW				×		×†		$O(T_{\max}^2)$
ED	×‡				×‡	×†		$O(T_{\max}^2)$
CED		×		×	×	×		$O(T_{\max}^3)$
FTH		×	×	×	×	×	×	$O(\max\{n,p\}T_{\max}\log T_{\max})$

†Par défaut, distance triviale $\rho(x,y) = \begin{cases} 0 & x = y \\ 1 & \text{else} \end{cases}$

‡ Variante Damerau-Levenshtein [52]. Sacrifie l'inégalité triangulaire et autorise les transpositions adjacantes.

Table 6.2 – Comparaison des propriétés principales des mesures sur les séquences sémantiques-temporelles

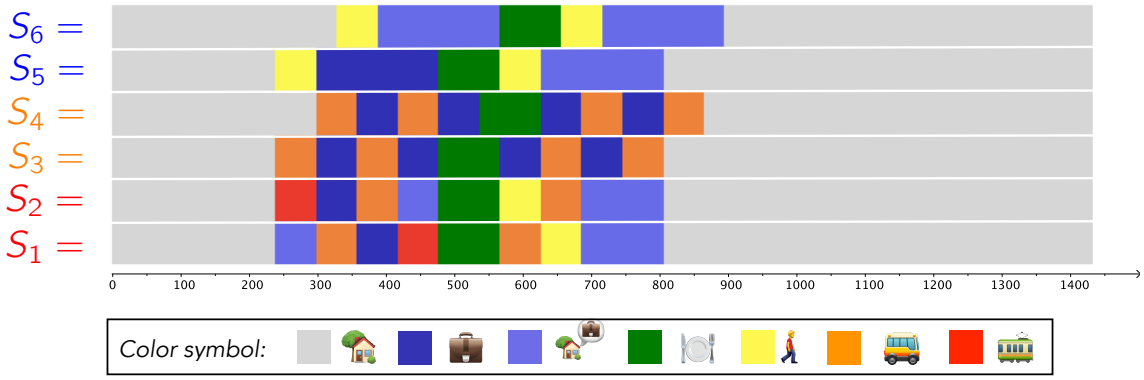


Figure 6.6 – Exemple de 6 séquences sémantiques-temporelles fictives

6.3 Expérimentations

Dans cette section, nous présentons une utilisation pratique de FTH pour la fouille de séquences de mobilité humaine afin d'identifier des comportements de mobilité basés sur les séquences sémantiques-temporelles d'activités. Ces expérimentations et le code de FTH sont entièrement disponibles sur notre Github [1] et Google Colab [2]. La première partie présente un exemple pilote qui illustre l'accomplissement des spécificités et propriétés montrées dans la section précédente par rapport aux autres mesures classiques de l'état de l'art. La seconde partie expose une comparaison entre les mesures FTH et Hamming sur des séquences sémantiques-temporelles réelles obtenues à partir d'un sondage EMD (Enquête Ménage-Déplacement) et démontre l'applicabilité de notre méthode à un cas d'usage réel. Dans la suite de cette section et dans un but de concision, nous appelons FTHΔ et FTHγ l'instanciation de FTH avec les fonctions de coût respectives Δ et γ.

6.3.1 Exemple pilote

La figure 6.6 présente un échantillon de 6 séquences sémantiques-temporelles inspirées de l'ensemble de données EMD. Ces séquences sémantiques-temporelles sont délibérément caricaturales afin de mettre en évidence les propriétés attendues des mesures étudiées. Chaque carré coloré représente une heure (60min) d'activité. Les séquences ont été construites par paires de la façon suivante :

- La paire (S_1, S_2) exhibe quelques permutations au sein des deux séquences. En outre, S_1 et S_2 sont composées des mêmes activités mais globalement mélangées. On remarque une transposition des activités "Travail" (💼) et "Bus" (🚌) mais aussi "Marche à pied" (🚶) et "Bus". On note également une permutation plus distante entre les activités "Tramway" (🚋) et "Travail à la maison" (🏠💼).

1. https://github.com/Clement-Moreau-Info/FTH
2. https://github.com/Clement-Moreau-Info/FTH/blob/main/fuzziee_temporal_hamming_dis.ipynb

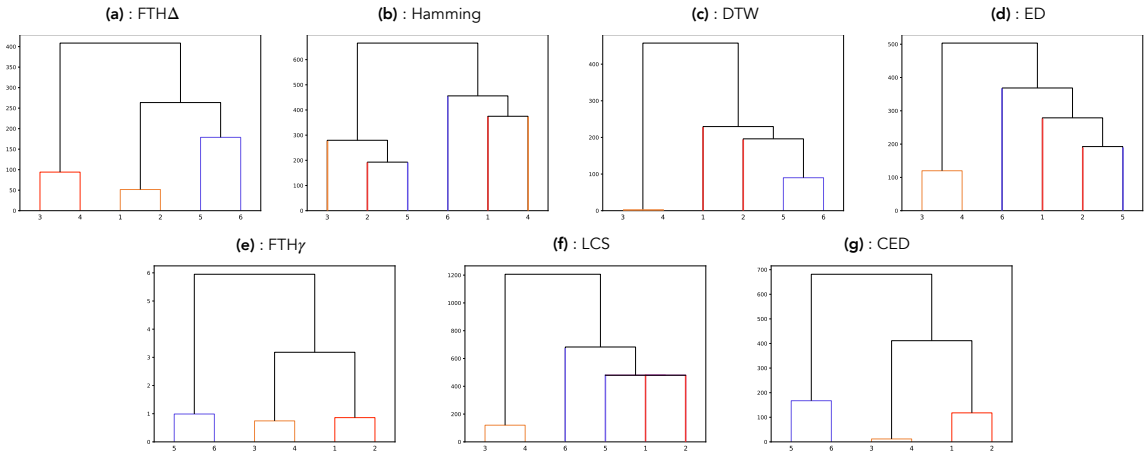


Figure 6.7 – Dendrogrammes des séquences sémantiques-temporelles de la figure 6.6 pour différentes mesures. Les couleurs indiquent la couleur de la paire d'origine.

- La paire (S_3, S_4) montre un décalage temporel. Les activités de S_4 sont identiques à celles de S_3 mais exécutées 60min plus tard.
- La paire (S_5, S_6) reprend l'idée de décalage temporel mais avec une durée plus importante de 90min. De plus l'activité matinale "Travail" dans S_5 est substituée par une activité similaire, "Travail à la maison" ( ), dans S_6.

Sur la base de ces séquences sémantiques-temporelles, la figure 6.7 représente les dendrogrammes construits à l'aide d'un clustering hiérarchique pour chaque mesure étudiée dans la sous-section 6.1.2. Le but pour chaque mesure est de retrouver les paires étiquetées ci-dessus. Les dendrogrammes ont été calculés en selon l'algorithme Linkage et le critère d'agrégation de Ward de la bibliothèque Scipy (1.4.1) de Python 3. La fonction temporelle de FTH est quant à elle fixée avec une frontière floue $\beta = 240$min.

Concernant l'analyse des dendrogrammes, on remarque tout d'abord que les séquences avec de petits décalages (paire orange) sont bien regroupées avec toutes les méthodes testées, à l'exception de la distance de Hamming, tandis que les séquences avec des décalages temporels plus importants et des activités similaires mais différentes (paire bleu) ne sont regroupées correctement qu'avec DTW, CED, FTHΔ et FTHγ ce qui confirme que ces méthodes sont robustes aux décalages temporels. Pour les permutations (paire rouge), les séquences sont bien regroupées avec FTHΔ, FTHγ et CED. Nous remarquons que LCS regroupe S_1, S_2 et S_5. Enfin, on voit que les séquences sémantiques-temporelles sont toutes correctement regroupées avec FTHΔ, FTHγ et CED.

Néanmoins, nous modérons ces résultats en rappelant que les exemples présentés dans cette section sont construits avec comme objectif particulier de mettre en défaut les méthodes afin de tester la validité des mesures sur les propriétés ciblées (permutations, répétitions, décalages temporels, homogénéité sémantique). En particulier, sur un ensemble de données réelles, on s'attend à ce que la distance de Hamming soit plus

performante. Par ailleurs, d'un point de vue computationnel, Hamming reste de loin le meilleur choix par rapport, par exemple, à la mesure CED qui, certes, fournit un clustering parfait sur cet exemple mais qui est en pratique inapplicable pour de grandes valeurs de $T_{\max}$.

6.3.2 Clustering de séquences sémantiques-temporelles de mobilité

Description du jeu de données Pour tester l'applicabilité de nos deux variantes de FTH par rapport aux autres mesures, nous les avons mis en oeuvre dans une tâche de clustering portant sur un échantillon aléatoire de 1200 séquences sémantiques-temporelles extraites du jeu de données Enquête Ménages-Déplacements (EMD) Rennes 2018 (voir section 8.2). Les symboles utilisés forment des activités et sont fortement similaires à ceux présentés dans la figure 6.2. Ces activités sont organisées en un graphe sémantique (i.e., taxonomie) et sont comparées par paire à l'aide de la mesure de similarité de Wu-Palmer [244]. Nous renvoyons à l'article [165] pour davantage de détails et une analyse exhaustive de cet EMD.

Méthodologie Afin de comparer la distance de Hamming avec FTHΔ et FTHγ, nous avons réalisé trois clustering hiérarchiques avec la même paramétrisation qu'à la section précédente dans le but d'étudier le ré-arrangement des clusters en fonction de la mesure utilisée. Basé sur la maximisation du score de Silhouette [199] et du saut d'inertie (voir section 4.2), nous avons fixé le nombre de clusters à 5 pour chaque processus de clustering. La frontière temporelle de la fonction temporelle μ_e est fixée à $\beta = 12$h afin de détecter les activités similaires au sein du matin et de l'après-midi.

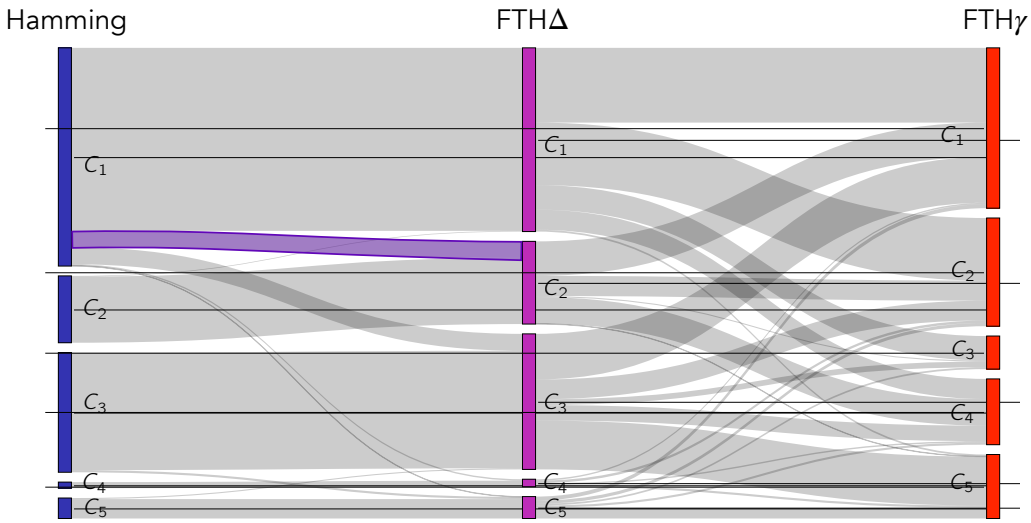


Figure 6.8 – Diagramme de Sankey illustrant les flux entre les 3 partitions formées par les mesure de Hamming, FTHΔ et FTHγ

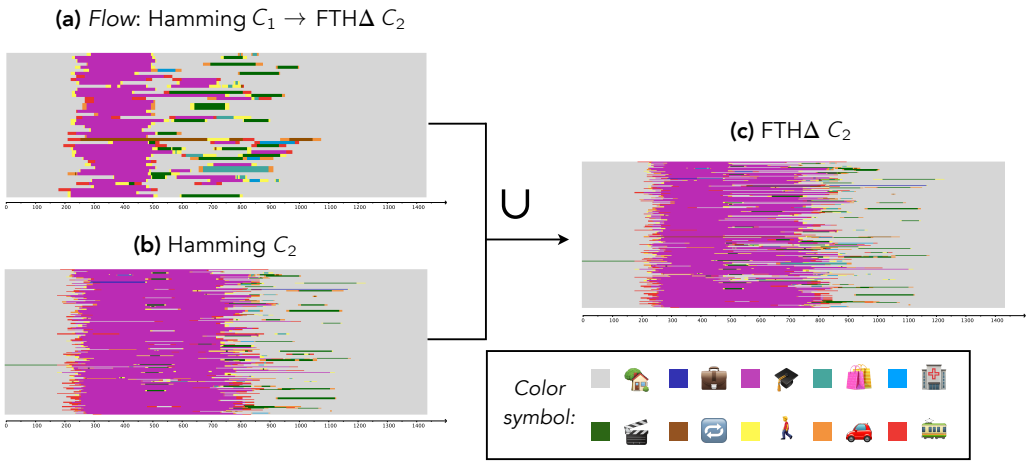


Figure 6.9 – Tapis de séquences dans (a) le flux depuis le cluster C_1 de Hamming vers le cluster C_2 de FTHΔ (b) Cluster C_2 de Hamming et (c) Union résultante dans FTHΔ

Analyses et résultats La figure 6.8 représente le diagramme de Sankey et flux entre les résultats de clustering. Nous observons 10,5% de ré-arrangements des séquences entre les clusterings Hamming et FTHΔ. Afin d'illustrer un de ces ré-arrangements, la figure 6.9 détaille le flux entre les clusters Hamming C_1 et FTHΔ C_2 coloré en violet dans la figure 6.8. Ce flux contient 46 séquences sémantiques-temporelles détaillées en (a) et concernent des étudiants (🎓) à leur école durant la matinée et effectuant, globalement, une activité de loisir (🎬) l'après-midi ; le cluster Hamming C_2 (b) concentre quant à lui des étudiants sur leur lieu d'étude toute la journée et ayant une activité de loisir plutôt en début de soirée. Ces séquences sont fusionnées dans le cluster FTHΔ C_2 (c), ce qui met en évidence la capacité de FTHΔ à effectuer des dilatations temporelles pourvu que le contexte soit similaire (ici un motif diurne étudiant avec une activité de loisir en deuxième partie de journée).

Concernant les flux entre FTHΔ et FTHγ, le grand nombre de ré-arrangements est dû au fait que FTHγ suit un paradigme de coût différent. Pour FTHγ, rappelons que le coût de transformation peut être élevé même si la durée de l'activité est courte. Par exemple, la transformation de 10min de marche peut être aussi importante que la transformation de 4h de travail. Nous prévoyons de travailler avec des experts pour analyser la pertinence de la mesure et les résultats des différents clusters afin de choisir la meilleure variante en fonction des besoins métiers.

Discussion

Dans ce chapitre, nous avons introduit une extension floue de la distance de Hamming pour les séquences sémantiques-temporelles appelée *Fuzzy Temporal Hamming* ou FTH. Cette nouvelle mesure améliore la distance de Hamming en introduisant une fenêtre temporelle floue afin d'être robuste aux distorsions temporelles comme les

décalages et permutations, et afin de saisir le contexte global autour d'une période donnée. Ces propriétés sont particulièrement nécessaires dans des domaines tels que l'analyse de la mobilité humaine dans le but d'extraire des comportements similaires.

En nous basant sur une adaptation de la fonction de coût d'opération d'édition de la distance de Hamming similaire à celle expérimentée pour CED dans le chapitre 6, nous avons prouvé que FTH satisfait les besoins précédents concernant les propriétés inhérentes à la mobilité et aux habitudes humaines tout en ayant une complexité de calcul compétitive par rapport aux autres mesures de l'état de l'art sur les séquences.

Enfin, FTH a été testée expérimentalement sur deux jeux de données. Un premier jeu de 6 séquences sémantiques-temporelles test aux structures délibérément caricaturales a été mis en oeuvre afin de vérifier les propriétés souhaitées par rapport aux autres mesures. Les résultats confirment que FTH les surpasse sur les capacités testées (timing, permutations, décalages et homogénéité sémantique). Le second jeu de données a mis en place des séquences sémantiques-temporelles réelles issues d'une EMD et a été utilisé pour comparer FTH avec la distance de Hamming dans une tâche de clustering. Nous avons montré que les FTH – coûts γ et Δ – ont la capacité de rassembler des séquences dont le contexte est proche. En perspective, nous proposons d'analyser plus en détail les clusters produits par les variantes de FTH, notamment la version de coût γ qui demande une analyse plus minutieuse – à la fois structurelle et compositionnelle – de la dimension temporelle. De plus, nous espérons tester notre nouvelle mesure sur des jeux de données plus grands, plus complexes et issus d'autres thématiques afin de vérifier les problématiques de passage à l'échelle et de généricité.

Chapitre 7

Une approche pour le clustering de séquences d'ensembles d'éléments sémantiques multi-dimensionnels

Publication

C. Moreau, A. Chanson, V. Peralta, T. Devogele, C. de Runz *Clustering Sequences of Multi-dimensional Sets of Semantic Elements*, ACM SAC (2021)

7.1 Aspect multi-dimensionnel de la sémantique

Jusqu'à présent, nous avons considéré que les séquences sémantiques (resp. temporelles-sémantiques) étaient constituées d'éléments atomiques dans le sens où chaque période temporelle donnée est associée à une unique activité. Or, il est pourtant fréquent d'effectuer plusieurs actions au même instant (e.g., être dans un bus et écouter de la musique), voire même que les éléments de la séquence soient décrits par plusieurs dimensions. La première sous-section illustre des exemples de tels phénomènes et les applications que ceux-ci peuvent avoir. La seconde sous-section propose un formalisme de représentation et une mesure pour la comparaison de telles séquences.

7.1.1 Contexte illustratif

Nous avons vu en section 2.1 que la modélisation séquentielle des données s'impose dès qu'il s'agit de représenter un phénomène ou des éléments qui évoluent à travers le temps. Lorsque ces éléments ne peuvent être qualifiés que par des informations qualitatives, alors la séquence prend la forme d'une séquence sémantique. De telles séquences permettent de modéliser la mobilité humaine d'un point de vue sémantique, les éléments en jeu représentent alors des lieux, des activités, des modes de transport, etc. Cependant, il est tout aussi envisageable de représenter par exemple des playlists musicales comme sur l'exemple de la figure 7.1, ou des logs de base de données comme vus en section 5.3. Lorsque les éléments en jeu sont complexes et dotés de nombreux

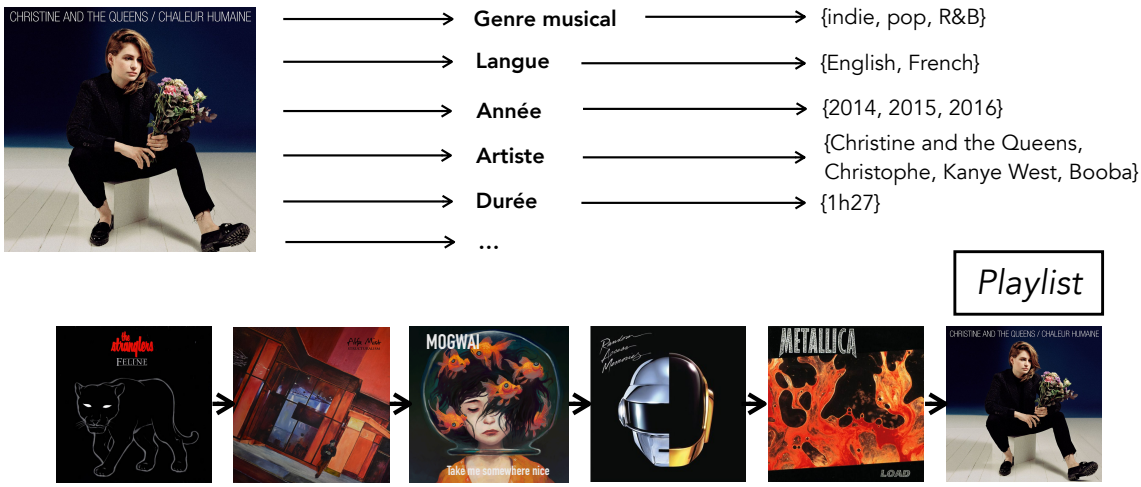


Figure 7.1 – Exemple de playlist musicale formant une séquence d'ensembles d'éléments sémantiques multi-dimensionnels

attributs sémantiques, comme dans le cas des musiques, le processus de comparaison des séquences devient plus complexe et abstrait. Dans le cas de la figure 7.1, les éléments possèdent plusieurs dimensions (genre, langues, années, artistes, etc.) où chacune de ces dimensions forme un ensemble de labels sémantiques qui doivent être comparés puis agrégés afin d'établir une proximité avec les autres éléments (e.g., album).

Revenons au contexte de la mobilité humaine. Nous ancrerons ici notre cadre de réflexion auprès du projet SMARTLOIRE dont l'objectif est d'offrir un ensemble d'outils numériques à destination des professionnels du tourisme et décideurs politiques pour la recommandation d'itinéraires et l'analyse de traces touristiques en région Centre-Val de Loire. En outre, la comparaison des parcours touristiques et l'extraction de comportements issus de ces traces permet aux acteurs du tourisme de mieux comprendre les dynamiques touristiques de la région et de saisir les intérêts et opportunités selon les profils socio-démographiques des individus. Ces connaissances permettent alors d'adapter les offres de recommandation et d'améliorer la visibilité de certains sites en ciblant au mieux les touristes susceptibles d'y être intéressés.

Considérons les deux scénarios suivants :

"Alice est en vacances dans la vallée de la Loire et loge dans un hôtel d'époque Renaissance. Elle quitte son hôtel pour visiter le matin un musée sur la religion au Moyen Âge. Elle pique-nique dans un parc pour se restaurer le midi. L'après-midi, Alice visite un château puis une église. Le soir, elle dîne dans un pub / restaurant avant d'assister à un spectacle son et lumière dans un château Renaissance. Le spectacle terminé, elle retourne à son hôtel."

"Bob est aussi en vacances dans la vallée de la Loire. Bob séjourne dans un camping. Le matin, il visite un château avec ses jardins fleuris. Il déjeune dans un bistrot.

L'après-midi, il assiste à un concert baroque dans une église gothique, puis fait une visite guidée de la ville à vélo. Le soir, Bob dîne à nouveau dans un bistrot avant de retourner au camping."

Quel est le degré de similarité entre ces séjours touristiques ? Nous avons vu dans la section 3.1 que la comparaison d'éléments sémantiques atomique est déjà un problème difficile. Or, cette difficulté est accrue lorsque les éléments à comparer ne sont plus unidimensionnels mais multidimensionnels, c'est-à-dire qu'ils sont définis au moyen de plusieurs concepts exprimant des sémantiques diverses et variées. Par exemple, comment comparer un "*Spectacle son et lumière dans un château Renaissance*" et un "*Concert baroque dans une église gothique*" ? La comparaison doit-elle prendre en compte le type d'événement (spectacle vs. concert), le lieu (château vs. église) ou le style architectural (renaissance vs. gothique) ? De plus, certains éléments peuvent concerner un ensemble de concepts pour une même dimension, par exemple un *château* avec un *jardin fleuri*.

Grâce aux ontologies de domaine capturant les besoins spécifiques métiers et à l'explosion du Web sémantique et Linked Open Data (LOD), la capacité de comparer avec précision des éléments sémantiques complexes semble plus que jamais possible [83]. Néanmoins, à notre connaissance, la question de la comparaison d'ensembles d'éléments sémantiques multidimensionnels, ambigus et spécifiques à un domaine reste un problème ouvert [100]. La question est encore plus piquante lorsqu'il s'agit de comparer des séquences d'ensembles multidimensionnels d'éléments sémantiques.

Dans la section suivante, nous proposons une approche pour la modélisation de telles séquences d'ensembles multidimensionnels d'éléments sémantiques ainsi que leur comparaison basée sur l'utilisation d'ontologies et de graphes de concepts tels que présentés section 3.1.2.

7.1.2 Comparaison d'éléments sémantiques multidimensionnels

On considère un ensemble $\mathscr{O} = \{O_1, ..., O_q\}$ de q ontologies (i.e., graphes de concepts) formant des ensembles de concepts structurés. Chaque graphe de connaissance décrit une famille de propriétés des éléments à comparer. Ainsi, leurs concepts sont disjoints, à l'exception du concept `all` (i.e., $i \neq j, O_i \cap O_j = \{\texttt{all}\}$) qui est commun à tous les graphes de connaissances. On définit alors un élément sémantique multi-dimensionnel comme un vecteur q-dimensionnel, où chaque dimension est un sous-ensemble de concepts d'un graphe de connaissance.

Définition 22 (Élément sémantique multi-dimensionnel)**.** *Soit $\Sigma = \times_{k=1}^{q} \mathcal{P}(O_k)$ avec $\times$ désignant le produit cartésien et $\mathcal{P}$ l'ensemble des parties. Un élément sémantique multi-dimensionnel $\sigma \in \Sigma$ est un vecteur à q dimensions où la k-ème composante (notée $\pi_k(\sigma)$) est un sous-ensemble de O_k.*

Une telle formalisation permet, par exemple, de se ré-approprier la modélisation de séquences comportementales de Cao [35] présentée section 4.1.1.

Dans les exemples suivants, nous revisitons les séjours d'Alice et de Bob en utilisant les concepts de l'ontologie DATAtourisme, esquissée figure 7.2 et décrite lors de la présentation du cas d'étude (section 7.2).

Exemple 9. *Considérons une des activités de Bob : Le matin, il visite un château avec des jardins fleuris (château de Chaumont). Considérons trois dimensions décrivant les activités touristiques : le lieu d'intérêt, le type d'événement et le style architectural, en prenant les concepts de trois ontologies de domaine. L'activité de Bob peut être formalisée comme suit :*

$$\sigma_1 = (\{Castle, ParkAndGarden\}, \varnothing, \{Renaissance\})$$

indiquant que l'activité se déroule à la fois dans un château et dans un parc/jardin, sans événement particulier[1] *(visite simple) et que le lieu est de style Renaissance.*

Formalisons maintenant une des activités d'Alice : Aller voir un spectacle son et lumière dans un château de la Renaissance (château de Blois). Cette activité peut être formalisée comme suit :

$$\sigma_2 = \langle \{Castle\}, \{VisualArtEvent\}, \{Renaissance\} \rangle$$

Afin de comparer des éléments (i.e. vecteurs), nous proposons de comparer séparément chaque dimension (i.e. sous-ensembles de concepts) et d'agréger les scores de similarité obtenus. Par ailleurs, notre méthode s'emploie à être capable de traiter les vecteurs aux dimensions incomplètes, ce qui est une situation très fréquente dans le cas des activités touristiques qui peuvent être mal renseignées.

Formellement, nous avons besoin d'une mesure de similarité entre parties d'ensemble (comme présenté en section 3.1.2.4). Pour ce faire, nous proposons d'utiliser la mesure d'Halkidi [94] $\zeta : \mathcal{P}(O) \times \mathcal{P}(O) \rightarrow [0, 1]$ définie telle que :

$$\zeta(X, Y) = \frac{1}{2} \left(\frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} \{sim(x, y)\} + \frac{1}{|Y|} \sum_{y \in Y} \max_{x \in X} \{sim(x, y)\} \right) \tag{7.1}$$

La fonction sim désigne ici la similarité simple entre deux concepts d'une même ontologie. Nous renvoyons à la section 3.1.2 pour un rappel sur l'ensemble des mesures et techniques disponibles. Ici, nous faisons le choix d'utiliser la similarité de Wu-Palmer [244] pour la comparaison entre concepts. Nous explicitons les raisons de ce choix en section 3.1.2 et dans la table 3.3, la similarité de Wu-Palmer semblant donner les

1. Précisions que la "visite simple" est considérée comme un événement $\varnothing$ afin d'éviter un doublon d'information. En effet, dans notre contexte de représentation des séjours touristiques, la présence de σ_1 au sein d'une séquence suffit à induire la visite du lieux représenté. De plus, une visite simple n'est pas considérée, au sens métier, comme un événement touristique majeur à la différence par exemple d'un Son et Lumière ou d'une visite costumée.

résultats les plus consensuels au regard des autres mesures abordées. Pour rappel, celle-ci est définie telle que :

$$sim_{wup}(x, y) = \frac{2 \times d(LCA(x, y))}{d(x) + d(y)} \tag{7.2}$$

Enfin, la similarité entre deux éléments de σ, notée $sim_\Sigma : \Sigma \times \Sigma \rightarrow [0, 1]$ est calculée comme l'agrégation des scores de similarité pour chaque dimension comme suit :

$$sim_\Sigma(\sigma, \sigma') = Agg_{k=1}^{q} \{\zeta(\pi_k(\sigma), \pi_k(\sigma'))\} \tag{7.3}$$

Où $Agg : [0, 1]^q \rightarrow [0, 1]$ désigne une fonction d'agrégation quelconque.

Comme soulevé précédemment, il n'est pas rare de devoir traiter des éléments incomplets, c'est-à-dire ayant des valeurs manquantes pour certaines dimensions. Par exemple, dans l'exemple 9, la deuxième composante de σ_1 n'est pas renseignée outre mesure. La gestion des valeurs manquantes est délicate car elles peuvent indiquer des valeurs inconnues, non pertinentes ou inexistantes. Pour faire face à ce problème, nous proposons l'utilisation de la fonction d'agrégation $average\_if$ qui calcule une moyenne mais en ignorant les dimensions où l'un des éléments a des valeurs manquantes (notées $\varnothing$). En d'autres termes, seules les dimensions k telles que $\pi_k(\sigma) \neq \varnothing \wedge \pi_k(\sigma') \neq \varnothing$ sont considérées. Les poids de la moyenne, si celle-ci est pondérée, sont équitablement répartis sur les autres dimensions. Notons que d'autres opérateurs d'agrégation seraient possibles (cf. [57] pour une étude de la plupart des opérateurs classiques), mais les caractéristiques de la moyenne (e.g. simplicité, absence d'élément neutre, absence d'élément absorbant) nous semblent pertinentes pour notre problème afin de prendre en considération toutes les dimensions de façon égale.

Exemple 10. *Soit les instances σ_1 et σ_2 définies dans l'exemple 9. Le type d'événement de σ_1 étant vide, le calcul de $sim_\Sigma(\sigma_1, \sigma_2)$ est tel que :*

$$sim_\Sigma(\sigma_1, \sigma_2) = \frac{1}{2}\zeta(\{Castle, ParkAndGarden\}, \{Castle\}) + \frac{1}{2}\zeta(\{Renaissance\}, \{Renaissance\})$$

Pour la dernière partie de l'équation, on a : $\zeta(\{Renaissance\}, \{Renaissance\}) = 1$. Pour la première partie, on calcule la similarité de Wu-Palmer entre les concepts. On rappelle que, $\forall x \in O, sim_{wup}(x, x) = 1$. Ainsi, on calcule uniquement $sim_{wup}(Castle, ParkAndGarden) = \frac{1}{2}$, en accord avec l'équation 7.2 et l'ontologie décrite figure 7.2.

Enfin, par l'équation 7.1, on a : $\zeta(\{Castle, ParkAndGarden\}, \{Castle\}) = 0.875$.

Finalement, $sim_\Sigma(\sigma_1, \sigma_2) = \frac{1}{2} \times 0.875 + \frac{1}{2} \times 1 = 0.9375$

Ainsi, grâce aux définitions précédentes, il est possible d'étendre la définition 1 de la séquence sémantique au cas où les symboles sont des éléments sémantiques multidimensionnels.

Exemple 11. *Revenons à l'exemple fil rouge d'Alice et Bob. Nous représentons la journée d'Alice d'une manière plus formelle à l'aide des définitions des éléments sémantiques multidimensionnels (def. 22) et des séquences sémantiques (def. 1) comme suit :*

$$\begin{aligned} S_{Alice} = \langle & (\{Hotel\}, \varnothing, \{Renaissance\}), \\ & (\{ReligiousSite, Museum\}, \varnothing, \{Medieval\}), \\ & (\{ParkAndGarden\}, \varnothing, \varnothing), \\ & (\{Castle\}, \varnothing, \{Renaissance\}), \\ & (\{Church\}, \varnothing, \{Roman\}), \\ & (\{PubAndBar, Restaurant\}, \varnothing, \varnothing), \\ & (\{Castle\}, \{VisualArtEvent\}, \{Renaissance\}), \\ & (\{Hotel), \varnothing, \{Renaissance\})\rangle \end{aligned}$$

Ce type de séquence peut ensuite être comparé à l'aide de la similarité sim_{Σ} et des mesures CED et FTH (dans le cas d'une dimension "durée") développées en chapitre 5 et 6.

7.2 Cas d'étude et expérimentations

Dans cette section, nous décrivons les expérimentations menées pour valider notre approche. Nous introduisons tout d'abord une étude de cas concernant la mobilité des touristes, encadrée par le projet SMARTLOIRE, puis nous présentons le protocole expérimental appliqué et discutons les résultats obtenus.

7.2.1 Description du cas d'étude et de l'ontologie DATAtourisme

Le projet SMARTLOIRE [2] fait partie d'une initiative régionale afin d'aider les utilisateurs à déterminer des itinéraires touristiques personnalisés dans la région de la vallée de la Loire [31]. En particulier, l'identification de clusters de comportements cohérents avec des schémas de visite similaires est essentielle pour guider les acteurs du tourisme sur les profils des utilisateurs mais aussi pour concevoir de meilleurs outils de recommandation basés sur les connaissances extraites. Nous représentons les itinéraires touristiques comme des séquences sémantiques où les symboles sont des éléments sémantiques multidimensionnels. En outre, chaque symbole est décrit par trois dimensions : (i) le lieu (POI), (ii) l'événement et (iii) style architectural du POI.

2. `https://smartloire.univ-tours.fr`

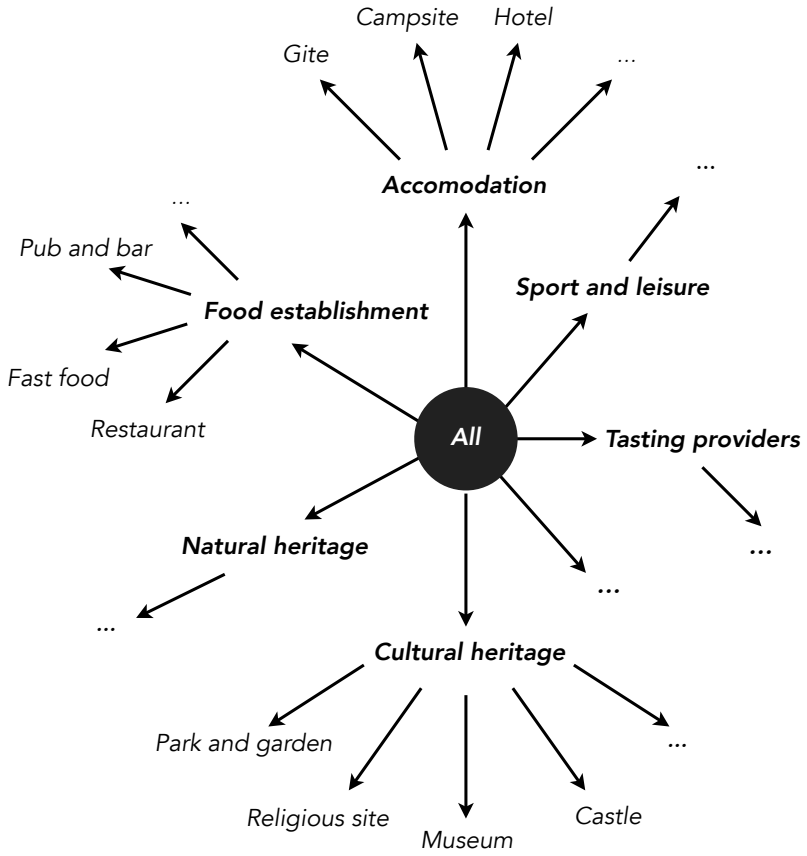


Figure 7.2 – Taxonomie des POI extraite de DATAtourisme

Le jeu de données de ces éléments touristiques est tiré de DATAtourisme [3], une ontologie nationale normalisée décrivant les entités touristiques. DATAtourisme est une vaste ontologie organisée principalement en deux grandes parties, *POI* et *Événement*. Ainsi, une instance touristique (c'est-à-dire un symbole) est décrite en fonction d'un POI et d'un événement, par exemple "un spectacle son et lumière dans un château de la Renaissance". En pratique, nous avons divisé DATAtourisme en ontologies plus petites, chacune servant de graphe de connaissances pour les trois dimensions sémantiques des symboles qui constitue les itinéraires (i.e., séquences sémantiques). La première est extraite du noeud PlaceOfInterest (POI) et de toutes ses sous-classes, la seconde regroupe plusieurs noeuds liés à des événements. Enfin, nous profitons du fait que certains POI ont des détails architecturaux pour créer une troisième ontologie construite autour des styles architecturaux des bâtiments. La figure 7.2 fournit une représentation résumée de l'ontologie POI extraite et des noeuds pertinents pour nos expériences. Les jeux de données complets peuvent être consultés sur notre GitHub [4].

Categorie	Concept	Nombre d'instances
Accomodation	Hotel	50
	Gite	50
	Camping	50
Food establishment	Fast Food	1
	Bar	29
	Restaurant	50
Nature Heritage	Nature	16
Cultural Heritage	Park and garden	24
	Religious site	29
	Museum	46
	Castle	47
Sport and leisure	Sport	50
Other (...)	...	50
Tasting Providers	Tasting	50
Total		542

Table 7.1 – Nombre d'instances utilisées dans l'ontologie POI

7.2.2 Profils touristiques et génération des données

Le projet SMARTLOIRE étant en cours de déploiement, nous ne disposons pour l'heure d'aucune trace réelle des activités des touristes, hormis de profils touristiques prototypiques décrits par les experts métiers. Dans ces conditions, nous avons donc choisi de générer des séquences artificielles issues, néanmoins, d'instances touristiques réelles issues de l'ontologie DATAtourisme. Ainsi, ce jeu de données nous permet de valider notre méthodologie et analyser la sensibilité de ses paramètres dans un environnement contrôlé.

À cette fin, nous avons sélectionné toutes les instances identifiées dans une zone de 50 km autour de la ville d'Amboise dans la vallée de la Loire. Nous avons obtenu environ 2500 instances différentes. Parmi celles-ci, nous avons sélectionné les 50 instances les plus décrites pour les 14 concepts les plus utilisés de l'ontologie POI illustrée dans la figure 7.2. La table 7.1 indique le nombre d'instances sélectionnées pour chaque concept de l'ontologie POI, soit un total de **542 instances** utilisées pour la génération des données. Les séquences artificielles ont été générées en utilisant un marcheur aléatoire markovien (i.e., chaine de Markov) qui donne une structure crédible aux activités quotidiennes de nos touristes virtuels. Le processus stochastique est absorbant et est défini sur les états décrits dans la figure 7.3.

Afin d'expliciter le processus de génération des séquences, on considère l'ensemble O de concepts structurés issus de l'ontologie POI, ainsi qu'un ensemble de profils

3. `https://framagit.org/datatourisme/ontology/`
4. `https://github.com/Clement-Moreau-Info/SAC2021`

id	État
1	*Start*
2	Activité matinale
3	Déjeuner
4	Activité après-midi
5	Hébergement
6	Activité nocturne
7	*End*

(a)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & p_1 & p_2 & 0 & p_3 & 0 & 0 \\ 0 & p_4 & p_5 & p_6 & p_7 & 0 & 0 \\ 0 & 0 & 0 & p_8 & p_9 & 0 & p_{10} \\ 0 & 0 & 0 & p_{11} & p_{12} & p_{13} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{14} & p_{15} \\ 0 & 0 & 0 & 0 & 0 & p_{16} & p_{17} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

(b)

Figure 7.3 – Description du marcheur aléatoire markovien (a) Table des états (b) Matrice stochastique

prototypiques prédéfinis Ψ. Chaque probabilité p_k dans la matrice stochastique est initialisée selon un profil $\psi \in \Psi$. De plus, chaque état s contient un ensemble de concepts $O_s \subset O$ tels que, pour chaque concept $c \in O_s$, la probabilité de sélectionner une instance σ de concept c dépend du profil ψ donné. On considère une probabilité de sélection de σ dans l'ensemble des concepts c selon une loi uniforme, soit $\frac{1}{|c|}$. Ainsi, pour un état s du marcheur aléatoire et un type de profil touristique ψ, la probabilité de sélectionner l'instance $\sigma \in c$ est telle que :

$$p(\sigma) = \frac{1}{|c|} \times p(c|\psi) \tag{7.4}$$

Pour les expérimentations de la section suivante, nous avons dressé 5 types de profil touristique partiellement inspirés de ceux identifiés par [228], chacun d'entre eux ayant une inclinaison naturelle pour certaines activités :

- *Les noctambules* ont une forte tendance à dormir le matin, à fréquenter les pubs et à pratiquer des activités nocturnes.
- *Les touristes culturels* visitent principalement de musées, châteaux et autres bâtiments remarquables. Ils mènent toujours une activité le matin et l'après-midi.
- *Les randonneurs* ne se logent que dans des campings et aiment visiter des espaces naturels.
- *Les jeunes couples* pratiquent des activités éparses et variées. Ils logent majoritairement dans des gîtes.
- *Les fins gourmets* ne dorment que dans des hôtels, déjeunent dans des restaurants et ont quelques prédispositions pour les activités de dégustation.

Notons que l'état 7 de la chaine de Markov est absorbant. Ainsi, ce processus génère un large éventail de modèles de séquences finis. À chaque étape de la marche aléa-

toire, une instance est choisie selon l'équation 7.4. La concaténation de ces instances produit une séquence artificielle représentant un séjour touristique d'une journée. Les détails sur l'initialisation des probabilités pour chaque profil sont disponibles dans notre GitHub [5]. Par ce procédé, nous avons généré **250 séquences** (50 de chaque profil) que nous exploitons dans la section suivante au sein d'un processus de clustering.

7.3 Clustering de séquences d'éléments multidimensionnels

Dans cette section, nous exploitons la mesure sim_{Σ} présentée dans la sous-section 7.1.2 pour le clustering de séquences d'éléments multidimensionnels. Plus précisément, sur la base de nos séquences artificielles aux comportements touristiques prédéfinis, notre but est l'élaboration d'un processus complet de clustering permettant d'extraire et discriminer proprement les séquences émanant de profils présentant des caractéristiques comportementales semblables.

7.3.1 Protocole et algorithmes

Sur la base des algorithmes détaillés en section 4.2, notre objectif est de déterminer empiriquement le plus adapté aux séquences d'éléments sémantiques multidimensionnels. En outre, à l'aide des bibliothèques SCIPY et SKLEARN de Python 3, nous avons testé les algorithmes :

- Clustering hiérarchique (critère de Ward)
- DBSCAN
- k-medoids
- Spectral

Nous ne pouvons pas exploiter directement k-means car nous ne disposons comme données numériques que les valeurs de notre distance entre les séquences.

Afin de disposer d'un point de comparaison, nous avons testé chacun des algorithmes selon différentes mesures : distance d'édition classique (Levenshtein), distance d'édition où le coût de l'opération de modification est égale à la similarité sim_{Σ} telle que définie équation 7.3 (Lev. + Ontologies) et la mesure CED initialisée avec la similarité sim_{Σ} et un noyau gaussien comme fonction d'encodage μ du vecteur temporel :

$$\mu(k) = \exp\left(\frac{1}{2}\left(\frac{k - k_{edit}}{\beta}\right)^2\right) \tag{7.5}$$

Ici, β est fixé empiriquement à $\frac{m}{2}$ où m correspond à la taille médiane (en nombre de symboles) des séquences.

5. `https://github.com/Clement-Moreau-Info/SAC2021`

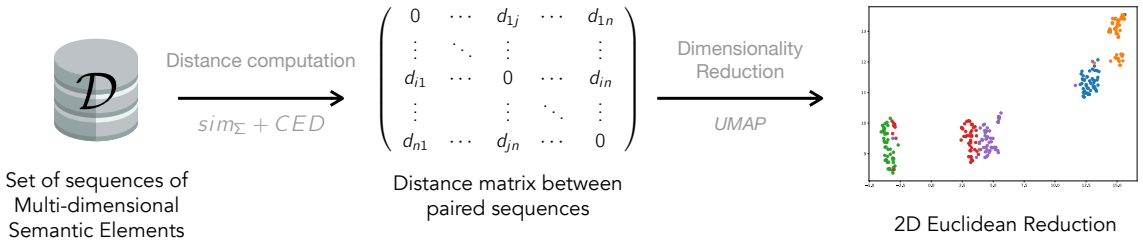


Figure 7.4 – Chaîne de traitement pour le clustering de séquences d'éléments multi-dimensionnels

De plus, nous explorons également la technique de réduction de dimensionnalité UMAP [148], en projetant notre jeu de données dans un espace euclidien 2D. Ainsi, nous avons pu tester les algorithmes *k*-means, Spectral et DBSCAN sur le jeu de données projeté. Notons que sur l'espace euclidien 2D produit en tant que projection d'UMAP, *k*-means, étant basé sur la même fonction objectif, permet de remplacer le clustering hiérachique selon le critère de Ward avec une complexité plus faible.

Concernant les paramètres UMAP, nous utilisons la librairie python UMAP-LEARN version 0.4.3 en utilisant les paramètres par défaut s'ils ne sont pas spécifiés, *min_dist* est fixé à 0.01, et *n_neighbors* à 25 (soit 10 % du jeu de données). La graine du générateur de nombres pseudo-aléatoires *random_state* est fixée à 42 pour rendre les résultats reproductibles. Toutes les expérimentations peuvent être reproduites en exécutant notre notebook [6] python dans Google Colab ou un environnement Jupyter.

Pour valider notre approche, l'objectif principal de nos expériences est d'évaluer dans quelle mesure les algorithmes de clustering et mesures associées sont capables de regrouper les séquences correspondant à un même profil. L'avantage d'utiliser des données générées ici est de connaître les classes des clusters, ce qui correspond à une forme vérité terrain. Nous évaluons alors notre approche en utilisant une métrique de qualité basée sur la performance de l'algorithme de clustering, ici l'Adjusted Rand Index (ARI) [193]. Contrairement aux mesures de qualité internes basées sur des critères topologiques (par exemple, Silhouette), cette mesure s'appuie sur les labels de vérité terrain dont nous disposons et nous permet ainsi d'obtenir une meilleure image de l'approche la plus performante.

7.3.2 Résultats expérimentaux

La table 7.2 présente les score d'ARI obtenus pour les différentes méthodes et mesure de clustering. Le meilleur score, 0.833, est obtenu par UMAP combiné à *k*-means ou Spectral, associé à CED. Sans projection UMAP, Spectral, associé à CED, surpasse les autres combinaisons avec un ARI de 0.649. Par ailleurs, CED surpasse les autres mesures pour tous les algorithmes de clustering rapportés. Nous remarquons que

6. `https://github.com/Clement-Moreau-Info/SAC2021/blob/main/Semantic_Trajectories_Clustering.ipynb`

	Levenshtein	Lev. + Ontologies	CED
DBSCAN	0.128	0.203	0.409
Hierarchical	0.315	0.171	0.483
k-medoids	0.300	0.170	0.550
Spectral	0.510	0.590	**0.649**
UMAP + DBSCAN	0.549	0.636	0.733
UMAP + *k*-means	0.659	0.673	**0.833**
UMAP + Spectral	0.665	0.680	**0.833**

Table 7.2 – Adjusted Rand Index des différents algorithmes et mesures pour le clustering des séquences touristiques simulées

lorsqu'il est associé à CED, le deuxième plus mauvais résultat est obtenu par le clustering hiérarchique (Hierarchical). Nous reviendrons sur ce résultat en discussion.

Il est intéressant de noter qu'en travaillant sur une matrice de similarité brute (sans projection UMAP), Spectral surpasse les autres méthodes, pour les trois mesures de similarité. Les scores ARI obtenus sont sensiblement plus élevés. Néanmoins, un résultat surprenant est que la projection réalisée par UMAP augmente les performances par rapport à n'importe quel algorithme et distance dans l'espace original, même en utilisant naïvement la distance de Levenshtein. Cela souligne le grand potentiel de la technique UMAP pour le regroupement de séquences sémantiques. De plus, la capacité de projeter un objet aussi complexe dans une représentation simple en 2D ouvre la possibilité de l'utiliser comme outil de visualisation pour les experts qui explorent les données.

La figure 7.5 met en évidence le comportement des trois mesures lorsqu'elles sont couplées avec UMAP. Cette représentation 2D des séquences touristiques montre l'éparpillement dans l'espace selon la mesure utilisée. Chaque couleur représente un profil touristique. Pour la projection CED (a), nous remarquons que les profils sont assez bien séparés et denses, la seule classe difficile à séparer des autres étant les

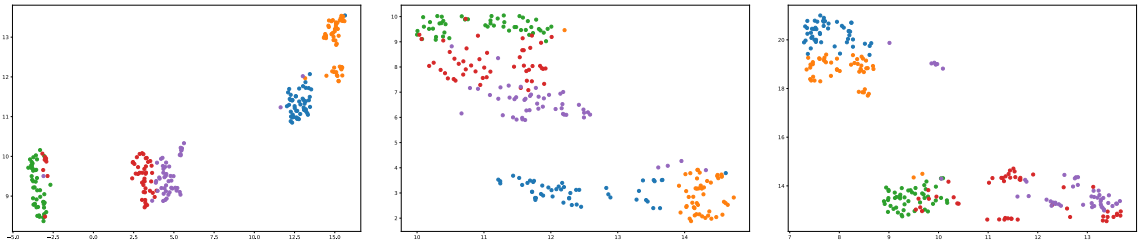


Figure 7.5 – Projection 2D de UMAP des 250 séquences en utilisant comme mesure de similarité (a) CED, (b) Lev. + Ontologies et (c) Levenshtein. Couleurs : bleu *Les randonneurs*, orange *Les noctambules*, vert *Les fins gourmets*, rouge *Les touristes culturels*, violet *Les jeunes couples*.

touristes culturel (rouge) avec quelques points mélangés avec les *fins gourmets* (vert) et relativement proche des *jeunes couples* (violet). Sur la projection de la mesure correspondant à Lev. + Ontologies (b), nous constatons immédiatement que les clusters de points sont moins denses à travers l'espace. Nous émettons l'hypothèse que la mesure Lev. + Ontologies, sans considération du contexte, n'est pas appropriée pour les séquences avec des valeurs manquantes. Nous remarquons à nouveau que la classe rouge est plus difficile à séparer des autres classes. Enfin, l'utilisation de la distance de base de Levenshtein (c) présente l'hétérogénéité, dans les clusters, la plus importante. En effet, plusieurs classes sont peu séparées : on retrouve ensemble les *jeunes couples* et les *fins gourmets*, mais aussi les *noctambules* (orange) et les *randonneurs* (bleu).

Discussion

Dans ce chapitre, nous avons proposé une nouvelle mesure et approche pour le clustering de séquences d'éléments sémantiques multidimensionnels, c'est-à-dire des séquences dont chaque élément est décrit pas plusieurs dimensions, chacune reliée à une ontologie spécifique, et où chaque dimension est décrite par un ensemble de valeurs. Ce type de séquences complexes est couramment utilisé dans de nombreux domaines tels que la modélisation de playlists musicales, de la mobilité ou plus généralement dès que les éléments dans les séquences peuvent être représentés de façon riche, sous plusieurs dimensions et par un aspect ensembliste. L'approche introduite comprend une mesure de similarité, basée sur des ontologies métiers, et une phase de clustering. Concernant la mesure de similarité, nous proposons à cet effet d'utiliser la mesure de similarité de Wu-Palmer combinée avec la mesure d'Halkidi et la fonction d'agrégation $average\_if$ pour comparer ces éléments sémantiques multidimensionnels complexes. Enfin, notre similarité peut-être ré-utilisée dans le cadre de la mesure CED pour la comparaison de séquences. Toutefois, notre approche reste modulaire et peut être utilisée avec d'autres techniques listées dans la partie État de l'art. Pour la phase de clustering, nous avons testé plusieurs algorithmes afin d'identifier le meilleur processus capable de traiter notre proposition. Nous avons constaté que l'utilisation du clustering hiérarchique sur le jeu de données testé a fourni des scores peu performants au vu des autres méthodes. Étant la méthode utilisée dans les chapitres précédents, nous pouvons nuancer et expliquer ce résultat par les explications suivantes :

- Les expérimentations menées dans les chapitres 5 et 6 ne concernait pas la variante de mesure testée ici pour les séquences d'éléments multidimensionnels. Ce résultat est, peut-être, une caractéristique de cette mesure et de la multidimensionnalité.
- L'expérimentation a été menée un unique jeu de données artificiel. Des réplications sur d'autres jeux de données (réels et avec une volumétrie plus importante) doivent venir corroborer ces résultats.
- Enfin, malgré ces failles, le clustering hiérarchique offre, grâce au dendro-

gramme, une forme de visualisation précieuse des clusters et une flexibilité quant au choix du nombre de clusters.

Néanmoins, nous avons constaté que la technique de réduction dimensionnelle UMAP combinée à un clustering spectral ou à un k-means surpasse les autres méthodes sur le jeu de données testé. À notre connaissance, il s'agit de la première utilisation de cette technique pour représenter les séquences sémantiques dans un espace appréhendable et visualisable.

Les expérimentations ont été menées dans le cadre du projet SMARTLOIRE. En conséquence, le jeu de données utilisé s'ancre dans le domaine touristique et utilise des instances réelles de l'ontologie DATAtourisme. Un modèle de Markov aléatoire est proposé afin de générer des séquences virtuelles de séjours touristiques en Centre-Val de Loire selon des profils prototypiques utilisateurs définis.

Les résultats obtenus par notre approche sont très prometteurs et montrent un score ARI de 0.83. En perspective, nous prévoyons d'appliquer notre approche à des séquences réelles d'éléments sémantiques multidimensionnels, notamment dans le cadre du projet MOBI'KIDS. De plus, lorsque nous disposons de la durée des éléments, il est pertinent de coupler notre similarité pour élément multi-dimensionnel sémantique à la mesure FTH, présentée dans le chapitre précédent. Ainsi, nous prévoyons de nouvelles expérimentations avec un jeu de données approprié et réel afin de réunir et tester l'ensemble de nos propositions dans un contexte unifié.

Chapitre 8

Une méthodologie pour l'analyse et la découverte de comportements

Publication

C. Moreau, T. Devogele, L. Etienne, V. Peralta, C. de Runz *Methodology for Mining, Discovering and Analyzing Semantic Human Mobility Behaviors*, *submitted (Dec. 2020) in* Data Mining & Knowledge Discovery (2020)

8.1 Cadre méthodologique

L'exigence de clarté au sein des procédés de découverte de connaissances et de machine learning est une préoccupation grandissante au sein de la littérature scientifique et du grand public [13]. De fait, lorsqu'un projet mêle différents acteurs et évolue sur des ensembles de données complexes, il nous semble important de faire preuve de pédagogie et de transparence quant aux découvertes effectuées par les procédés algorithmiques. Or, ces qualités ne tiennent que si l'on est capable d'expliquer à la fois les algorithmes et données en jeu. En conséquence, nous pensons que l'ajout d'indicateurs et de visualisation en charge de la description statistique des données (préliminaire et post-process) peut permettre une meilleure compréhension pour les experts et les utilisateurs des processus d'extraction de connaissances mis en place. Dans cette section, nous revenons sur les différentes motivations de tels ajouts méthodologiques puis nous développons la mise en place d'un cadre d'étude pour l'analyse et la découverte de comportements appliqué aux séquences de mobilité sémantique.

8.1.1 Motivations

La capacité à définir des types de comportements au sein d'un ensemble de séquences sémantiques dans un environnement non supervisé, c'est-à-dire où l'on a une *absence de vérité terrain* quant à la nature des différents profils comportementaux à découvrir est une problématique majeure que nous avons jusqu'alors passée sous silence. Contrairement aux algorithmes d'apprentissage supervisé où l'on dispose d'une base de données d'apprentissage, annotée par des experts, et dont le but est de généra-

ser raisonnablement les connaissances apprises sur des données encore non observées, les méthodes de clustering que nous utilisons (voir section 4.2) ne disposent pas de telles connaissances préalables. L'objectif des méthodes d'apprentissage non supervisé consiste alors à trouver une redescription des données disponibles pour en faire ressortir les régularités intéressantes et éventuellement inattendues. Dès lors, il est très difficile d'évaluer objectivement la réussite de ces algorithmes, d'autant plus lorsque les données sont complexes et mêlent des informations de différentes natures comme dans le cas des séquences de mobilité sémantique. En outre, les choix du nombre de classes (i.e., clusters) à retenir et de la mesure à utiliser font partie des problèmes difficiles de l'apprentissage non supervisé.

Il existe plusieurs critères reposant généralement sur une interprétation géométrico-topologique pour évaluer la qualité des clusters découverts (voir section 4.2). Par exemple le fait que les classes soient suffisamment denses et séparées entre elles comme le suggère l'indice de silhouette [199]. Or, même si ces critères peuvent être un indice pour mesurer la qualité du processus de clustering en tant que tel, ils ne possèdent que peu de potentiel pour l'interprétation des clusters en une logique métier, une symbolique claire et évidente pour l'utilisateur final. Il nous semble primordial et impératif de pouvoir comprendre aisément le sens sous-jacent de chaque cluster établi par le processus de clustering, non seulement pour des raisons éthiques afin de communiquer avec certitude, véracité et efficacité les connaissances découvertes mais aussi dans un objectif d'amélioration du processus (e.g., détecter les potentielles erreurs, biais du système). Dans les deux cas, le jugement d'un expert (concepteur ou métier) peut permettre de valider la pertinence des résultats du processus de clustering, apporter un éclairage contextuel, voire de guider l'exploration des données dans leur amélioration.

Ainsi, nous proposons ici un cadre méthodologique d'analyse (*data pipeline*) interactif et explicatif afin de permettre aux différents experts, métiers et scientifiques de la donnée, de réfuter ou valider la pertinence des comportements découverts. L'intégration de données contextuelles telles que les variables soci-démographiques des individus doit être possible à l'issue du processus d'extraction des connaissances. Un tel cadre de travail s'inscrit dans une perspective pluridisciplinaire et *human in the loop* où l'expertise humaine et les découvertes liées aux méthodes d'intelligence artificielle viennent s'enrichir mutuellement offrant du même coup une transparence et une intelligibilité quant aux comportements extraits.

Eu égard aux techniques et méthodes abordées en État de l'art (chapitre 4), nous proposons dans la figure 8.1 une méthodologie qui reprend et enrichit les concepts développés dans la figure 4.1 (section 4.2), notamment sur les aspects d'analyse du jeu de données et des clusters. Nous détaillons en section suivante les différentes analyses dans le cadre de la découverte de comportements dans les séquences de mobilité sémantique. Bien qu'orientée sur l'analyse de séquences d'activités, cette approche méthodologique demeure générique et applicable à tout type de contexte.

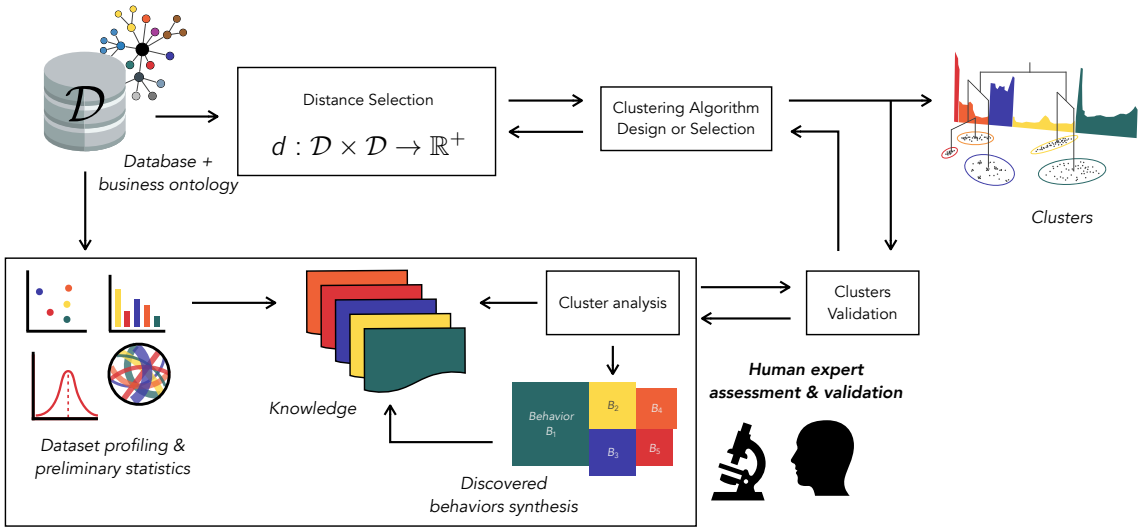


Figure 8.1 – Méthodologie de découverte de connaissance et d'analyse / *data pipeline* pour les séquences de mobilité sémantique

Préalablement, nous proposons de relier les données sémantiques disponibles à une ontologie permettant à la fois de faciliter les processus de comparaison et les raisonnements sur les données (voir section 3.1). L'ajout majeur consiste en la définition d'un cadre de travail pour l'analyse des données où la connaissance est subordonnée par le jugement humain de l'expert et/ou du concepteur d'algorithme, au sein même du processus de découverte de connaissances. Ce cadre de travail est constitué principalement de deux aspects : (i) d'une phase d'analyse statistique préliminaire et d'un profilage du jeu de données afin de bien comprendre leur nature, les lois statistiques, corrélations qui sous-tendent les données mais aussi la qualité globale du jeu de données. (ii) Un ensemble d'indicateurs complémentaires (détaillé table 4.3, section 4.3.1) afin d'analyser, visualiser et résumer les caractéristiques de chaque cluster en termes de comportement et d'en obtenir une image mentale compréhensible.

Dans la sous-section suivante, nous présentons plus en détail les indicateurs et modes de visualisation retenus pour l'analyse des séquences de mobilité sémantique.

8.1.2 Analyse de séquences de mobilité sémantique

Les séquences de mobilité sémantique sont difficiles à analyser en raison de leur combinaison des dimensions temporelle (qui peut être abordée selon l'ordre ou la durée des activités dans la séquence) et sémantique. Comme nous l'avons vu à la section 3.2, les séquences sémantiques de la mobilité humaine ont tendance à suivre des lois statistiques précises. La fréquence de visite des lieux induit une répétition des activités qui peut être modélisée par une loi de Zipf. Les séquences sont principalement structurées par quelques motifs topologiques (daily patterns) et sont caractérisées

par une faible entropie et une distribution de Poisson quant au nombre d'activités effectuées.

Par conséquent, pour assurer la qualité d'un ensemble de données de séquences de mobilité sémantique en termes de propriétés susmentionnées et en obtenir une compréhension préliminaire nous proposons l'utilisation d'un ensemble d'indicateurs, détaillés table 8.1.2 et basé sur ceux décrits table 4.3. Aussi, bien que cette étude se soit concentrée sur les séquences de mobilité, la méthodologie proposée est générique et peut être utilisée pour analyser tout type de jeu de données de séquences sémantiques. Au demeurant, une méthodologie analogue a été mise en oeuvre dans [166] pour l'étude et le profiling de comportements d'exploration de base de données.

Dans la section suivante, nous présentons une analyse exhaustive – de l'étude préliminaire des données à la découverte de comportements de mobilité – menée à l'aide de la méthodologie présentée et appliquée au contexte de la mobilité sémantique. Cette étude est réalisée sur l'ensemble des données récoltées auprès de l'enquête Ménage-Déplacement Rennes 2018.

Id	Technique	Mode de Visualisation	Utilisé pour		Exemple
			Jeu de données gobal	Clusters	
		Distribution des fréquences			
1	Distribution des longueurs	Histogramme, boîte à moustaches	×	×	Figs. 8.4, 8.11
2	Distribution des états	Histogramme, Barres empilées	×	×	Figs. 8.3, 8.12
		Transitions			
3	Matrice Origine-Destination	Diagramme de flux circulaire	×	×	Figs. 8.5, 8.14, 8.15
4	Daily patterns	Réseaux et histogramme	×	×	Fig. 8.6
		Désordre			
5	Entropie & prédictibilité	Diagramme de densité	×		Fig. 8.8
6	Symboles distincts	Boîte à moustaches	×		Fig. 8.7
		Lien statistique			
7	Résidus de Pearson	Diagramme mosaïque		×	Fig. 8.13
		Dispersion			
8	Rayon et diamètre	Table		×	Tab. 8.3
9	Silhouette	Table		×	Tab. 8.3
		Résumé			
10	Éléments prototypiques	Emojis sequence		×	Tab. 8.4
11	Nuage de mots	Mosaïque de mots		×	Fig. 8.17

Table 8.1 – Indicateurs retenus pour l'analyse des séquences de mobilité sémantique

8.2 Cas d'étude

Pour mettre à l'épreuve la méthodologie proposée, nous avons utilisé un ensemble de séquences de mobilité réelles obtenues à partir d'un sondage national nommé Enquête Ménage-Déplacement [1] (EMD). L'objectif des EMD est de fournir un instantané des déplacements effectués par les résidents d'une zone métropolitaine donnée afin d'aider les urbanistes, acteurs sociaux et politiques à mieux comprendre les comportements de mobilité, gérer l'installation et le déploiement de nouvelles infrastructures urbaines et mesurer leur impact et les changements induits dans le temps. Dans cette section, nous décrivons les données EMD en matière de volumétrie, de représentativité statistique et de méthodologie de recueil. L'ensemble de données est complété par une ontologie de domaine décrivant la sémantique des activités que nous détaillons. Sur la base de la méthodologie précédente, une étude statistique et une analyse générale de l'ensemble de données apporte une vision globale quant à la qualité et la compréhension des données. Enfin, la section finale présente le processus d'extraction des clusters de mobilité ainsi qu'une interprétation en termes de comportements à l'aide des indicateurs précédemment sélectionnés table 8.1.2.

L'ensemble des expérimentations et données peuvent être trouvées sur notre Github [2].

8.2.1 Les données EMD

Description de l'EMD Rennes 2018 Le jeu de données étudié "EMD Rennes 2018" est une une enquête Ménage-Déplacement réalisée dans la ville de Rennes et ses environs (département Ille-et-Vilaine). L'enquête a été réalisée de janvier à avril 2018 pendant les jours de semaine. Les données représentent 11 000 personnes (âgées d'au moins cinq ans) issues de 8 000 ménages. Cet échantillon est considéré comme statistiquement représentatif de 500 000 ménages et d'un million de résidants. Les détails de la méthodologie de collecte des données et leur qualité sont discutés dans [36], et une synthèse des résultats de l'enquête EMD Rennes 2018 est présentée dans [15].

Le jeu de données est constitué d'un ensemble de séquences de mobilité sémantique décrivant les activités réalisées par une personne sur 24h. La table 8.2 énumère les différentes étiquettes d'activités recensées dans les séquences de mobilité EMD. Deux classes principales sont représentées : les activités d'arrêt (STOP) et les activités de déplacement (MOVE). La première correspond aux activités statiques quotidiennes telles que "rester à la maison", "travailler" et "faire les courses". La seconde représente les activités de transport telles que "marcher" ou "conduire une voiture". Les séquences sont alors définies sur la base du paradigme STOP-MOVE de Spaccapietra et al. [222]. Chaque activité d'arrêt systématiquement est suivie d'une (ou plusieurs en cas d'intermodalité) activité de déplacement.

1. https://www.cerema.fr/fr/mots-cles/enquete-menage-deplacement-emd
2. https://github.com/Clement-Moreau-Info/EMD2018_DMKD

Col.	Act. agrégée	Emoji	Label et description
			STOP
	Maison		**1** : Domicile princial ; **2** : Résidence secondaire, hôtel
	Travail		**11** : Travail sur le lieu d'activité principal ; **12** : Travail à domicile ; **13** : Travail sur un autre lieu ; **43** : Recherche d'emploi ; **81** : Tournée professionnelle
	Étude		**21** : Crèche ; **22** : Primaire ; **23** : Collège ; **24** : Lycée ; **25** : Université ; **26** : Étude mobile (Primaire) ; **27** : Étude mobile (Collège) ; **28** : Étude mobile (Lycée) ; **29** : Étude mobile (Université)
	Achat		**30** : Visite d'un commerce ; **31** : Visite d'une grande surface ; **32** : Visite d'un centre commercial ; **33** : Achat petit ou moyen commerce ; **34** : Achat centre commercial ; **35** : Récupérer achats en drive
	Santé et administration		**41** : Soin de santé ; **42** : Démarche administrative
	Loisir		**51** : Sport, culture ou activité sociale ; **52** : Promenade et lèche-vitrine ; **53** : Restaurant / Cantine ; **54** : Visite familiale ou amicale ; **82** : Tournée shopping (suite consécutive de plus de trois activités 30)
	Accompagnement		**61**, **63** : Accompagner quelqu'un ; **62**, **64** : Récupérer quelqu'un ; **71**, **73** : Déposer quelqu'un à un arrêt de transport (gare, bus) ; **72**, **74** : Récupérer quelqu'un à un arrêt de transport
	Autre		**91** : autre (préciser en notes)
			MOVE
	Mobilité douce		**100** : marche ; **110** : location de vélo ; **111** : vélo ; **112** : passager vélo ; **193** : Skateboard, rollers et trottinette ; **194** : Fauteuil roulant ; **195** : Mobilité électrique douce (Trottinette, vélo, segway)
	Mobilité motorisée		**113** : Conducteur moto ($< 50cm^3$) ; **114** : Passenger moto ($< 50cm^3$) ; **115** : Conducteur moto ($\geqslant 50cm^3$) ; **114** : Passager moto ($\geqslant 50cm^3$) ; **121** : Conducteur voiture ; **122** : Passager voiture ; **161** : Taxi ; **171** : Véhicule professionnelle ; **181** : Conducteur camion (pour activité 81) ; **182** : Passager camion (pour activité 81)
	Transport en commun		**131** : Bus urbain ; **133** : Métro ; **138, 139** : Autre mode de transport public urbain ; **141**, **142** : Transport public départementaux ; **151** : Train
	Autre		**191** : Engin maritime et bateau ; **192** : Avion ; **193** : Autre mode (engin agricole, quad, etc.) ;

Table 8.2 – Description des activités de l'EMD Rennes 2018

Tout au long de cette section, nous utiliserons les codes d'activités et emojis pour représenter les séquences, à la fois dans les exemples et lors de l'analyse des séquences réelles. Parmi les 11 000 séquences de l'ensemble des données (correspondant aux 11 000 personnes interrogées), nous avons filtré celles qui ne contenaient aucun MOVE (correspondant aux personnes qui sont restées chez elles toute la journée). Nous avons ainsi obtenu un jeu de données final de 10 005 séquences de mobilité.

Exemple 12. *Considérons les activités suivantes effectuées par Alice durant la journée : "Elle part de son domicile à pied jusqu'à la station de métro afin de se rendre à son travail. Elle passe la matinée à son bureau, puis se rend à pied au restaurant. Alice retourne à son bureau à pied, elle travaille l'après-midi puis rendre directement en bus à son domicile."*

La séquence d'activités d'Alice S peut être modélisée à l'aide des concepts d'activités de la table 8.2 telle que $S = \langle 1, 100, 138, 11, 100, 53, 100, 10, 131, 1 \rangle$ *ou dans sa version agrégée en utilisant les emojis* $S_{agg} = \langle$🏠, 🚶, 🚃, 💼, 🚶, 🎬, 🚶, 💼, 🚃, 🏠$\rangle$.

L'étude du jeu de données EMD Rennes 2018 s'inscrit dans nos projets à deux niveaux :

- D'une part, sa construction est très semblable au jeu de données MOBI'KIDS. De fait, la plupart des codes activités sont les mêmes dans les deux projets car liées à la mobilité urbaine. Ainsi, l'analyse des jeunes populations (enfants et adolescents) nous a permis d'établir quelques résultats de référence pour MOBI'KIDS.
- Elle s'incorpore comme le pendant statistique auprès d'une collaboration avec la thèse d'Aline Meunin [152] portant sur l'élaboration d'un environnement de visualisation pour l'analyse multi-points de vue des mobilités quotidiennes. Nous renvoyons à ses travaux [153, 152] pour une analyse métier, plus à caractère géographique et socio-démographique des EMD.

Enfin, précisons que bien que disposant des informations, nous n'aborderons pas l'étude des durées des activités pratiquées [3]. Nous évoquons des pistes d'analyse en section 8.3.3. En conséquence, nous axons notre démarche vers une approche compositionnelle [150] de la mobilité, c'est-à-dire que nous accordons plus d'importance à l'ordre et à l'exécution des activités observée plutôt qu'au budget temps associé. En conséquence, nous ré-utilisons les définitions et modèles proposés lors du chapitre 5 portant sur CED. Nous prévoyons dans de futurs travaux l'étude et l'incorporation des durées liées aux activités grâce à une étude faite à l'aide de FTH (voir 6).

Ontologie de domaine EMD Les concepts d'activités détaillés dans la table 8.2 sont structurés dans un graphe de connaissances (i.e., ontologie) représenté figure 8.2. L'ensemble des liens du graphe sont de type hiérarchique tel que $x \rightarrow y$ indique

3. Ce choix est motivé à la fois pour des raisons de concision du discours, de complexité mais est aussi du à la chronologie des méthodes développées pendant la thèse. La prise en compte de la durée (chapitre 6) ayant été abordée après la mise au point de la méthodologie exposée dans ce chapitre.

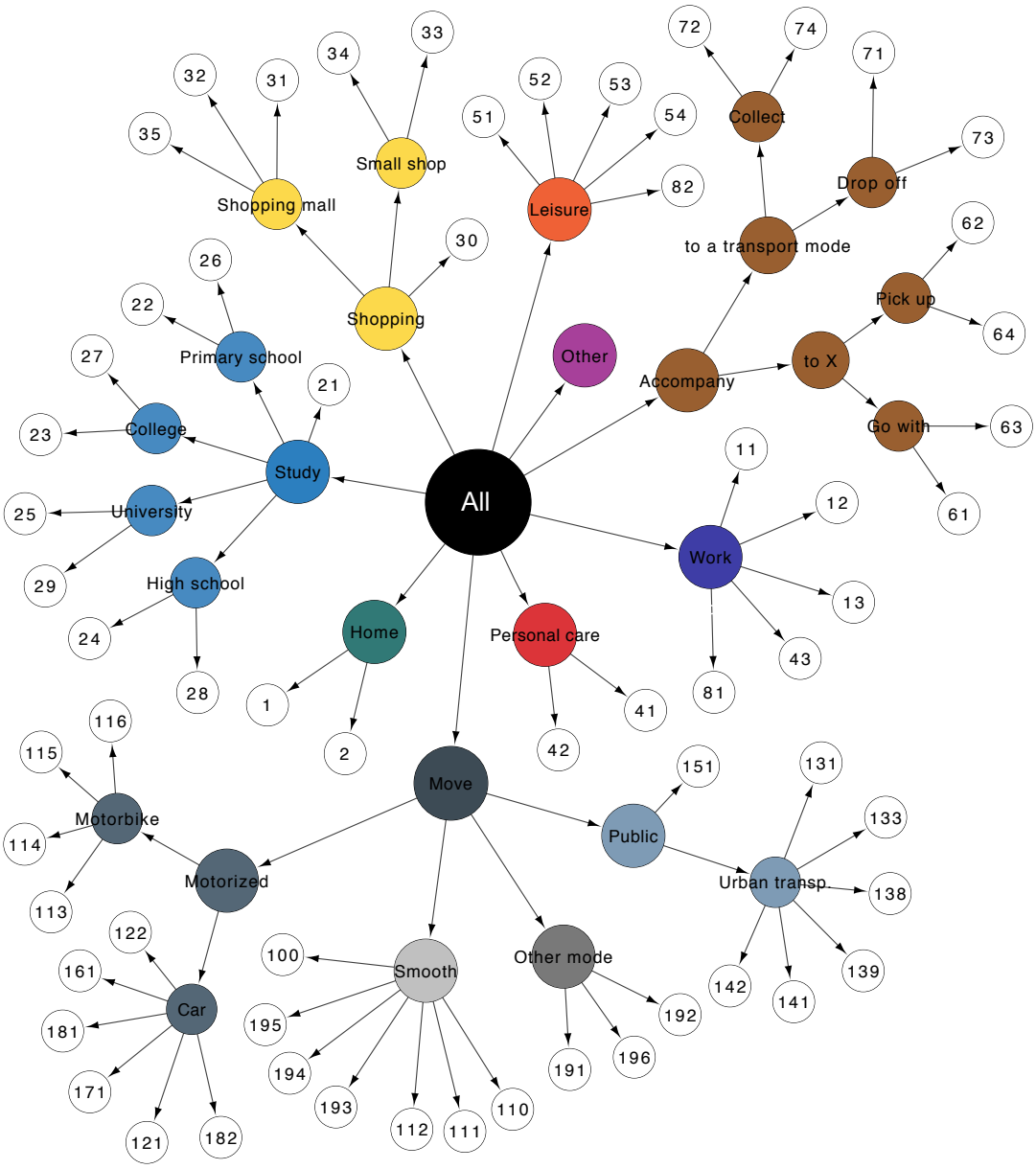


Figure 8.2 – Ontologie de domaine de l'EMD Rennes 2018

que le concept y est un concept spécifique de x (i.e., x est un holonyme de y). Par exemple le concept 133 (Métro) est un concept spécifique de "Transport urbain" qui est lui-même un concept spécifique de "Transport public", etc.

Ce graphe a été construit suivant la nomenclature de l'*Harmonised European Time-Use Survey* (HETUS) [72] qui décrit une taxonomie normée des activités quotidiennes et en accord avec des experts métiers (sociologues et géographes). Ainsi, chaque couleur correspond à une méta-catégorie représentant l'activité agrégée de premier niveau pour les STOP et de second niveau pour les MOVE. D'autres possibilités d'organisation des concepts peuvent être envisagées, chacune d'entre elles faisant référence à un contexte d'étude particulier ou un besoin spécifique métier ; la structure du graphe influençant les mesures de similarité entre les concepts comme détaillé en section 3.1.2.

8.2.2 Étude statistique des données EMD 2018

Afin d'obtenir une vision holistique des données capable de fournir à la fois du sens, un contrôle qualité, mais aussi une compréhension préliminaire de l'ensemble des séquences de mobilité sémantique, nous analysons ces données en utilisant l'ensemble d'indicateurs décrit table 8.1.2, section 8.1.1.

Dans un premier temps, nous étudions la fréquence d'apparition globale de chaque activité au sein des séquences. Par commodité nous avons séparé les activités STOP et MOVE selon leur catégorie. La figure 8.3 présente la distribution de fréquences de chaque activité dans l'ensemble de données. Comme prédit dans [219], cette distribution suit une loi de Zipf qui peut être visualisée dans les sous graphiques respectifs (b) et (d), indiquant que très peu d'activités concentrent l'essentiel de la mobilité humaine. De façon prévisible, les trois activités STOP les plus fréquentes sont 1 (domicile), 10 (travail) et 33 (achats en petit ou moyen commerce). Pour les activités MOVE, les activités les plus fréquentes sont 121 (conducteur voiture), 100 (marche à pied) et 122 (passager voiture). En outre, cette figure met en évidence les principales activités qui structurent la mobilité.

Le second indicateur concerne la longueur des séquences (i.e., ici le nombre d'activités réalisées par jour par un individu). Du à la représentation des séquences selon le modèle STOP-MOVE, nous constatons que très peu des séquences ont une longueur paire. Pour pallier à cette spécificité des séquences, nous considérons des intervalles de longueur I_k. La figure 8.4 présente la distribution du nombre d'activités pratiquées par séquence sur l'ensemble des données. La courbe verte représente la fonction de masse de la probabilité $P(|S| \in I_k)$ estimée selon une approximation Poissonnienne avec un paramètre λ obtenu à partir d'une estimation par maximum de vraisemblance ($\lambda = 1.36$). Une boîte à moustaches vient compléter cette analyse. Nous remarquons par ces deux graphiques que la majorité des séquences sont composées entre 5 et 13 activités avec une médiane de 9 activités par séquence.

Une autre méthode d'analyse des séquences sémantiques porte sur l'étude des tran-

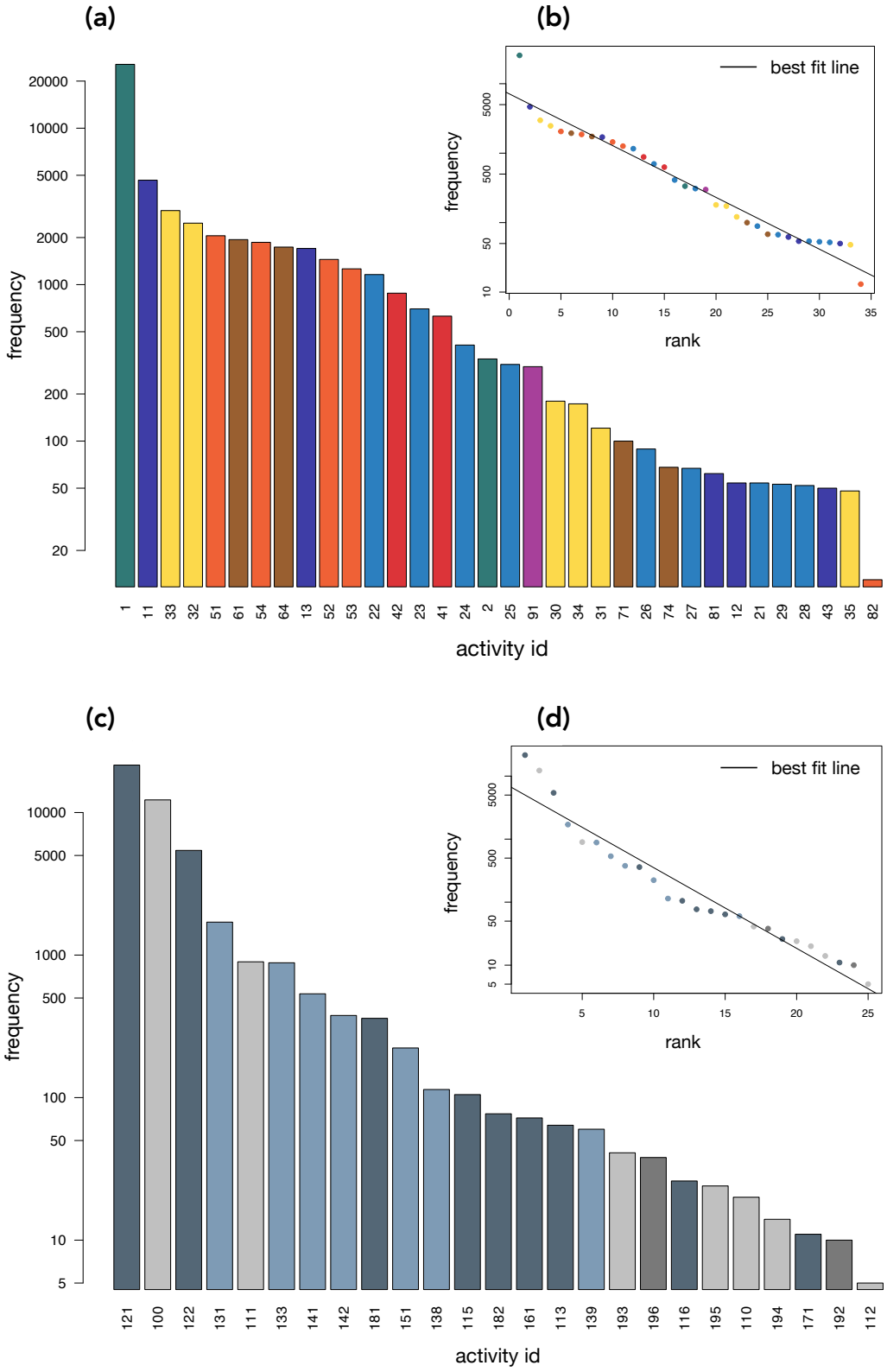


Figure 8.3 – Distribution de fréquences (échelle logarithmique) des activités STOP (a) et MOVE (c). Adéquation à un modèle Zipfien (b) et (d), les points correspondent aux activités dans le graphique à barres ci-dessous

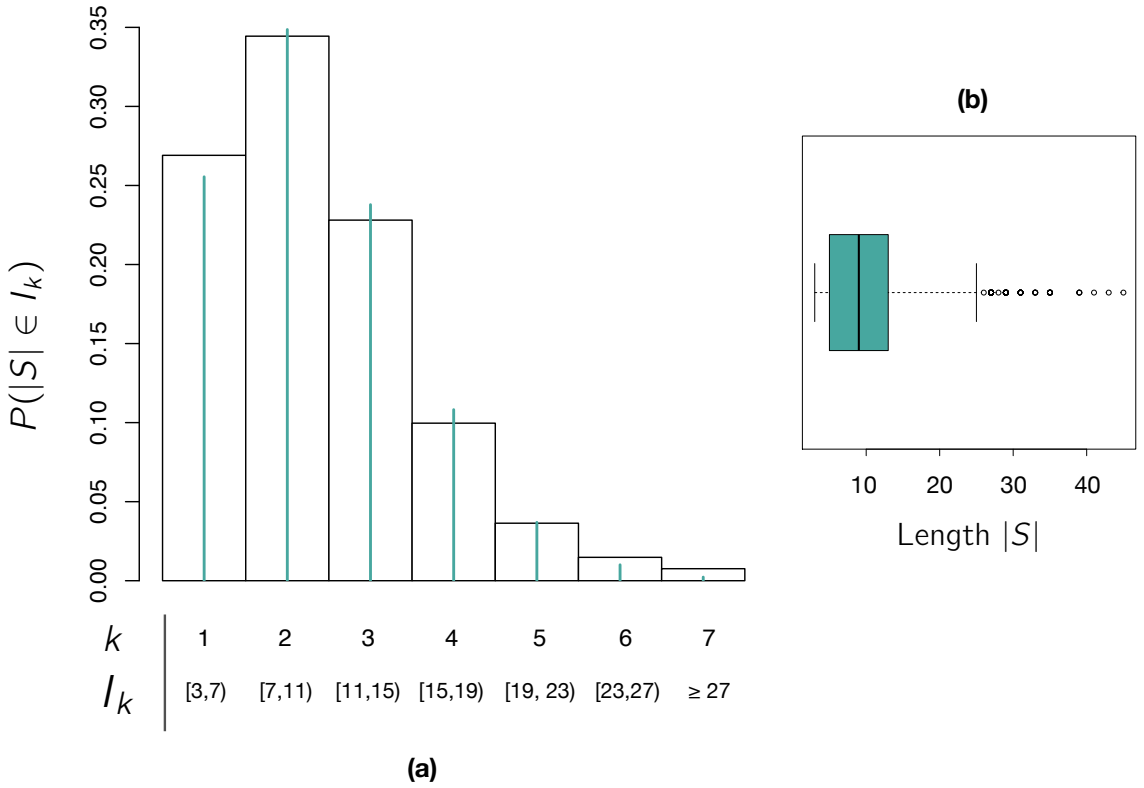


Figure 8.4 – Distribution du nombre d'activités par séquence (a) La distribution de $|S|$ pour un intervalle $I_{k\in\{1...7\}}$ est estimée par une loi de Poisson $P(|S| \in I_k) \approx \frac{1.36^k e^{-1.36}}{k!}$ (b) Boîte à moustaches du nombre d'activités par séquence

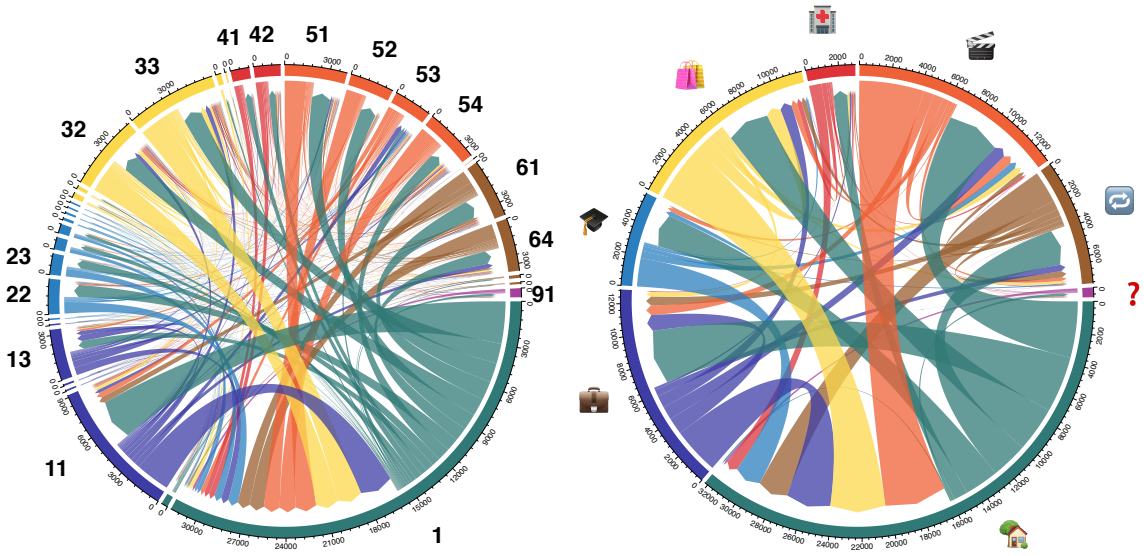


Figure 8.5 – Diagramme de flux entre deux activités STOP consécutives (i.e., connectées par un MOVE) (a) avec toutes les activités (b) avec les activités agrégées

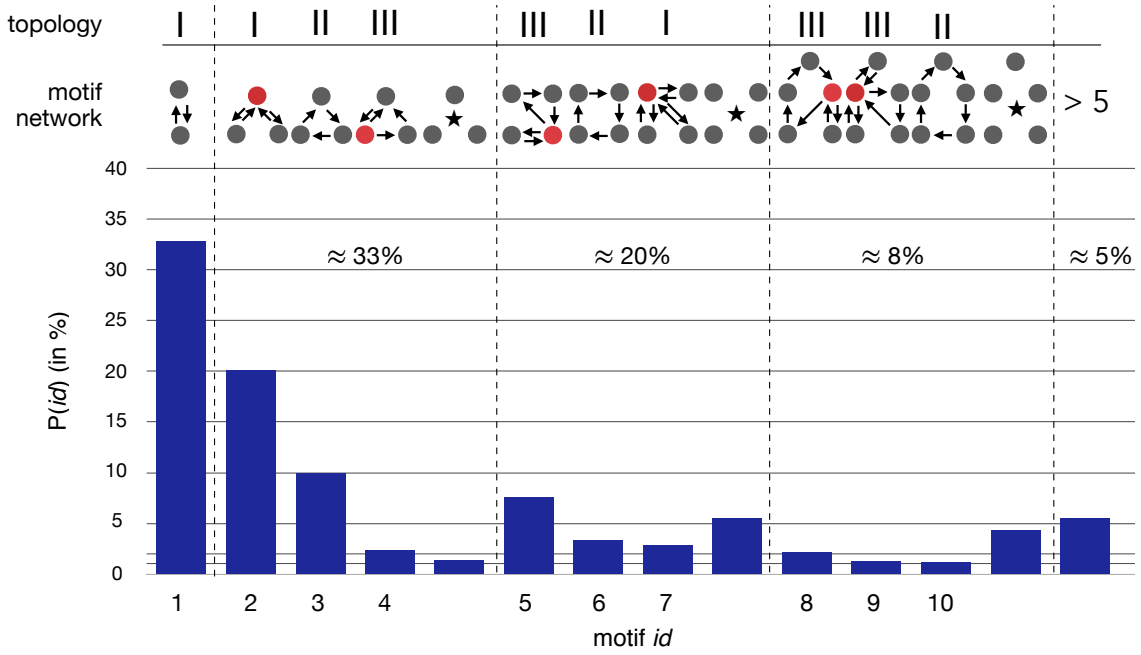


Figure 8.6 – Daily patterns. Les motifs sont regroupés en fonction de leur taille (séparés par des lignes pointillées). Les motifs notés $\star$ incluent tous les autres motifs avec $k \in \{3, 4, 5\}$ noeuds. Pour chaque groupe, nous montrons la fréquence qu'un motif donné possède k noeuds. Les noeuds centraux sont mis en évidence en rouge. Les motifs sont classés selon trois règles indiquant les propriétés topologiques : (I) les graphes avec des oscillations entre deux noeuds, (II) les graphes avec des cycles de 3 noeuds ou plus et (III) les graphes qui combinent les deux propriétés précédentes (I) et (II)

sitions entre les symboles en utilisant une matrice origine-destination. La figure 8.5 présente les transitions entre deux activités STOP consécutives dans l'ensemble des données à l'aide de diagrammes de flux circulaires (*Chord diagram*) [101]. L'ontologie nous permet de visualiser ces flux selon différents niveaux de granularité. Les flux au niveau le plus fin de l'ontologie sont détaillés en (a) tandis que la figure en (b) en présente une version agrégée. On constate que l'activité domicile principal (1 / 🏠) joue un rôle majeur et centralise la plupart des transitions.

Dans le contexte de la mobilité quotidienne, les transitions ont également été étudiées en termes de réseaux de mobilité individuels (*daily patterns*) afin d'identifier les motifs topologiques récurrents. Sur la base des travaux de Schneider et al. [206] et de l'algorithme présenté section 4.3.1 dans la catégorie "Transitions", nous avons extrait les principaux motifs des séquences. Comme le montre la figure 8.6, les motifs extraits et les fréquences sont cohérents avec les résultats présentés dans [206] (voir figure 3.3 (e)). Nous illustrons les trois motifs les plus fréquents pour les groupes de graphes dotés de k noeuds avec $k = 3, 4$ et 5 noeuds. Les graphes anonymes, notés $\star$, illustrent la proportion des autres graphes à k noeuds et non identifiés. Globalement, on peut voir que les motifs les plus fréquents ont moins de quatre noeuds et sont ceux

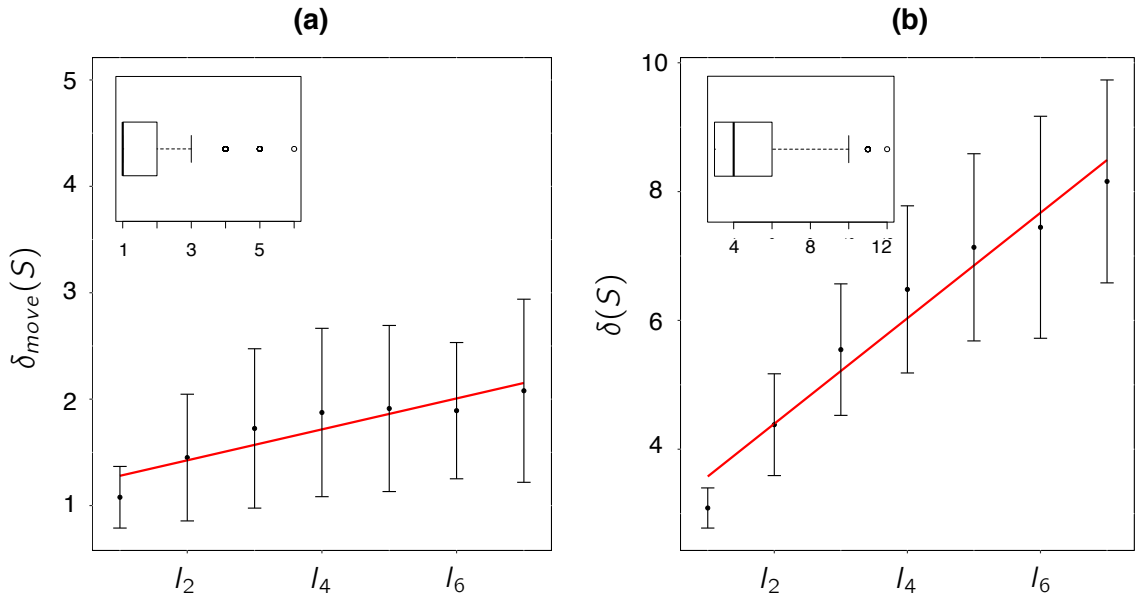


Figure 8.7 – Corrélation et régression linéaire entre les intervalles de longueurs I_k et le nombre d'activités distinctes (a) MOVE $\delta_{move}(S)$, (b) et total $\delta(S)$ par séquence S. La boîte à moustaches est montrée pour δ et δ_{move}. Le coefficient de corrélation est respectivement (a) $\rho = 0.4$ (b) $\rho = 0.8$. Le coefficient de régression linéaire a de $ax + b$ est respectivement (a) $a = 0.04$ (b) $a = 0.21$

présentant des oscillations (étiquettes I et III). Environ 87% des séquences suivent l'un des 10 motifs identifiés. En outre, le fait que les graphes fréquents contiennent de nombreuses formes d'oscillations démontre la présence de répétitions d'activités STOP dans les séquences.

Une autre technique permettant d'étudier la répétition et la régularité d'une séquence sémantique S consiste à calculer le nombre de symboles distincts $\delta(S)$ qu'elle contient. La figure 8.7 présente la corrélation entre la longueur d'une séquence $|S|$ et son nombre d'activités distinctes $\delta(S)$. L'axe horizontal représente la longueur de l'intervalle défini dans la figure 8.4 et l'axe vertical le nombre d'activités MOVE distinctes (a) d'activités distinctes $\delta_{move}(S)$ (resp. (b) (nombre d'activités distinctes STOP + MOVE δ) moyen avec écart-type. On peut voir que $\delta_{move}(S)$ reste globalement stable avec très majoritairement un ou deux modes différents pour toute longueur de séquence. Ce résultat est approuvé par la boîte à moustaches à l'intérieur du graphique qui illustre des points aberrants à partir de 3 modes de déplacement différents dans la même séquence. Le graphique (b) montre également que la plupart des activités sont répétées (On a $a < 1$ pour le coefficient d'interpolation linéaire $ax + b$.)[4]. Toutefois, le graphique exhibe également une corrélation forte entre le nombre d'activités distinctes et le nombre total d'activités effectuées en une journée (longueur de la séquence) indiquant que les activités STOP sont moins répétées que les activités MOVE.

4. Ce résultat est du en partie aux activités MOVE mais est confirmé dans le cadre d'analyse de δ_{stop}, non présentée ici.

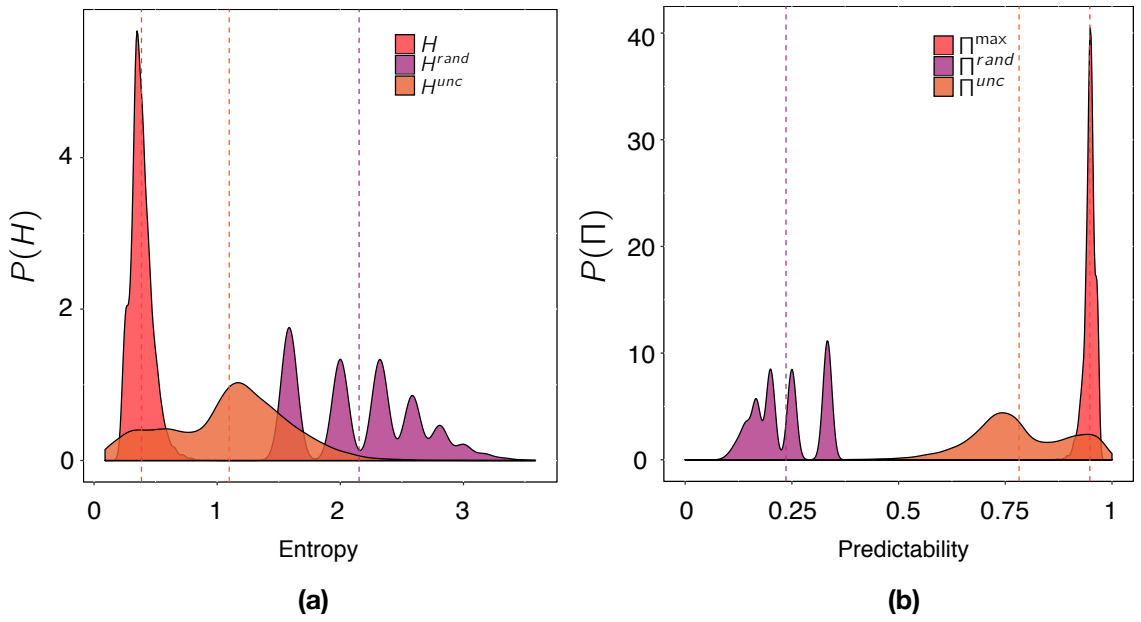


Figure 8.8 – Entropie et prédictibilité des séquences, les lignes en pointillés représentent la moyenne (a) Fonction de densité de probabilité de l'entropie réelle H, de l'entropie aléatoire H^{rand}, et de l'entropie non corrélée H^{unc} (b) Fonction de densité de probabilité des prédictibilités $\Pi^{\max}$, Π^{rand}, et Π^{unc}

La boîte à moustaches indique un 3ème quartile à partir de 5 activités différentes (STOP+MOVE) ce qui est cohérent la figure 8.6 des daily patterns qui indique qu'une large part ($\approx 66\%$) des séquences est composée de moins de 4 activités STOP [5]. Ces résultats, en accord avec l'intuition, viennent corroborer nos hypothèses et les études sur le caractère hautement répétitif de la mobilité et des activités humaines.

Enfin, l'entropie et la prédictibilité des séquences de mobilité peuvent être étudiées pour déterminer le degré de prévisibilité des individus. La figure 8.8 illustre la distribution respective pour ces deux variables. Les résultats obtenus sont consistants avec ceux donnés par [220] et présentent une faible incertitude de moins de deux activités concernant la future activité d'un individu [6]. Il convient de noter que ces résultats sont cohérents avec ceux présentés précédemment pour les valeurs de $\delta(S)$. La prévisibilité dans le cas aléatoire est $\Pi^{rand} \approx 0.24$, or on peut voir sur la boîte à moustaches de la figure 8.7 (b) que le nombre médian d'activités différentes dans une séquence est de 4. Cela indique qu'en choisissant l'activité de façon aléatoire sur celle qui compose la séquence, nous avons typiquement une prédictibilité d'1 activité sur 4, soit ≈ 0.25. Toutefois, contrairement aux résultats de Song et al, notre distribution $P(\Pi^{unc})$ culmine approximativement à $\Pi^{unc} \approx 0.78$ ce qui est relativement proche de la valeur de $\Pi^{\max}$. Cette constatation peut s'expliquer par le petit nombre d'activités

5. Ce résultat monte à $\approx 86\%$ si l'on compte les séquences dotées de 4 STOP distinctes.

6. La moyenne de l'entropie réelle H étant de 0.4 (pointillés rouge sur la figure 8.8), on en déduit que la quantité d'information nécessaire pour déduire l'activité est de $2^{0.4} \approx 1.4$.

distinctes dans les séquences mais aussi par le nombre, relativement faible, d'activités dans l'ensemble de données ainsi que les lois de Zipf qu'elles suivent (Figure. 8.3). Cela nous permet de prédire certaines activités clés (par exemple, maison, voiture, travail, marche à pied) avec un degré de certitude confortable. Enfin, la prédictibilité réelle $P(\Pi^{max})$ vient parachever nos observations en atteignant un pic avoisinant 0.95, démontrant que disposer de l'historique de mobilité de l'individu ainsi que ces transitions fournissent un très haut degré de prédictibilité quant à l'activité future de celui-ci.

8.2.3 Clustering de séquences de mobilité sémantique et découverte de comportements

Cette section décrit l'application du processus de clustering et de découverte de comportements au jeu de données EMD Rennes 2018. La première sous-section détaille l'ensemble des méthodes et algorithmes utilisés pour le regroupement des séquences de mobilité sémantique. La deuxième sous-section présente la phase de découverte et d'interprétation des comportements à partir des caractéristiques saillantes des clusters fournies par les indicateurs discutés en section 4.3.1 et résumés dans la table 8.1.2. Une discussion des résultats obtenus et des méthodes alternatives conclut cette section.

8.2.3.1 Processus de clustering : initialisation et validité

Nous continuons de suivre le processus méthodologique de découverte de connaissances établi sur la figure 8.1. Les données d'activités de l'EMD étant structurées

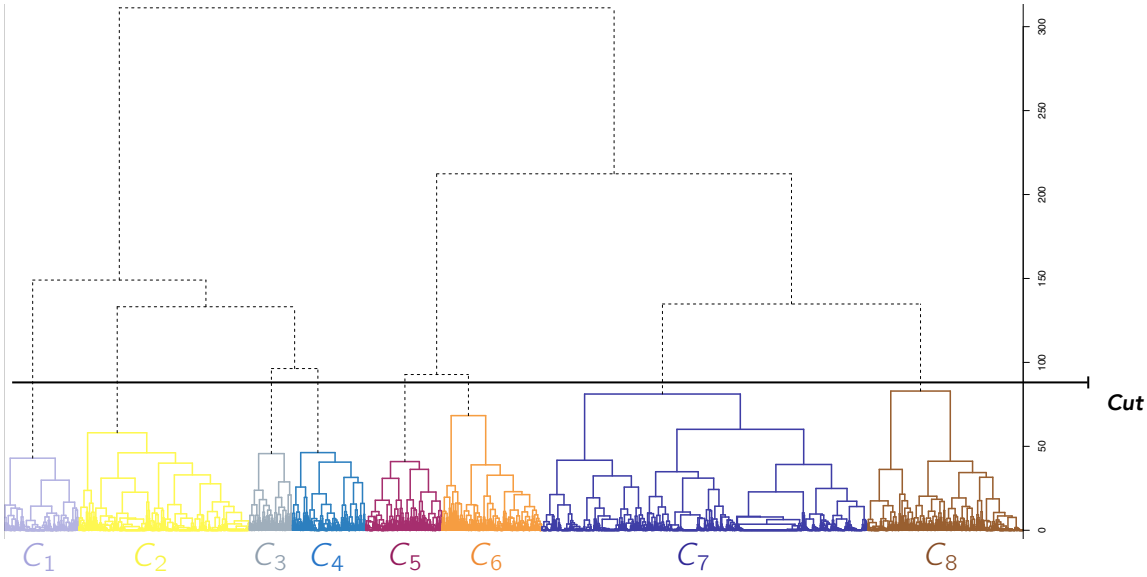


Figure 8.9 – Dendrogramme du jeu de données EMD 2018 via la mesure CED. La coupe engendre 8 clusters

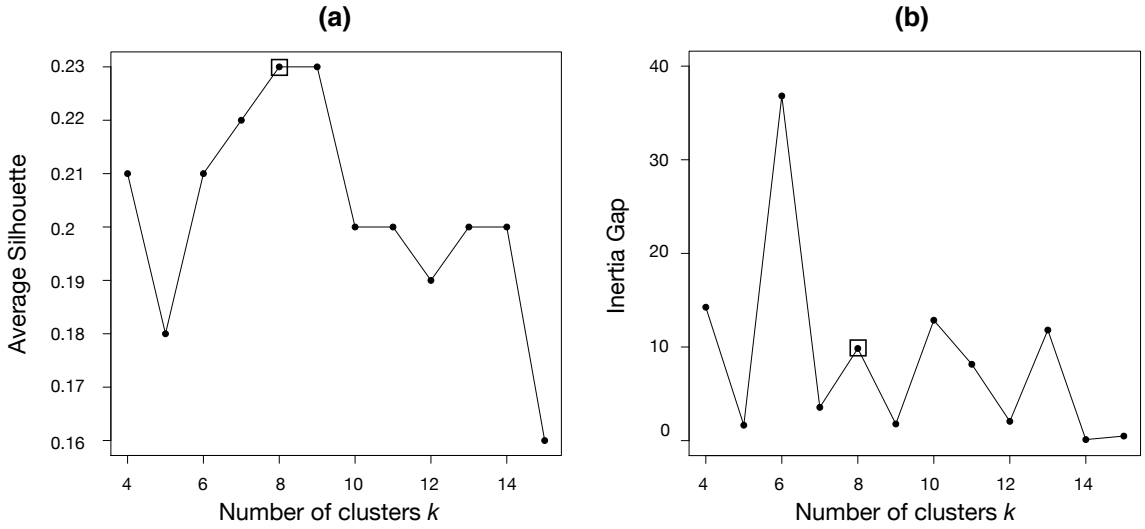


Figure 8.10 – Indices de qualité de clustering en fonction du nombre de clusters k retenus (a) silhouette moyen (b) Saut d'inertie

en ontologie (voir figure 8.2), nous utilisons la similarité de Wu-Palmer [244], déjà éprouvée dans les expérimentations et chapitres précédents, afin de comparer ces labels entre eux. Nous utilisons la mesure CED pour comparer les séquences de mobilité sémantique entre elles et adoptons la fonction μ définie équation 7.5 pour l'encodage du vecteur temporel et l'algorithme de regroupement hiérarchique Agglomerative Nesting (AGNES) [168] de la librairie `hclust` de R pour le clustering des séquences. Le dendrogramme et les clusters formés sont présentés figure 8.9. Dans notre étude, ne disposant pas de données annotées, nous avons utilisé deux indices de qualité interne pour sélectionner le nombre optimal de clusters. Le premier est le *saut d'inertie* qui représente l'écart entre deux agglomérations successives dans l'arbre hiérarchique ; plus l'écart est grand, plus le changement de structure des clusters est important. Le deuxième critère utilisé est l'indice *silhouette* qui reflète la compacité et la séparation des clusters. Cet indice est pertinent dans le cadre d'une utilisation d'un algorithme basé sur la maximisation de l'inertie inter-classe comme k-means ou le critère de Ward dans notre cas. Silhouette est défini dans l'intervalle $[-1, 1]$, une valeur plus élevée de l'indice indiquant un meilleur résultat.

La figure 8.10 présente les graphiques de silhouette moyen et du saut d'inertie selon le nombre de clusters k sélectionné. Les valeurs modérément faibles de silhouette peuvent être attribuées à la topologie particulière associée à CED combinée au critère de Ward ainsi qu'à la présence de valeurs aberrantes. Toutefois, nous relativisons nos résultats par rapport à l'état de l'art où Jiang et al., dans [113], obtenaient sur une tâche de clustering de séquences de mobilité sémantique semblable un score de silhouette ≈ 0.31 en usant d'une distance euclidienne combinée à une réduction par ACP et un k-means. Étant donné la nature des méthodes employées, le score de silhouette aurait du être plus élevé ce qui témoigne ici de la difficulté de la tâche de regroupement de tels objets.

Cluster C_i	$\|C_i\|$	% (au total)	$Sil(C_i)$	$diam(C_i)$	$diam^{95\%}(C_i)$	$rad(C_i)$	$rad^{95\%}(C_i)$
1	738	7.4	0.41	8.85	5	5.04	3.44
2	1673	16.7	0.37	20.03	8	12.66	4.68
3	423	4.2	0.01	20.81	7.74	7.95	5.53
4	719	7.2	0.12	26.64	7.21	12.51	5.7
5	747	7.5	0.18	23.42	6.9	9.86	5.35
6	981	9.8	0.1	24.34	8.15	14.59	6.11
7	3199	32	0.29	20	5.57	11.09	4.09
8	1525	15.2	0.07	28.5	7	10.14	4.65

Table 8.3 – Cardinal, silhouette, diamètre et rayon des clusters extraits

Néanmoins, nous nous référons aux graphiques de la figure 8.10 afin de déterminer le nombre optimal de clusters. Selon des critères métiers, nous exigeons un minimum de 5 clusters afin d'avoir un niveau de détails suffisant durant la phase d'analyse des clusters. Le graphique (a) suggère le choix de 8 ou 9 clusters. Le graphique (b) encourage fortement le choix de 6 clusters, mais 8, 10 ou 13 clusters sont également des choix pertinents. D'après ces résultats, nous avons décidé de retenir **8 clusters** pour la suite de l'analyse. Au demeurant, le choix de 6 clusters pour une analyse comptant moins de classes (et donc moins exhaustive), ou de 10 voire 13 clusters pour une analyse de plus grande ampleur reste tout à fait pertinent et réalisable.

Des informations supplémentaires concernant chaque cluster telles que les proportions, les indices de silhouette, les diamètres et les rayons sont données table 8.3. La mention 95% indique que nous avons filtré 5% des valeurs les plus extrêmes de la distribution. Ainsi, nous avons calculé les valeurs de diamètre et de rayon [7] des clusters telles que :

$$diam(C_i) = \max_{x,y \in C_i} \{d(x,y)\} \quad ; \quad diam^{95\%}(C_i) = P^{(95\%)}_{x,y \in C_i} \{d(x,y)\} \tag{8.1}$$

et

$$rad(C_i) = \max_{x \in C_i} \{d(x,m)\} \quad ; \quad rad(C_i)^{95\%} = P^{(95\%)}_{x \in C_i} \{d(x,m)\} \tag{8.2}$$

où m désigne l'élément médoid de C_i tel que calculé à l'équation 4.2 et $P^{(i\%)}$ désigne le i-ème percentile.

En outre, la réduction importante des valeurs entre les colonnes $diam(C_i)$ (resp. $rad(C_i)$) et $diam^{95\%}(C_i)$ (resp. $rad^{95\%}(C_i)$) montre la présence d'éléments statistiquement aberrants très en marge des points "centraux" des clusters.

7. Pour notre problème, ne disposant pas du centroid de l'ensemble des données, nous approximons celui-ci par le medoid tel que détaillé dans l'équation 8.2.

8.2.3.2 Découverte de comportements et interprétation des clusters

Dans cette section, nous réutilisons les indicateurs statistiques présentés table 8.1.2, enrichis de tests de significativité afin de déduire des caractéristiques notables des clusters extraits dans la section précédente. Cette phase d'analyse des clusters est cruciale dans le processus d'intégration des connaissances car elle permet à la fois de comprendre les modèles de comportements constitués par l'algorithme de clustering, mais aussi de vérifier la validité et l'interprétabilité de ces modèles. En outre, si l'analyse des clusters ne permet pas de conclure quant à certains types de comportements, alors notre méthodologie (voir figure 8.1) propose une rétro-propagation des résultats au niveau du choix de l'algorithme de clustering, voire de la distance.

Distribution des longueurs de séquences En premier lieu, nous analysons les longueurs (i.e., nombre d'activités) des séquences à l'intérieur des clusters. La figure 8.11 présente les boîtes à moustaches pour les longueurs de séquences dans chaque cluster. Comparées à la distribution des longueurs et à la boîte à moustaches de l'ensemble des données (diagramme le plus à gauche), on peut voir que les clusters C_1, C_2 et C_7 contiennent des séquences relativement courtes avec une longueur médiane de 6 à 7 activités, correspondant aux intervalles I_1 et I_2 dans la distribution établie figure 8.4. En revanche, les clusters C_5, C_6 et C_8 contiennent des séquences de mobilité plus longues mais aussi avec une plus grande dispersion des longueurs. De même, les groupes C_3 et C_4 ont des longueurs de séquences légèrement plus haute que la médiane et correspondant aux intervalles I_2 et I_3. Le chevauchement des boîtes à moustaches (c'est-à-dire l'existence de plusieurs clusters dont les séquences ont des longueurs d'ordre similaire) et la distribution des valeurs aberrantes au sein des diffé-

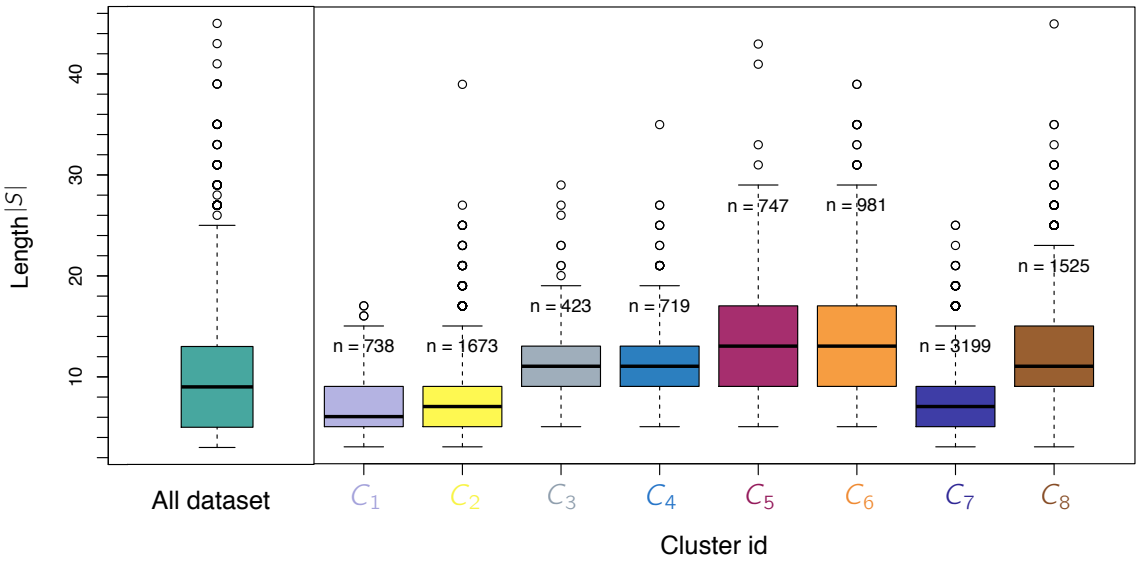


Figure 8.11 – Boîtes à moustaches des longueurs des séquences pour chaque cluster $C_{i \in \{1...8\}}$

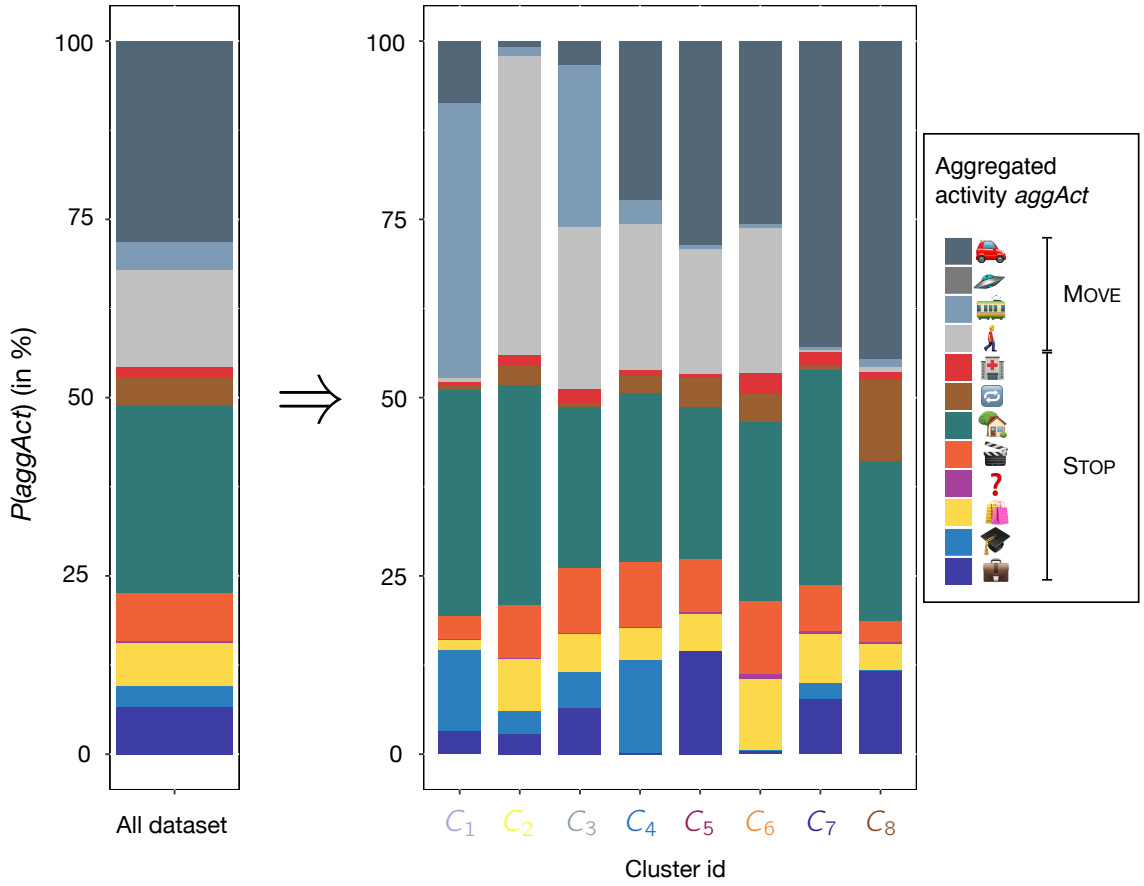


Figure 8.12 – Diagrammes à barres empilées de la répartition des activités agrégées pour chaque cluster $C_{i \in \{1...8\}}$

rents clusters indiquent que la longueur des séquences ne semble pas être un critère majeur de regroupement des séquences. Nous pensons qu'il s'agit d'un avantage de CED, par rapport aux autres mesures d'Optimal Matching, qui peut ainsi regrouper des séquences de longueurs différentes mais aux activités similaires.

Distribution des activités Concernant les distributions d'activités, la figure 8.12 présente les proportions d'activités agrégées[8] sous forme d'un diagramme à barres empilées. Un effet intéressant que l'on peut observer sur ce graphique est le fort effet de discrimination et de stratification des clusters en fonction des activités (MOVE). Par exemple, les activités de type "Transport motorisé" forment la quasi entièreté des MOVE pour les clusters C_7 et C_8. Les clusters C_2 à C_6 se distinguent par leur importante proportion d'activités de mobilité douce, tandis que C_1 et C_3 contiennent de nombreuses activités liées au transport public. Également, quelques catégories d'activités STOP forment un trait distinctif de certains clusters. Par exemple, les activités scolaires sont particulièrement présentes dans les clusters C_1 et C_4, tandis

8. Nous présentons les graphiques selon les activités agrégées de l'ontologie afin d'éviter une surcharge cognitive du lecteur.

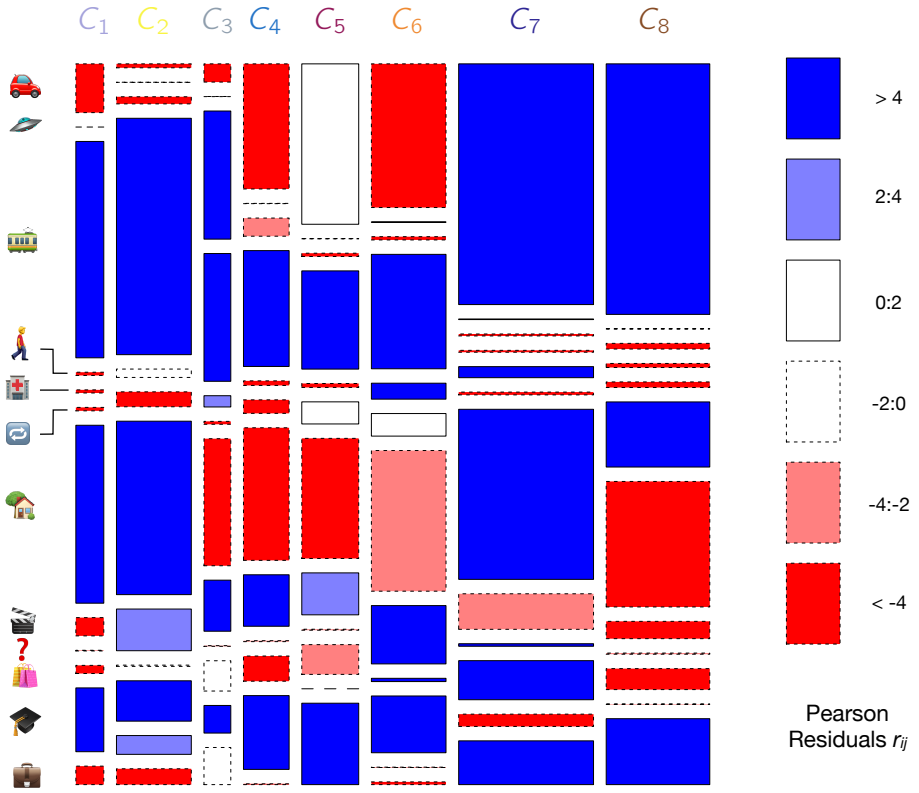


Figure 8.13 – Diagramme mosaïque et résidus de Pearson entre les activités agrégées et les clusters. V de Cramér = 0.3

que les activités professionnelles se retrouvent davantage dans les clusters C_5, C_7 et C_8. Par ailleurs, les activités d'accompagnement sont également courantes dans le cluster C_8. Enfin, soulignons que certains clusters se distinguent par leur absence de certaines activités, c'est le cas de C_6 qui ne répertorie que très peu d'activités liées au travail ou aux études.

Liens statistiques Clusters / Activités Afin de préciser ces sur et sous-représentations d'activités de façon quantitative, nous pouvons nous tourner vers le diagramme mosaïque représenté figure 8.13 où les résidus de Pearson indiquent l'écart de chaque cellule par rapport à l'indépendance des variables. Nous renvoyons à la section 4.3.1 pour un rappel sur les résidus de Pearson.

Nous rappelons plusieurs règles pour l'interprétation des diagrammes mosaïques. Ici, chaque ligne représente une activité agrégée $aggAct_i$ et chaque colonne un cluster C_j. Notons c_{ij} la cellule ligne i, colonne j correspondant à l'intersection des modalités $aggAct_i$ et C_j :

- La hauteur de c_{ij} est proportionnelle au nombre de $aggAct_i$ sous la condition d'être dans C_j [9].

9. en termes probabilistes, on noterait $P(aggAct_i|C_j)$.

- La largeur de c_{ij} est proportionnelle à la taille de C_j.
- La surface de c_{ij} est proportionnelle au nombre d'évènements $aggAct_i \cap C_j$.

La couleur d'une cellule c_{ij} indique la valeur du résidu de Pearson r_{ij} correspondant. Une cellule bleue indique une sur-représentation de l'activité agrégée $aggAct_i$ dans C_j. Une cellule rouge indique une sous-représentation. Sur la base de cette visualisation, il est commode de détecter les singularités de chaque cluster. Par exemple, on peut immédiatement voir qu'environ 30% du cluster C_1 est composé d'activités liées aux "Transport en commun". De plus, sur la base des résidus, nous pouvons facilement identifier, sous couvert d'une validité statistique, les activités caractéristiques d'un cluster, ainsi que celles sous-représentées. Par conséquent, la figure 8.13 complète et valide notre analyse précédente basée sur le diagramme à barres empilées. Le coefficient V de Cramér [50] calculé fournit quant à lui une information concernant la puissance de l'association entre les clusters et les activités agrégées. La plutôt bonne valeur de V (0.3) souligne la qualité de l'association entre nos clusters et les activités réalisées dans les séquences de mobilité sémantique. Le faible nombre de cellules blanches dans le diagramme mosaïque vient renforcer ce constat.

Analyse des transitions À propos des transitions entre les activités, les figures 8.14 et 8.15 présentent les différents diagrammes de flux pour chaque cluster. Précision ici que dans un souci de présentation et de concision des graphiques, nous avons filtrés les flux minoritaires de chaque cluster [10]. De plus, en conséquence de l'analyse présentée figure 8.7 qui indiquait que le mode de transport demeure globalement stable au sein des séquences, nous représentons ici uniquement les transitions entre deux activités STOP consécutives. Les flux sont représentés entre deux activités feuilles de l'ontologie afin d'explorer en détail le contenu des clusters. Par exemple, on peut voir que le cluster C_1 contient des activités liées aux études allant du collège (23) à l'université (25), alors que le cluster C_4 contient majoritairement des écoliers (22). Une explication de cette division peut être tirée du fait que le cluster C_1 concentre principalement ses activités MOVE autour des transports publics (voir figure 8.12), or ce mode de transport est très commun pour de nombreux adolescents et jeunes adultes. En particulier, cette association entre transport en commun et étude secondaire montre une étape majeure dans le processus d'autonomie des jeunes individus. À l'inverse, les enfants plus jeunes sont principalement accompagnés à l'école par leurs parents en voiture ou à pied, ce qui peut être observé au niveau du cluster C_4. Cette analyse est renforcée par la table 8.4, où les prototypes de centralité de C_1 mettent en évidence des séquences où l'individu se rend à son lieu d'étude en transport en commun puis rentre à son domicile par le même mode de mobilité. Le cluster C_4 lui concentre des déplacements en voiture (en tant que passager) et à pied, mais comprend aussi certaines activités STOP de loisir (51) et restauration/cantine (53).

10. Soit $OD = \{t_{ij}\}$ la matrice origine-destination avec t_{ij} le nombre de transitions entre les activités a_i et a_j, alors, nous représentons les flux dont l'effectif est supérieur à 5% du flux maximal, soit $t_{ij} > 0.05 \times \max_{i,j}\{t_{ij}\}$

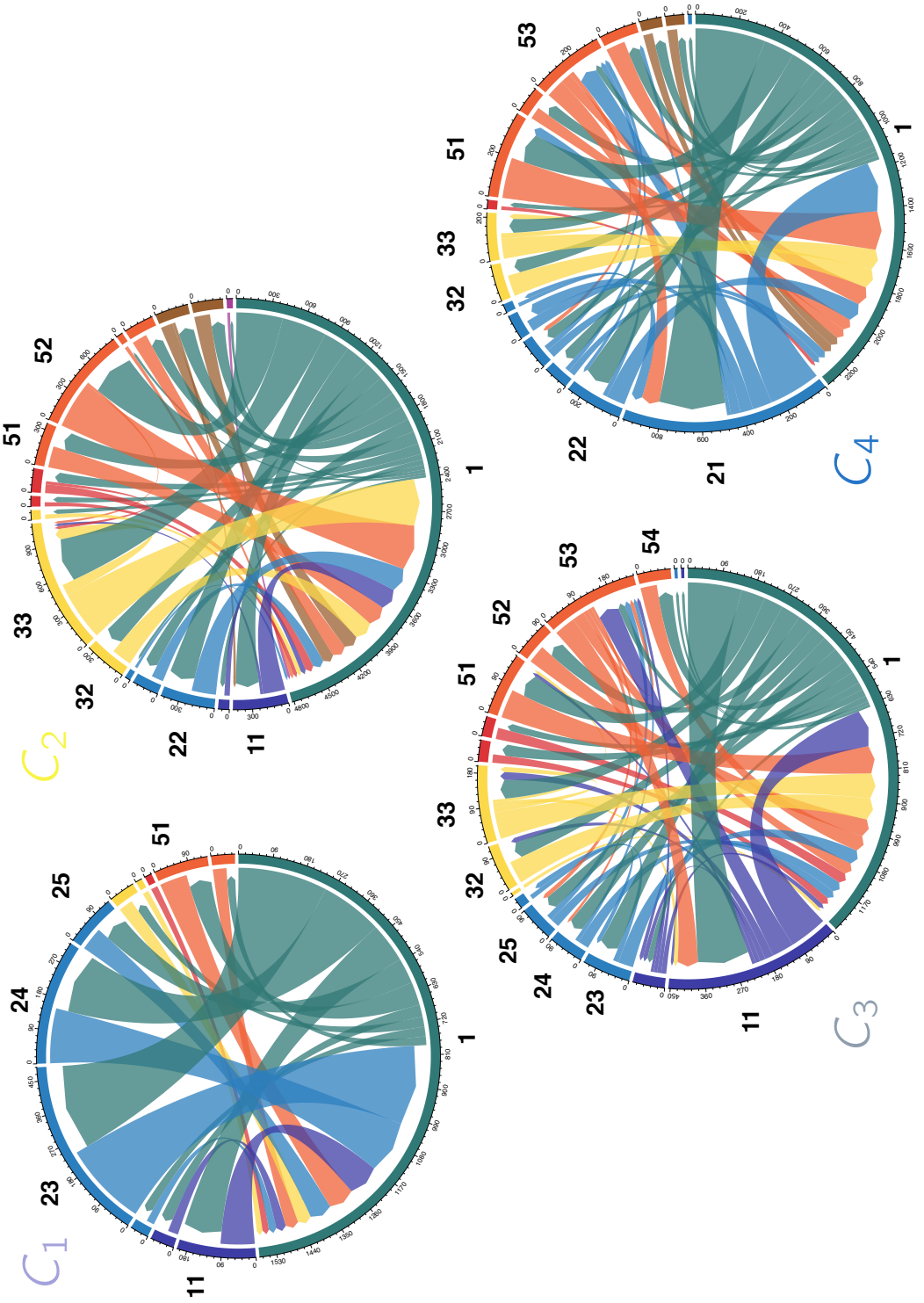


Figure 8.14 – Diagramme de flux des clusters $C_{i \in \{1...4\}}$

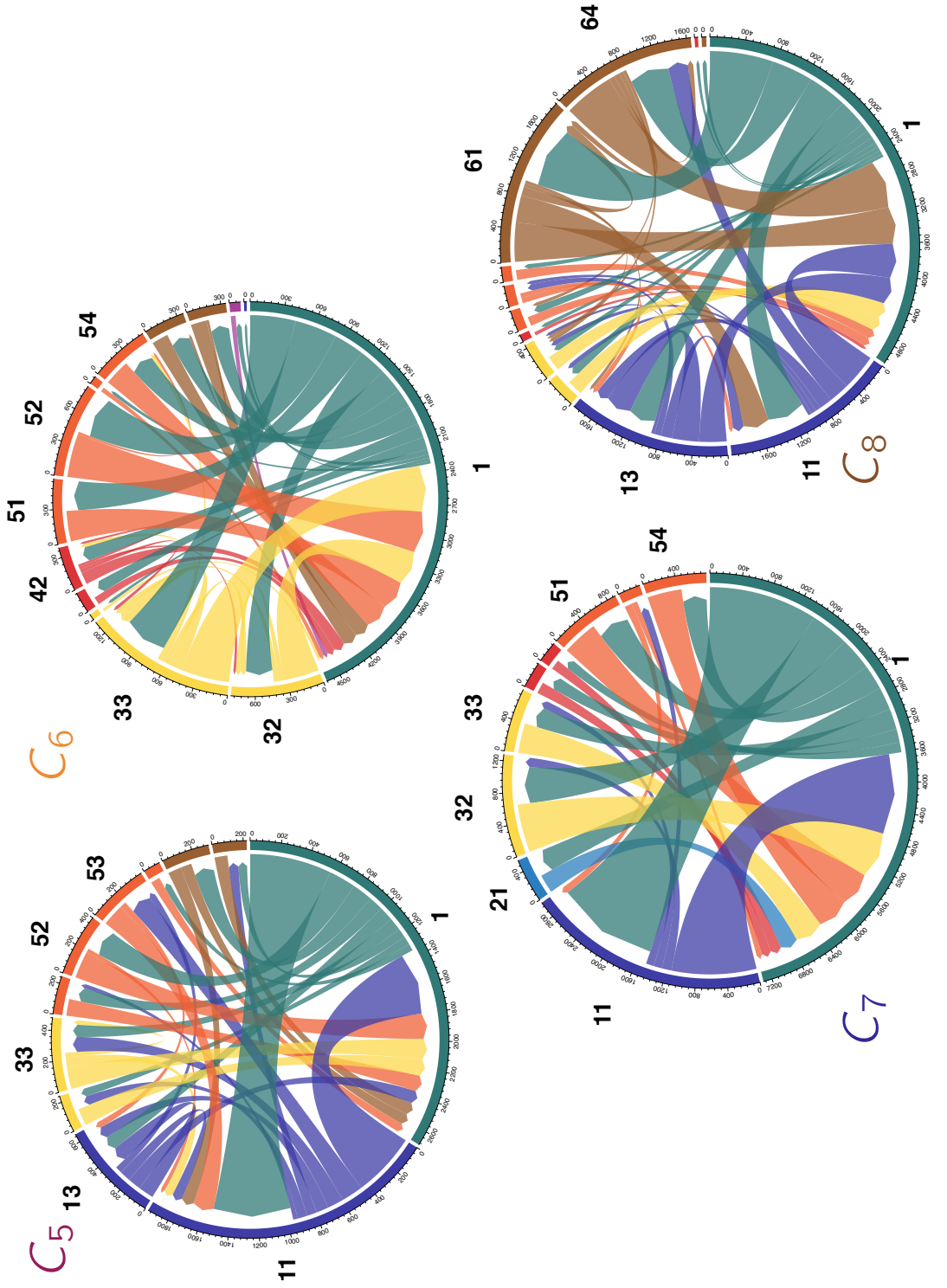
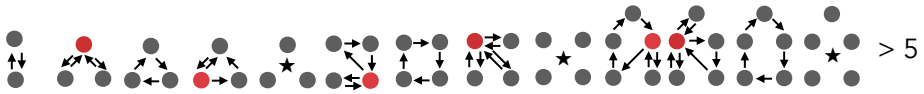


Figure 8.15 – Diagramme de flux des clusters $C_{i \in \{5..8\}}$

Cluster C_i	Medoid	Mode
1	⟨🏠, 🚃, 🎓, 🚗, 🏠⟩ ⟨1, 131, 23, 122, 1⟩	⟨🏠, 🚃, 🎓, 🚃, 🏠⟩ ⟨1, 141, 23, 141, 1⟩
2	⟨🏠, 🚶, 🛍️, 🚶, 🏠⟩ ⟨1, 100, 33, 100, 1⟩	⟨🏠, 🚶, 🛍️, 🚶, 🏠⟩ ⟨1, 100, 33, 100, 1⟩
3	⟨🏠, 🚃, 🚃, 💼, 🚶, 🎬, 🚃, 🏠⟩ ⟨1, 131, 133, 11, 100, 53, 131, 1⟩	⟨🏠, 🚃, 🎓, 🚶, 🎓, 🚶, 🎓, 🚃, 🏠⟩ ⟨1, 141, 23, 100, 27, 100, 23, 141, 1⟩
4	⟨🏠, 🚗, 🎓, 🚶, 🎬, 🚗, 🏠⟩ ⟨1, 122, 22, 100, 51, 122, 1⟩	⟨🏠, 🚗, 🎓, 🚶, 🏠⟩ ⟨1, 122, 22, 100, 1⟩
5	⟨🏠, 🚗, 💼, 🚶, 🎬, 🚶, 💼, 🚗, 🏠⟩ ⟨1, 121, 11, 100, 53, 100, 11, 121, 1⟩	⟨🏠, 🚗, 💼, 🚶, 🎬, 🚶, 💼, 🚗, 🏠⟩ ⟨1, 121, 11, 100, 53, 100, 11, 121, 1⟩
6	⟨🏠, 🚶, 🛍️, 🚶, 🏠, 🚗, 🎬, 🚗, 🏠⟩ ⟨1, 100, 33, 100, 1, 121, 52, 121, 1⟩	⟨🏠, 🚗, 🛍️, 🚗, 🏠, 🚶, 🎬, 🚶, 🏠⟩ ⟨1, 121, 33, 121, 1, 100, 52, 100, 1⟩
7	⟨🏠, 🚗, 💼, 🚗, 🏠⟩ ⟨1, 121, 11, 121, 1⟩	⟨🏠, 🚗, 💼, 🚗, 🏠⟩ ⟨1, 121, 11, 121, 1⟩
8	⟨🏠, 🚗, 🔁, 🚗, 💼, 🚗, 🔁, 🚗, 🏠⟩ ⟨1, 121, 61, 121, 11, 121, 64, 121, 1⟩	⟨🏠, 🚗, 💼, 🚗, 🏠⟩ ⟨1, 121, 13, 121, 1⟩

Table 8.4 – Séquences prototypiques centrales pour chaque cluster $C_{i \in \{1 \ldots 8\}}$

Concernant les clusters d'individus au comportement guidé par l'exercice d'une activité professionnelle (c'est-à-dire C_5, C_7 et C_8), nous observons certaines nuances parmi chacun d'entre eux. Dans le cluster C_5, le comportement typique semble être celui d'un individu qui se rend au travail en voiture (11, 13) puis se rend à pied dans un lieu de restauration (53) pour déjeuner avant de retourner au travail puis de rentrer chez lui en voiture. Ce scénario est corroboré par la séquence de mobilité medoid et la figure 8.16 qui présente les daily patterns de chaque cluster. En C_5, nous remarquons une tendance à l'oscillation entre deux activités avec un noeud central (ici le travail). Certaines activités périphériques peuvent également être ajoutées à la séquence sémantique telles que faire des courses (32, 33) après le travail, se promener ou faire du lèche-vitrine (52), ou encore accompagner/déposer quelqu'un (61, 64). Le cluster C_7 représente des individus qui se rendent majoritairement au travail (11) en voiture avant de rentrer chez eux, ce qui est révélé par les épaisses transitions visibles sur le diagramme de flux entre le domicile (1) et le lieu d'activité professionnelle (11). Cette interprétation est cohérente avec l'analyse de la longueur des séquences qui sont très courtes dans ce cluster et où la grande majorité des daily patterns sont pourvus d'une unique oscillation. Ce comportement de mobilité est le plus fréquent (près d'$1/3$ du jeu de données) et peut être interprété comme une routine quotidienne consistant à aller au travail en voiture, à faire occasionnellement des achats en centre commercial (32), activités de loisirs ou de sociabilité (51, 54) puis à rentrer chez soi. Le cluster C_8 concentre lui des individus qui accompagnent et récupèrent (61, 64) quelqu'un avant et après le travail (11, 13) avec une possible mobilité autour du



motif *id*	1	2	3	4		5	6	7		8	9	10			$\sum_j$
C_1	542	84	63	8	8	14	11	1	4	2	0	0	1	0	738
C_2	912	321	147	28	16	61	29	39	50	13	9	5	27	16	1673
C_3	15	168	51	10	6	48	13	14	35	8	1	9	28	17	423
C_4	65	219	40	18	6	89	34	34	56	29	19	17	44	49	719
C_5	7	213	12	9	11	95	6	33	108	38	16	5	81	113	747
C_6	38	312	27	28	13	110	14	85	60	33	40	8	88	125	981
C_7	1578	496	481	84	51	182	90	26	114	20	10	14	37	16	3199
C_8	147	211	179	46	26	165	142	54	134	73	35	56	131	126	1525
$\sum_i$	3304	2024	1000	231	137	764	339	286	561	216	130	114	437	462	10005

Figure 8.16 – Carte de chaleur avec résidus de Pearson des daily patterns pour chaque cluster $C_{i\in\{1...8\}}$. V de Cramér = 0.25

lieu d'activité professionnelle (13). De même que les déplacements de C_7, la mobilité de C_8 est quasi intégralement effectuée en voiture. De plus, dans ce cluster, les séquences de mobilité sont relativement longues et forment des schémas complexes avec généralement quatre activités STOP ou plus.

Enfin, certains clusters se distinguent par l'absence de certains éléments communs. Par exemple, les individus du cluster C_6 ne travaillent pas ou n'étudient pas et ont tendance à consacrer leur temps principalement aux activités de shopping ou de loisir. De plus, la boîte à moustaches de la figure 8.11 révèle que les séquences de mobilité de C_6 sont longues. Enfin, les clusters C_2 et C_3 sont particulièrement caractérisés par leurs activités MOVE. Les individus de C_2 se déplacent presque exclusivement à pied pour effectuer une seule activité, principalement des achats. Comparées aux séquences de mobilité de C_6, ces séquences sont relativement courtes et basées sur une seule oscillation entre le domicile et un autre lieu. Le comportement typique de C_3 est plus difficile à cerner en termes de STOP mais semble être caractérisé par des individus qui utilisent à la fois les transports en commun et la marche à pied pour se déplacer.

8.2.3.3 Résumé des comportements découverts

Grâce aux analyses précédentes des clusters, nous sommes en capacité d'extrapoler un type de comportement pour chaque cluster. La table 8.5 résume les huit comportements découverts, où les colonnes "Act. typiques", "Nb. Act" et "Réseau de mobilité"

Data : Ensemble des clusters $\mathcal{C} = \{C_1, ..., C_k\}$
Result : (ActTypiques, NbAct, RMob)
for $C_i \in \mathcal{C}$ **do**
  ▷ L'activité x est typique de C_i si son résidu de Pearson dans le cluster est supérieur à 4 (i.e., écart positif à la norme avec degré de confiance de 99%) et x fait partir du mode ou du medoid.
  $\text{ActTypiques}(C_i) \leftarrow \{x | x \in \Sigma \wedge \text{PearsonResiduals}(x, C_i) \geqslant 4 \wedge (x \in \text{mode}(C_i) \vee x \in \text{medoid}(C_i))\}$
  ▷ Label qualitatif du nombre d'activités dans les séquences (i.e., longueur). I_k réfère aux intervalles de la figure 8.4.
  $$\text{NbAct}(C_i) \leftarrow \begin{cases} \text{“Peu”} & \text{Si } median(\{|S| : S \in C_i\}) \in I_1 \\ \text{“Moyen”} & \text{Si } median(\{|S| : S \in C_i\}) \in I_2 \\ \text{“Bcp.”} & \text{Sinon} \end{cases}$$
  ▷ G est l'ensemble des daily patterns de C_i avec un résidu de Pearson $\geqslant 4$.
  $G \leftarrow \{id | \text{ PearsonResiduals(motif } id, C_i) \geqslant 4\}$
  ▷ On qualifie ces daily patterns.
  $\text{RMob}(C_i) \leftarrow \varnothing$
  for $id \in G$ **do**
    $$\text{RMob}(C_i) \leftarrow \text{RMob.}(C_i) \cup \begin{cases} \text{“Aller-retour”} & \text{Si } id = 1 \\ \text{“Aller-retour double”} & \text{Si } id = 2 \\ \text{“3-cycle”} & \text{Si } id = 3 \\ \text{“Complexe”} & \text{Si } |Node(id)| \geqslant 4 \\ \text{“Autre”} & \text{Sinon} \end{cases}$$
  end
end

(correspondant aux daily patterns) fournissent des qualitatifs simples et compréhensibles exprimant les caractéristiques saillantes globales des clusters. Par exemple, le qualificatif “Aller-retour” dans la colonne “Réseau de mobilité” signifie qu'une proportion significative des individus possède un graphe de mobilité isomorphe au motif 1, c'est-à-dire constitué d'un aller-retour (i.e., oscillation simple).

Par souci de concision, les activités typiques ont été extraites au niveau des activités agrégées (en utilisant des emojis). Enfin, la colonne "Comportement" contient une étiquette caricaturale défini manuellement du comportement, principalement dans le but de susciter une image mentale, des émotions et idées simples attachées à un stéréotype capable de recouvrir en bonne partie l'ensemble des séquences du cluster. Toutefois, une analyse des variables socio-démographiques doit venir corroborer nos hypothèses. Nous développons ces pistes dans la section suivante. Ainsi, nous pouvons résumer les comportements dans les clusters comme suit :

Cluster C_i	% (au total)	Act. typiques	Nb. act.	Réseau de mobilité	**Comportement**
1	7.4	{🚃, 🎓}	Peu	Aller-retour	*Les adolescents*
2	16.7	{🚶, 🛍}	Peu	Aller-retour	*Les promeneurs*
3	4.2	{🚃, 🚶, 🎓, 🎬}	Moyen	Aller-retour double	*Les transports mixtes*
4	7.2	{🚶, 🎓, 🎬}	Moyen	Aller-retour double	*Les écoliers*
5	7.5	{🚶, 💼, 🎬}	Bcp.	Aller-retour double, Complexe	*Les actifs décontractés*
8	9.8	{🚶, 🛍, 🎬}	Bcp.	Aller-retour double, Complexe	*Les shopping addicts*
7	32	{🚗, 💼}	Peu	Aller-retour, 3-cycle	*Routine quotidienne*
12	15.2	{🚗, 💼, 🔁}	Moyen	Complexe	*Les parents actifs*

Table 8.5 – Résumé synthétique des comportements découverts

- Le cluster C_1 contient une majorité de séquences de mobilité courtes, avec un seul aller-retour entre le domicile et le collège/lycée ou l'université et une utilisation importante des transports publics tels que les bus. Ce groupe est principalement constitué de comportements de mobilité évoquant des *adolescents*.
- Le cluster C_2 est caractérisé par des individus ayant une mobilité douce, principalement marche à pied, très importante mais dont les séquences sont dotées de peu d'activités de STOP. Leurs activités STOP préférées sont liées au shopping, commerce et autres flâneries. Nous les appelons les *promeneurs*.
- La principale caractéristique des individus et séquences en C_3 est qu'ils combinent des modalités de transports en commun et douces telles que la marche, souvent de façon successive. Ainsi, l'intermodalité est une caractéristique importante au sein de ce cluster. Nous les nommons les *transports mixtes*.
- Les *écoliers* sont principalement regroupés dans le cluster C_4 où l'on constate une forte proportion d'activités d'école primaire, suivies d'activités sportives ou culturelles. Ces individus se déplacent principalement à pied ou en voiture.
- Le comportement archétypale du cluster C_5 semble être celui d'un individu travaillant et déjeunant en extérieur, typiquement au restaurant, ou pratiquant une activité de promenade/lèche-vitrine entre deux activités professionnelles. Ces *actifs décontractés* réalisent leur mobilité en voiture entre le domicile et le travail et en marche à pied entre le travail et le lieu de restauration/loisir.
- Les individus du cluster C_6 possèdent la caractéristique principale de pratiquer de nombreuses activités de shopping et de loisir en lieu et place de travailler ou d'étudier. Nous appelons ces individus les *shopping addicts*.

- Le cluster C_7 est le plus grand cluster et contient près d'$^1/_3$ de l'ensemble des données. Les individus de C_7 effectuent principalement de courtes séquences de mobilité représentant un aller-retour en voiture entre leur domicile et leur lieu professionnel. Ce comportement, avec ses activités élémentaires (voiture, travail, et parfois achats dans un centre commercial) et ses schémas de mobilité, évoque une *routine quotidienne*.
- Enfin, C_8 représente un comportement similaire à celui de C_7 mais où les individus transportent généralement une autre personne en voiture, la dépose avant de partir travailler, puis la récupère après le travail. Ce comportement peut être assimilé à celui de parents accompagnant leur(s) enfant(s) à l'école le matin avant d'aller travailler puis qui les récupèrent le soir. Ainsi, nous désignons ces individus sous le nom de *parents actifs*.

La figure 8.17 présente un résumé graphique des clusters et des comportements correspondants. La surface de chaque carré est proportionnelle à la taille du cluster associé. Les couleurs et la composition font référence au dendrogramme de la figure 8.9.

Enfin, précisions ici que les indicateurs de résumé présentés dans cette section n'ont pas pour but de substituer les analyses précédemment présentées section 8.2.3.2 mais forment plutôt une aide à la compréhension globale, une porte d'entrée pour les experts métiers vers une étude plus précise et des indicateurs plus détaillés mais aussi plus complexes.

Discussion

Dans cette sous-section, nous avons présenté l'analyse et les résultats du processus de clustering selon la méthodologie introduite dans la section 8.1. Celle-ci a permis la découverte de plusieurs modèles intéressants et cohérents de mobilité sémantique. Ceux-ci sont résumés dans la table 8.5. Dans une démarche d'amélioration de notre processus de découverte, nous soulignons ici plusieurs perspectives, problématiques et alternatives pouvant être considérées lors de travaux futurs.

Tout d'abord, le choix de la mesure de similarité / dissimilarité a un impact significatif sur les résultats du clustering. Dans ce cas d'étude, nous avons utilisé la mesure CED, dont plusieurs paramètres doivent être définis au préalable : la mesure de similarité entre les activités, l'ontologie et le vecteur temporel sont autant de paramètres qui influencent les résultats du clustering. Nous avons ajusté expérimentalement chaque paramètre et nous nous sommes référés aux connaissances métiers pour la construction de notre ontologie. Pour autant, l'utilisation d'autres mesures présentées en section 3.3 est possible.

Prochainement, nous souhaiterions également mettre à l'épreuve la mesure FTH [11] définie chapitre 6, afin de prendre en compte la notion de durée au sein des EMD.

11. FTH n'a été développée qu'après la réalisation du cas d'étude sur les EMD.

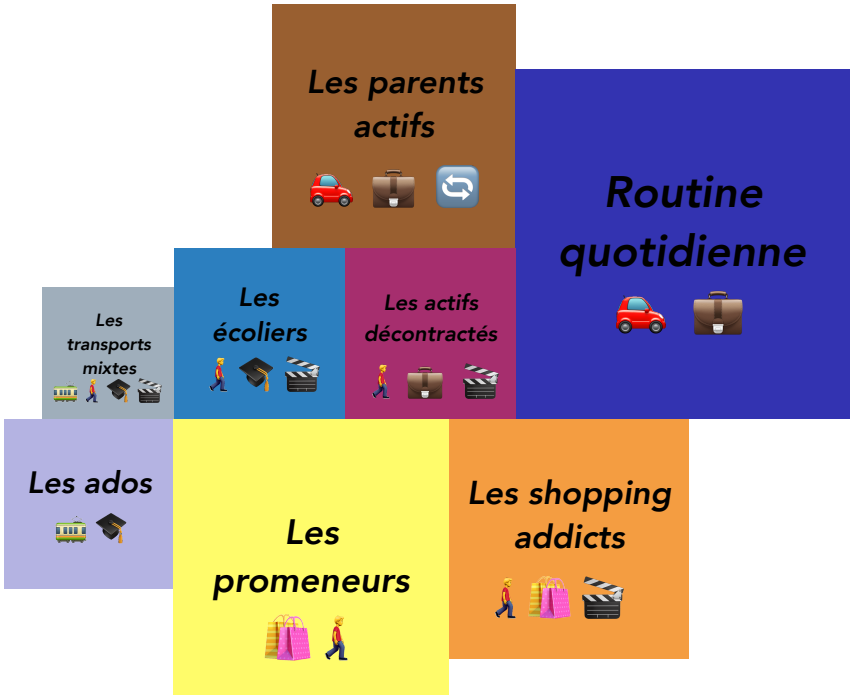


Figure 8.17 – Résumé graphique sous forme d'une mosaïque de mots de l'ensemble des clusters et comportements découverts

Toutefois, l'analyse des clusters issus de la mesure FTH réclame l'incorporation d'indicateurs supplémentaires afin de capter l'emprise temporelle des activités dans une perspective, non plus compositionnelle du temps, mais structurelle. Cette analyse peut être menée en inspectant non plus la fréquence des activités menées mais plutôt les budgets-temps réels associés (e.g., compter 30min + 10min de 🚗 plutôt que 2 occurrences de 🚗). Nous développons ces aspects dans la section suivante.

Le deuxième point discuté concerne le choix de l'algorithme de clustering. Comme indiqué dans la table 8.3, les diamètres et rayons des clusters indiquent la présence de quelques valeurs aberrantes et les scores de silhouette suggèrent que les clusters ne sont pas (ou partiellement) convexes. Par conséquent, l'utilisation d'un algorithme de clustering basé sur la densité tel que DBSCAN ou OPTICS, combiné à une étude plus complète de l'espace topologique et des relations de voisinage via UMAP pourrait nous aider à obtenir des clusters plus denses et à détecter les valeurs aberrantes. L'algorithme spectral est également une piste intéressante mais suppose une étude préalable de la normalisation optimale du laplacien de la matrice de distance pour être pleinement efficace [12].

Le troisième point relève du niveau d'analyse des activités dans l'ontologie. Pour éviter une surcharge cognitive lors de la visualisation, la section 8.2.3.2 ne présente que les résultats portant sur des activités agrégées. Une analyse détaillée au niveau

12. Plusieurs études et tests sont actuellement en cours. Notamment des expérimentations via UMAP ont été et sont en cours de réalisation au moment où nous rédigeons cette discussion.

des activités feuilles de l'ontologie serait également pertinente et pourrait aider à affiner les comportements découverts mais suppose la mise en place de méthodes de visualisation adaptées à l'affichage de nombreuses informations (e.g., l'ensemble de tous les labels de l'ontologie) tout en préservant la cognition et la facilité de lecture de l'utilisateur.

À propos des analyses effectuées, compte tenu des résultats sur l'aspect des clusters et de leur très probable non convexité, nous suggérons l'utilisation des deux indicateurs (brièvement commentés en section 4.3.1) supplémentaires suivants :

- Dans un but de robustesse et de meilleure représentativité, les séquences prototypiques centrales (mode et medoid) devraient être substituées par un modèle s'appuyant sur la notion de typicalité, c'est-à-dire faisant ressortir à la fois les caractéristiques partagées par l'endogroupe et qui les distinguent des autres clusters. L'usage de *prototypes flous* tels que définis par Lesot et al. dans [137] pourrait aider à obtenir des séquences prototypiques plus représentatives de la nature intrinsèque des clusters et du comportement qu'ils incarnent.
- L'utilisation d'éléments *contre-factuels* tels que définis équation 4.17 pourrait aider à mieux comprendre les différences entre clusters et les caractéristiques propres à chacun. Combiné à des indicateurs géométriques tels que le rayon, diamètre, etc. ces éléments pourraient aussi permettre d'appréhender de façon plus claire et contrastive la topologie de l'espace et des clusters formés.

Pour autant, sur la base de la méthodologie proposée, nous sommes en mesure d'analyser et d'extraire des comportements de mobilité précis de clusters d'un point de vue socio-cognitif et urbain. En outre, la méthodologie proposée facilite la compréhension des clusters, est utile pour l'évaluation et la détermination des paramètres optimaux (mesure et algorithme) du processus de clustering et permet ainsi d'échapper à l'empirisme.

Nous estimons cette approche est bénéfique à la fois aux experts métiers pour qui la synthèse et l'assimilation des connaissances découvertes sont simplifiées, mais aussi aux experts techniques qui ont ainsi une meilleure compréhension des processus et une validation, par les experts et l'intuition, des modèles extraits. En outre, l'ensemble des indicateurs et méthodes de visualisation présentés fournissent un cadre potentiel riche pour l'interaction entre experts techniques et métiers et une approche *human in the loop* permettant à tout moment d'améliorer le processus de découverte de connaissances grâce au savoir des experts.

Pour aller plus loin et rendre opérationnel ce souhait de collaboration entre experts métiers (dans notre cas : sociologues, psychologues géographes, etc.) et techniques, nous présentons dans la section suivante l'outil SIMBA (SematIc Mobility Behavior Analysis), une application web dédiée à la fouille et à l'analyse dynamique de tout type de séquences de mobilité sémantique

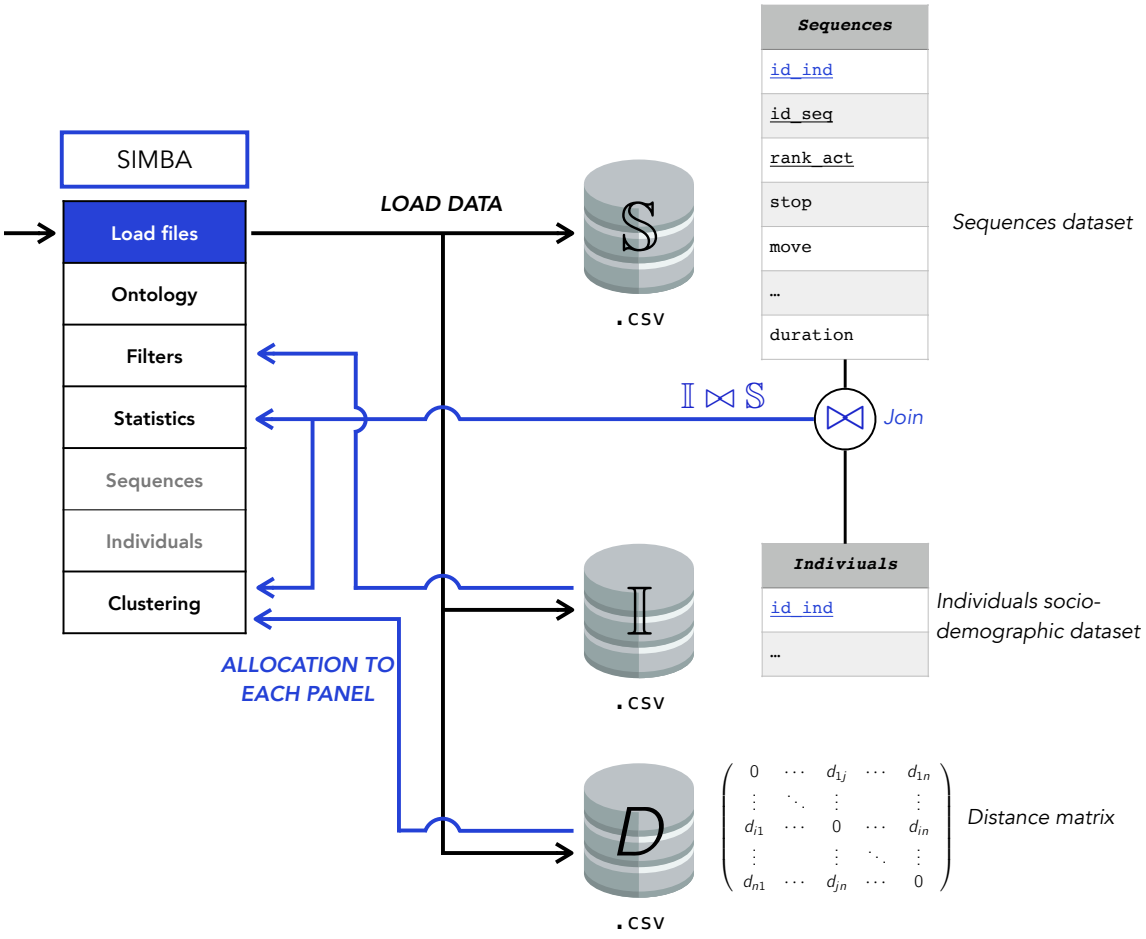


Figure 8.18 – Chargement des fichiers dans SIMBA et allocation

8.3 SIMBA : Un outil d'aide à la fouille et l'analyse visuelle de séquences de mobilité sémantique

SIMBA pour *SematIc Mobility Behavior Analysis* est une application web développée en R Shiny [13] basée sur notre méthodologie de découverte de connaissances et intégrant la plupart des indicateurs et techniques présentés dans notre cas d'étude EMD (section 8.2). La conception de SIMBA a été pensée en collaboration avec des experts métiers issus du collectif MOBI'KIDS et à fait l'objet d'un stage de développement de Master 2. Bien qu'imaginée initialement pour permettre l'analyse des données issues du projet MOBI'KIDS, SIMBA se veut être une plate-forme générique pour l'exploration interactive de tout type de séquence de mobilité sémantique. Dans un premier temps nous détaillons l'architecture de l'application Web d'un point de vue général ainsi que l'ergonomie liée à l'interface de visualisation et au chargement des fichiers. La seconde partie aborde les fonctionnalités de l'application dédiées à la fouille et l'ex-

13. https://shiny.rstudio.com/

ploration des données. Nous terminons en abordant les pistes d'amélioration futures et perspectives de développement de SIMBA à court et long terme.

L'ensemble du code et la documentation de SIMBA est disponible sur notre Github [14].

8.3.1 Architecture globale de l'application

Dans une recherche de simplicité et de minimalisme, l'application SIMBA a été conçue selon un ensemble de 5 sections principales qui permettent la navigation et l'interaction avec les données :

- Chargement des données (*Load files*)
- Chargement et visualisation des ontologies (*Ontology*)
- Filtrage des données (*Filters*)
- Statistiques générales (*Statistics*), dotée de deux sous onglets :
  - Sur les séquences (*Sequences*)
  - Sur les individus (*Indiviuals*)

14. `https://github.com/Clement-Moreau-Info/SIMBA`

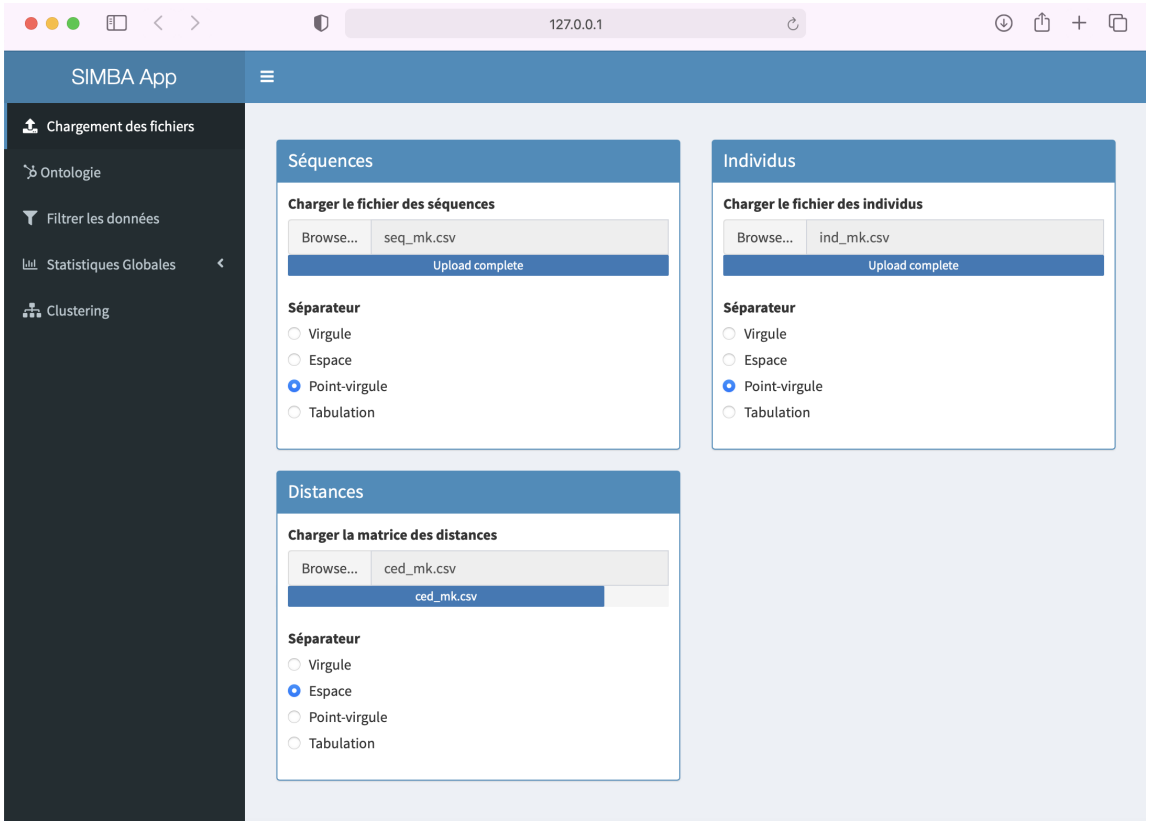


Figure 8.19 – Page de chargement des fichiers de SIMBA

id_ind	id_seq	rank_act	stop	move	duration
1	1	1	1		600
1	1	2		121	210
1	1	3	2		630
2	2	1	1		240
2	2	2		100	15
2	2	3	31		20
2	2	4		100	20
2	2	5	1		1145
1	3	1	1		1440

Table 8.6 – Exemple de données de la table `Sequences` ($\mathbb{S}$)

- Fouille et clustering interactif (*Clustering*)

La figure 8.18 détaille le processus de chargement des données depuis la page *Load files* et la figure 8.19 présente l'interface de chargement dans SIMBA. L'application requiert 3 fichiers de données (au format .csv) pour être pleinement opérationnelle :

- L'ensemble des séquences (noté $\mathbb{S}$). La table possède le schéma de base :

  $$\texttt{Sequences}(\texttt{id\_ind}, \underline{\texttt{id\_seq}}, \underline{\texttt{rank\_act}}, \texttt{stop}, \texttt{move}, ..., \texttt{duration})$$

  Le couple (`id_seq`, `rank_act`) fournit la clé primaire de la table et désigne respectivement l'identifiant de séquence et rang de l'activité dans la séquence. La colonne `id_ind` renseigne l'identifiant de l'individu ayant réalisé la séquence. Le schéma est inspiré de la modélisation des séquences sémantiques-temporelles telle que détaillée chapitre 6. Dans un souci de lisibilité, les dimensions `stop` et `move` ont été séparées [15]. Le symbole `...` indique que le schéma peut être complété par des dimensions supplémentaires optionnelles (ex. lieux, accompagnement, etc.) comme explicité en chapitre 7 pour la prise en compte d'éléments multi-dimensionnels. La colonne `duration` indique la durée totale de l'activité. La table 8.6 présente un exemple de données pour 3 séquences fictives décrites sur 24h (la colonne `duration`) est en minutes. Les numéros dans les colonnes `stop` et `move` correspondent aux identifiants de labels de l'EMD décrits table 8.2.
- L'ensemble des informations socio-démographiques sur les individus (noté $\mathbb{I}$). La table `Individual` possède la clé primaire `id_ind` permettant d'effectuer une jointure avec la table `Sequences`. Hormis cette clé, la table ne possède pas de colonne imposée.

15. Ces dimensions peut être réunies sous un unique label `activity` selon l'usage et le besoin métier.

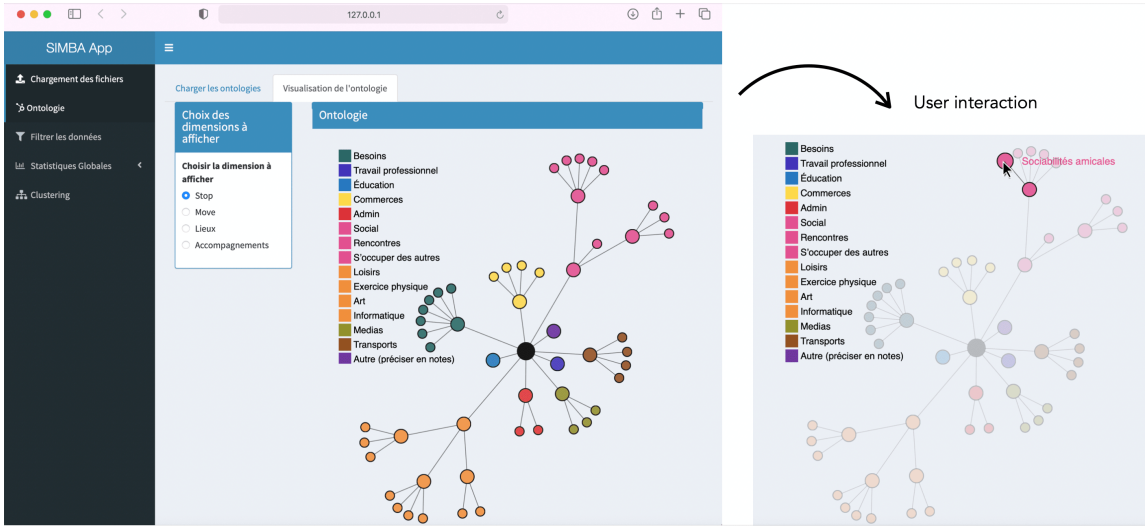


Figure 8.20 – Visualisation d'ontologie dans SIMBA

- La matrice de distance D entre chacune couple de séquences est décrite selon une représentation matricielle à 2 dimensions. Celle-ci est pré-calculée et chargée par l'utilisateur pour éviter des temps de calcul potentiellement importants.

Chacun de ces fichiers est ensuite assigné aux sections correspondantes tel qu'illustré sur la figure 8.18. La jointure pratiquée entre les table `Sequences` et `Individuals`, $\mathbb{S} \bowtie \mathbb{I}$, permet d'établir des statistiques et filtres généraux en fonction des variables socio-démographiques des individus. Nous détaillons ce point plus loin dans cette section.

Concernant le chargement et la visualisation des ontologies, ceux-ci sont réalisés dans la section *Ontology*. Une ontologie doit être assignée par dimension de la table `Sequences`. La figure 8.20 montre le rendu visuel et l'interaction avec une ontologie chargée dans SIMBA. Nous utilisons la librairie `NetworkD3` [16], surcouche de la librairie Javascript `D3.js`, pour modéliser et visualiser les graphes. L'ontologie permet de définir une notion de hiérarchie entre les activités ainsi que les couleurs attribuées lors des visualisations pour les sections *Statistics* et *Clustering*.

La troisième section de SIMBA, dédiée au filtrage des données, permet à l'utilisateur l'ajout de conditions logiques afin de sélectionner les individus et séquences obéissant à l'ensemble des critères exigés.

La figure 8.21 représente le processus de filtrage et de mise à jour des fichiers de l'application. L'ensemble des filtres applicables, noté σ, concernent l'ensemble des données contextuelles socio-démographiques $\mathbb{I}$. On obtient ainsi une vue des données, $\mathbb{I}_\sigma$, combinée ensuite par jointure à l'ensemble des séquences $\mathbb{S}$ tel que $\mathbb{I}_\sigma \bowtie \mathbb{S}$. Cet ensemble est ensuite alloué aux sections *Statistics* et *Clustering* pour le traitement

16. `https://cran.r-project.org/web/packages/networkD3/networkD3.pdf`

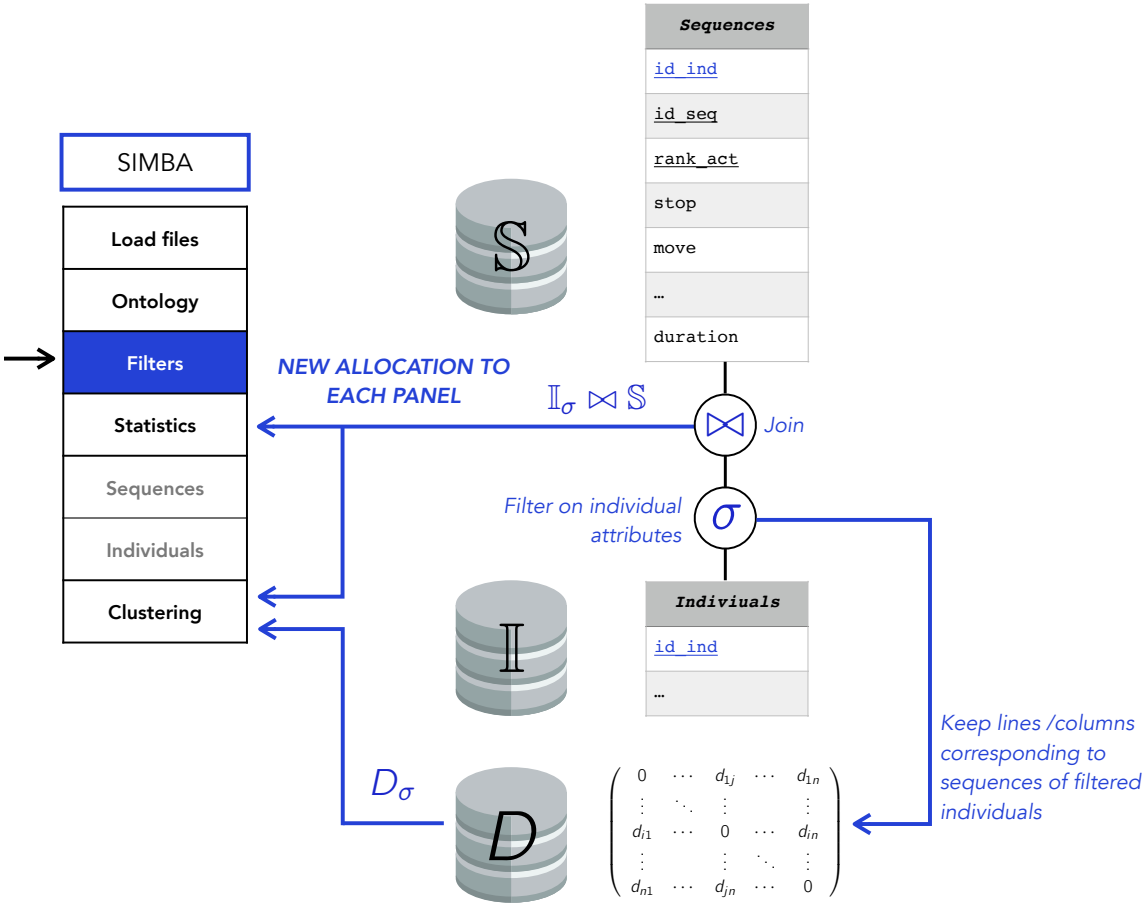


Figure 8.21 – Filtrage des données dans SIMBA

et l'analyse de données. Parallèlement, la matrice de distance D est également filtrée afin de conserver uniquement les couples de séquences issus des individus de $\mathbb{I}_\sigma$. On obtient une matrice D_σ ré-allouée à la section *Clustering*. La figure 8.22 montre l'interface de la section *Filters* de SIMBA. On voit ici plusieurs filtres appliquées sur les données MOBI'KIDS où l'utilisateur a conservé les individus *enfants* (`pers_parente = Un enfant`) et de sexe *féminin* (`pers_sexe = Femme`). La partie droite de l'interface fournit une vue sur les données filtrées.

Dans la section suivante, nous abordons les fonctionnalités de l'application dédiées à la fouille et l'exploration de données de séquences de mobilité sémantique et reprenant, en partie, les outils abordés lors du cas d'étude de l'EMD Rennes 2018 (section 8.2).

8.3.2 Fouille et analyse de séquences de mobilité sémantique dans SIMBA

La section *Statistics* est décomposée en deux sous-parties : (i) Un sous-onglet *Sequences* reprend sous un format interactif les indicateurs développés en section 8.2. On redécouvre :

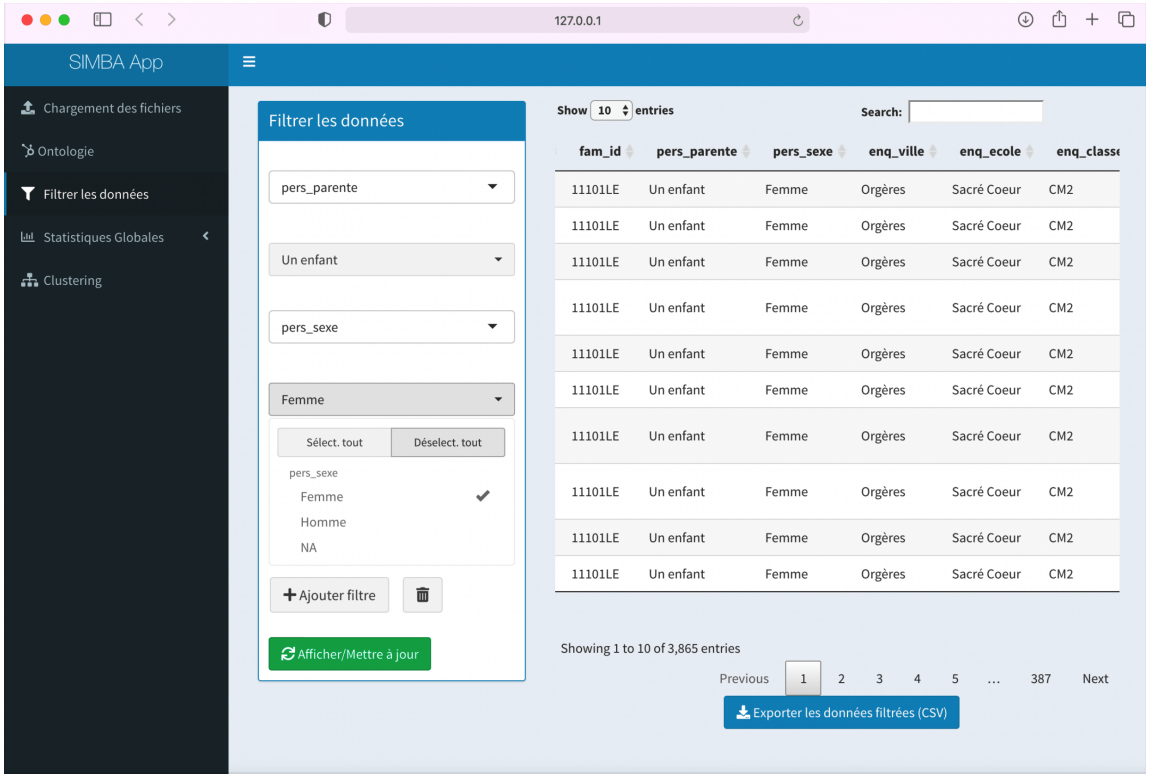


Figure 8.22 – Filtrage des données selon des critères socio-démographiques dans SIMBA

1. La *longueur des séquences* qui analyse la distribution du nombre d'activités au sein des séquences.
2. La *distribution des fréquences* des activités par dimension.
3. Les *diagrammes de flux* exhibant les transitions importantes par dimension.
4. La distribution de densité de l'*entropie & prédictibilité* des séquences.

La figure 8.23 montre la page dédiée à l'analyse fréquentielle des activités au sein des séquences. Le haut de la page fait figurer les onglets pour chacun des indicateurs présents. Les deux graphiques montrent ici la distribution Zipfienne des activités STOP issues du jeu de données MOBI'KIDS. Le choix de la dimension à analyser peut être changé en bas de la page à gauche et en accord avec les dimensions disponibles dans la table `Sequence`.

Le second sous-onglet (ii) permet d'effectuer une analyse bivariée des variables catégorielles de la table `Individuals`. Cette analyse présente la table de contingence et le diagramme mosaïque avec résidus de Pearson des variables sélectionnées. La figure 8.24 montre une analyse effectuée en croisant les variables `enq_ecole` représentant les écoles d'origine des enfants enquêtés et `pers_sexe` qui indique si l'individu est de sexe biologique Homme ou Femme. Nous remarquons ici une disparité statistiquement significative (i.e., biais des données) concernant les genres au sein des différentes

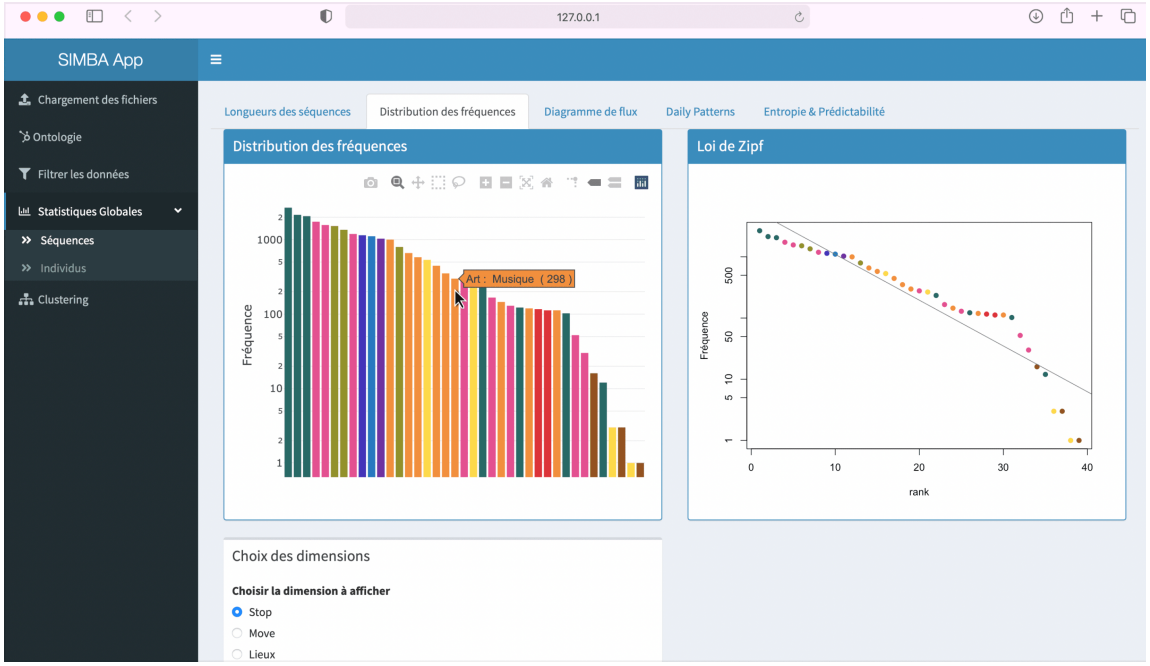


Figure 8.23 – Analyse de la fréquence des activités STOP au sein des séquences sémantiques de MOBI'KIDS à l'aide de SIMBA

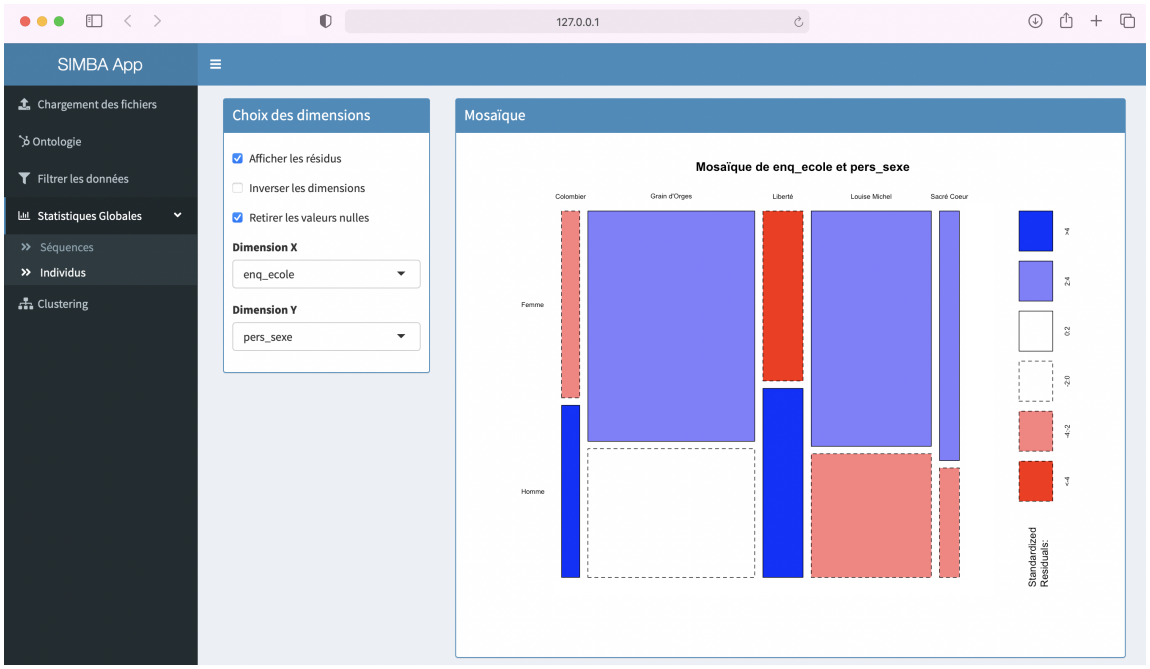


Figure 8.24 – Analyse bivariée des individus enfants enquêtés en fonction du sexe biologique et de l'établissement scolaire à l'aide SIMBA

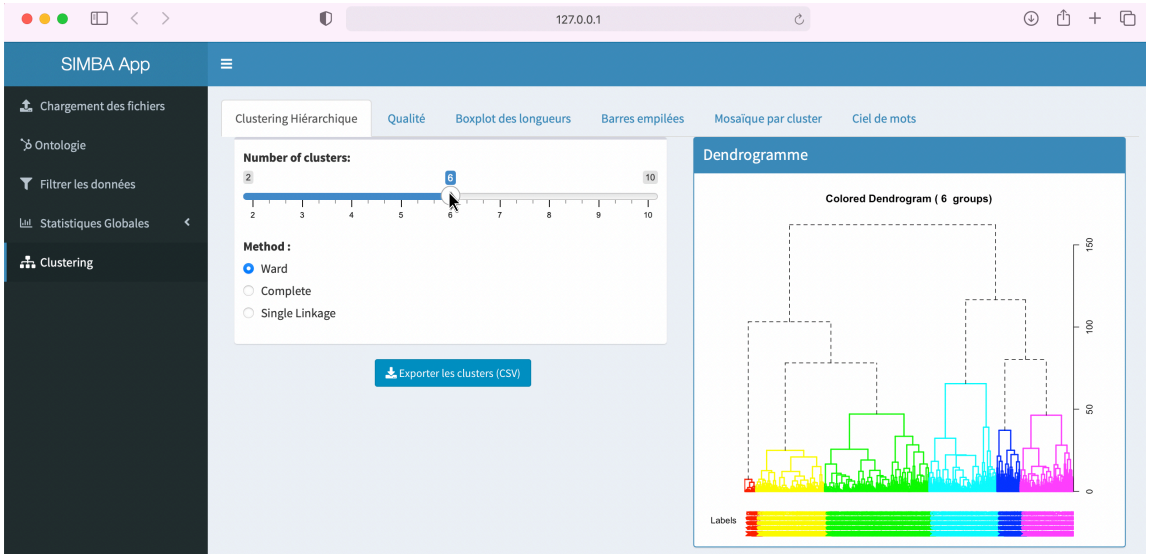


Figure 8.25 – Clustering hiérarchique dans l'application SIMA

écoles. Par exemple, il y a plus d'individus garçons enquêtés à l'école Liberté que de filles.

La dernière section concentre les outils liés au *clustering* et à la fouille interactive de comportements issus des séquences sémantiques. Pour l'heure, SIMBA propose la réalisation d'un clustering hiérarchique selon les trois critères principaux d'agrégation (Ward, Single, Complete) décrits en section 4.2. Le nombre de clusters est géré dynamiquement par l'utilisateur. Du processus de clustering résulte une table `Cluster[id_seq, id_clust]` (notée $\mathbb{C}$) qui associe à une séquence un numéro de cluster. Enfin, un bouton *export* permet le téléchargement d'un fichier .csv de la forme $(\mathbb{S} \bowtie \mathbb{I}_\sigma) \bowtie \mathbb{C}$ qui reprend l'ensemble des colonnes afin de poursuivre l'analyse de façon indépendante.

La figure 8.25 présente la page consacrée au clustering des séquences. L'utilisateur choisit le nombre de clusters, le critère d'agrégation et dispose de l'affichage en direct du dendrogramme et des clusters formés. En tant qu'algorithme dont le résultat est visualisable, nous pensons que le clustering hiérarchique est la méthode la plus appréhendable par l'utilisateur pour notre tâche de regroupement de séquences sémantiques, c'est pourquoi elle est disponible en priorité dans SIMBA. La section propose 5 autres onglets pour l'analyse des clusters :

1. L'analyse de la *qualité* des clusters.
2. Les *boîtes à moustaches des longueurs* des séquences en nombres d'activités par cluster.
3. L'analyse des activités par cluster et par dimension à l'aide de diagramme à *barres empilées*.

Figure 8.26 – Indicateurs de qualité de clustering dans SIMBA

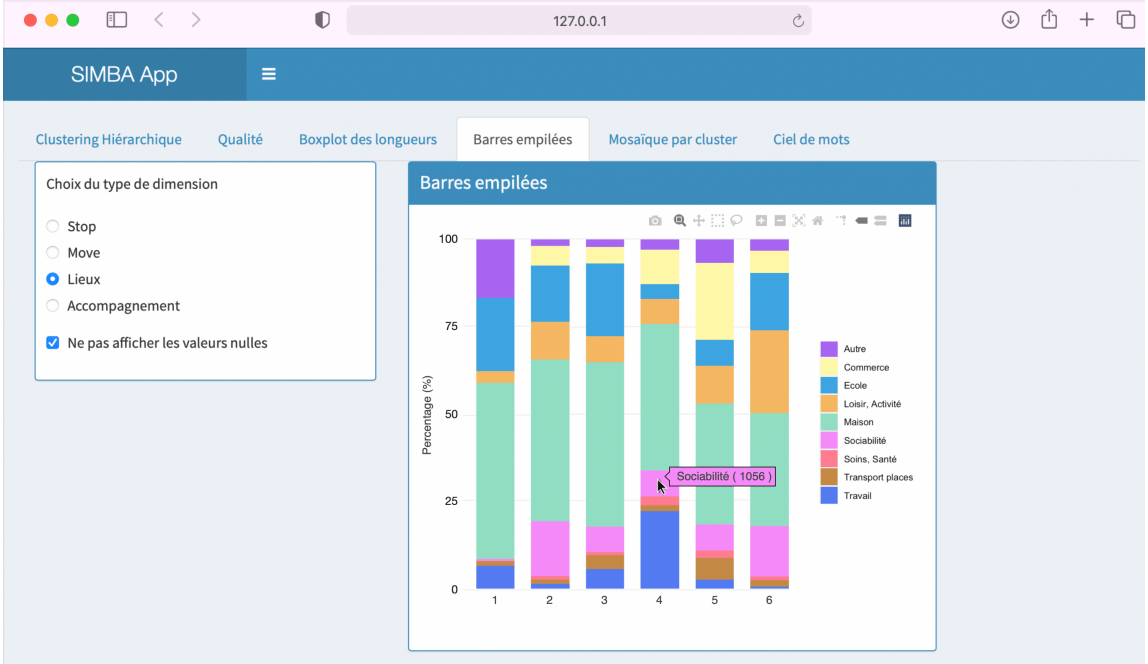


Figure 8.27 – Diagramme de barres empilées issu du clustering dans SIMBA

4. L'analyse bivariée par *diagramme mosaïque* entre les clusters et une variable qualitative sélectionnée par l'utilisateur.
5. Un résumé des clusters par *ciel de mots*. Nous revenons sur cet indicateur en perspective, section 8.3.3.

L'analyse des clusters et des indicateurs suit le même schéma que celui établi en section 8.2.3.2 lors du cas d'étude. La figure 8.26 présente par exemple l'onglet dédié

à l'analyse de la qualité des clusters formés pour le nombre fixé par l'utilisateur. On retrouve l'ensemble des indicateurs présentés dans la table 8.3 (silhouette, diamètre, rayon, etc.) à gauche de la page et les graphiques de silhouette moyen et de saut d'inertie à droite. La figure 8.27 montre la visualisation par diagramme à barres empilées, selon la dimension lieux, des séquences de mobilité du jeu MOBI'KIDS pour chaque cluster extrait.

Ces indicateurs visuels, par leur complémentarité et leur niveau de détail visent à satisfaire la compréhension de l'utilisateur sur les différents aspects des clusters de séquences sémantiques. Pour l'heure, ces indicateurs forment une aide à traduction des clusters en comportements intelligibles par des experts métiers ; néanmoins, comme nous l'avons vu en section 8.2, SIMBA n'implémente qu'une partie des indicateurs et solutions proposées. Nous évoquons en section suivante les pistes de développement pour faire de notre application une plate-forme complète pour la découverte automatique de comportements dans les séquences de mobilité sémantique.

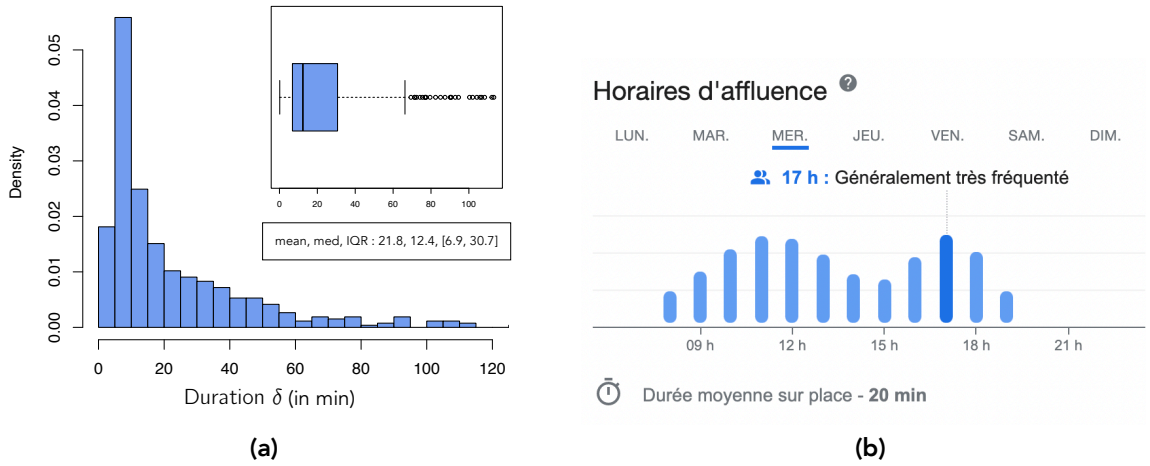
8.3.3 Perspectives de développement de SIMBA

Encore à l'aube de son développement, SIMBA forme, pour l'heure, un prototype pour la découverte et l'explication de comportements au sein de séquences de mobilité sémantique. L'application reprend en partie les indicateurs référencés table 8.1.2 et abordés section 8.2. Toutefois, son ambition est de devenir un environnement complet pour l'aide et l'analyse interactive de séquences sémantiques génériques. En conséquence, nous proposons les pistes d'améliorations technologiques suivantes :

Chargement de la matrice de distance Actuellement, il est demandé à l'utilisateur de pré-calculer et fournir la matrice de distance correspondant au jeu de données chargé. Ceci est justifié dans notre cas par les temps de calcul importants liés à l'exécution de CED.

Nous pensons toutefois laisser cette fonctionnalité tout en proposant à l'utilisateur un calcul dynamique d'une matrice de distance correspondant au jeu de données qu'il soumet. Pour se faire, nous comptons réutiliser la bibliothèque `TraMineR` (abordée en section 3.3.2) qui fournit un ensemble important de mesures pour la comparaison de séquences de données qualitatives.

Optimisations de certains indicateurs Pour l'heure, bien que fonctionnel, le calcul des indicateurs *daily patterns* et d'*entropie & prédictibilité* est subordonné à des scripts Python annexes du fait de l'indisponibilité de certaines librairies sous R. L'appel à distance de ces scripts via le package `Reticulate` ralentit considérablement leur exécution ($\approx$ 40s pour l'exécution des daily patterns sur le jeu EMD de 10 005 séquences). Bien que ces temps restent acceptables, il serait souhaitable pour des raisons de maintenance, de simplicité, de structure du code et d'efficacité de repasser l'intégralité des fonctionnalités de SIMBA sous R.



(a) (b)

Figure 8.28 – Exemples de perspective pour la prise en compte de la dimension temporelle dans SIMBA (a) Distribution des durées δ pour l'activité "Faire des courses" dans le jeu de données MOBI'KIDS (b) Exemple de fil d'horaires d'affluence d'un centre commercial dans Google

Intégration du temps La perspective temporelle a peu été prise en compte dans SIMBA. Toutefois, son intégration sous la forme de durée pourrait être effectuée sans peine ni modification majeure de l'application. Une première piste possible concerne l'analyse de la fréquence des activités ; ce graphique peut être détourné pour non plus représenter la fréquence des activités, mais le temps alloué à chaque activité. Si l'on reprend la modélisation des séquences sémantiques-temporelles, chaque barre représente ici, pour tout $x \in \Sigma$, la quantité $count(x) = |\{x_{ik} | S_i \in \mathbb{S}, x_{i_k} = x\}|$, ce qui revient à compter [17] le nombre de symboles identiques à x dans toute la base de séquences sémantiques $\mathbb{S}$. Dans un cadre basé sur la durée, on aurait alors $count_\delta(x) = \sum_{S_i \in \mathbb{S}} \sum_{(x_{ik}, \delta_{ik}), x_{ik} = x} \delta_{ik}$, on somme les durées des activités [18].

Il serait également pertinent de modéliser la distribution de probabilité liée à la durée des activités. Par exemple, on peut imaginer vouloir connaître la distribution liée à la durée de l'activité "Faire des courses", voire aux méta-activités (e.g. "Achat") grâce à une navigation dynamique dans l'ontologie afin de régler le niveau de granularité sémantique.

Enfin, un dernier indicateur statistique et visuel lié au temps serait la distribution de probabilité des horaires d'affluence par activité sur l'axe $[0, T_{\max}[$ (supposé $T_{\max} = 24h$). Ce graphique permettrait d'obtenir un référentiel quant à la temporalité standard en fonction des activités conduites. Les graphiques (a) et (b) présentés figure 8.28 résument respectivement les deux dernières propositions.

17. Ou en SQL `SELECT COUNT(A) FROM Sequences WHERE A = x;` avec $A \in \{\texttt{stop}, \texttt{move}, \ldots\}$.
18. En SQL `SELECT SUM(Durations) FROM Sequences WHERE A = x;` avec $A \in \{\texttt{stop}, \texttt{move}, \ldots\}$.

Umap et autres méthodes de clustering Compte tenu des résultats prometteurs d'UMAP, nous pensons intégrer la possibilité à l'utilisateur de pratiquer une réduction dimensionnelle via cette technique afin de visualiser les données en 2-dimensions. Un corollaire de cet ajout serait celui des différents algorithmes testés dans le chapitre 7.

Ajouts d'indicateurs pour l'analyse des clusters Certains indicateurs présents dans la phase d'analyse des clusters de la section 8.2 n'ont pas été intégrés à SIMBA, principalement du à des problématiques de visualisation et d'interface (notamment les différents diagrammes de flux par cluster et la carte de chaleur des daily patterns). Des solutions de visualisation adaptées doivent être trouvées.

De plus, nous comptons ajouter les indicateurs liés à l'étude de la dimension temporelle abordés ci-dessus ainsi qu'une visualisation par tapis de séquences (voir figure 6.9) afin d'analyser les notions de temporalité et de simultanéité des activités au sein des clusters.

Visualisation et calcul des séquences prototypiques Dans le cas d'étude de l'EMD, nous avons proposé la mise en place d'emojis afin de couvrir, par le biais d'une image, un ensemble de concepts complexes ce qui permet d'alléger la représentation des séquences.

Cependant, la mise en place de séquences prototypiques dans le cas de séquences d'éléments multi-dimensionnels (comme pour le jeu MOBI'KIDS) est difficile, tant pour des raisons computationnelles que de visualisation et de surcharge cognitive. Tout d'abord, une telle visualisation supposerait que l'utilisateur spécifie un emoji ou image exprimant les concepts supérieurs de l'ontologie ce qui est un investissement certain de la part de l'utilisateur lorsque l'alphabet et l'ontologie des activités sont de taille importante.

Le second problème est lié l'interprétabilité et au calcul des éléments prototypiques dans le cas de séquences d'éléments multi-dimensionnels. Lorsque les dimensions deviennent nombreuses, la représentation des séquences devient rapidement absconse, même à l'aide d'emojis, mais le calcul devient lui aussi peu pertinent. En outre, un phénomène analogue à la *malédiction de la dimensionnalité* [146] fait qu'en haute dimension, les éléments ont tendance à être très éloignés les uns des autres [19] dans l'espace topologique ce qui conduit la plupart des éléments prototypiques (en tout cas mode et medoid) à être de mauvais représentants pour l'ensemble des séquences d'éléments multi-dimensionnels d'un cluster.

De façon générale, la visualisation d'indicateurs et de statistiques permettant de décrire des séquences d'ensembles d'éléments sémantiques multidimensionnels est

19. Nous conseillons la vidéo suivante de Stéphane Mallat au collège de France sur le comportement des données en haute dimensionnalité `https://www.youtube.com/watch?v=WGIJOwvYmfE&ab_channel=Coll%C3%A8gedeFrance`

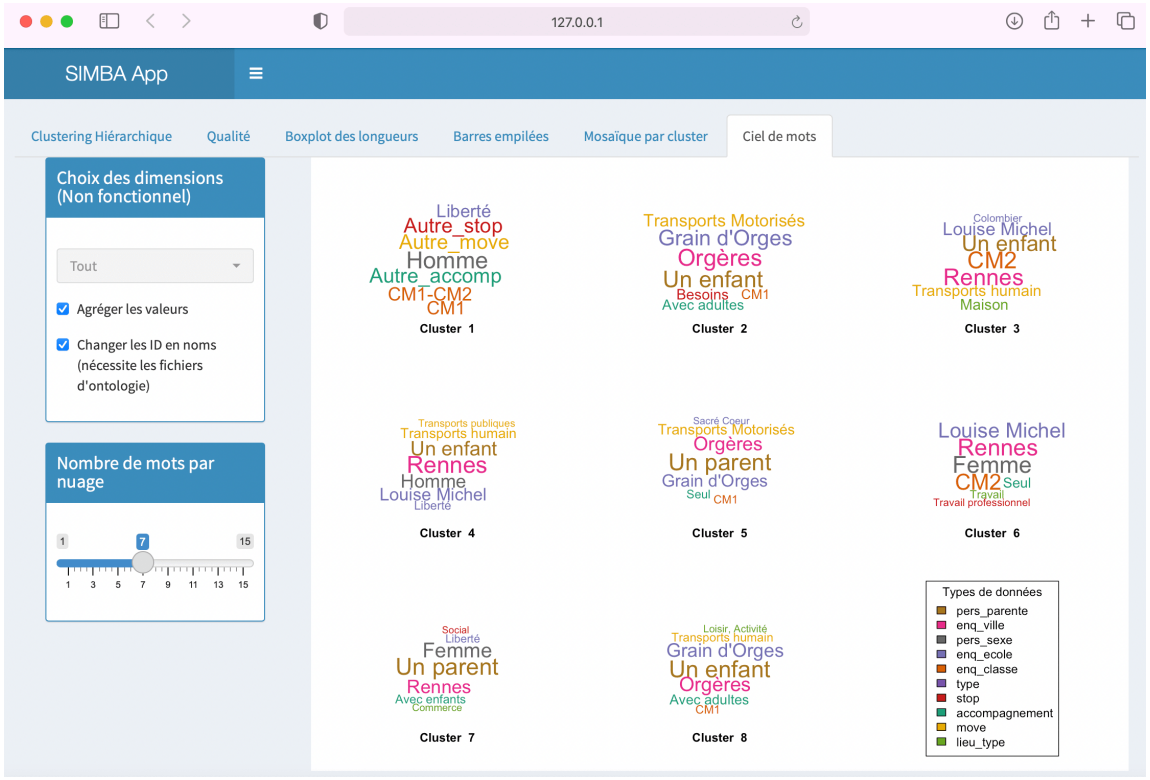


Figure 8.29 – Prototype de résumé par ciel de mots

un challenge difficile faisant intervenir de lourdes notions en statistiques, analyse de données et IHM.

Résumé automatique de comportements Enfin, concernant la phase de résumé automatique des clusters en comportements intelligibles, nous avons vu en section 8.2.3.3 que les méthodes utilisées étaient semi-automatiques en ce sens qu'elles sont supportées par les indicateurs visuels utilisés lors de la phase d'analyse et doivent induire une part d'interprétation et validation issue des experts humains.

Conscients que le degré de familiarité avec les concepts statistiques et les graphiques utilisés est variable parmi les experts et utilisateurs potentiels de SIMBA, nous proposons la mise au point de résumés linguistiques permettant de guider l'interprétation et faciliter la lecture de certains graphiques complexes (mosaïque, flux, etc.). De façon plus générale, nous songeons à la mise en place d'indicateurs de premier niveau, c'est-à-dire alliant simplicité de lecture, concision mais conservant une forme d'exhaustivité qui permettrait de dresser un profilage général des clusters. Des premières pistes autour des nuages de mots où les termes saillants sont extraits selon leur significativité statistique et leur fréquence d'apparition sont encore à l'étude. Ainsi, un nuage représente un cluster et ceux-ci sont agencés de façon à constituer un *ciel de mots* de façon à fournir une vision panoramique des comportements découverts depuis les clusters. Un stage de Master 2 a permis l'élaboration d'un premier prototype visible

sur la figure 8.29. Des tests utilisateurs et études approfondies d'un point de vue cognitif et du ressenti humain sont en cours afin d'adapter les modes de visualisation et contrôler la pertinence des informations rendues à l'utilisateur et trouver un juste équilibre entre qualité de restitution, ergonomie et esthétisme.

Discussion

Au cours de ce chapitre, nous avons présenté une méthodologie complète pour l'extraction de comportements et le profilage de données depuis un ensemble de séquences de mobilité sémantique. Cette méthodologie s'appuie sur la structuration en ontologie des activités dont sont composées les séquences ainsi que d'un ensemble d'indicateurs en charge de décrire la structure des données et de mettre en lumière les différentes singularités parmi les clusters.

La méthodologie a été appliquée sur un ensemble de données réelles de 10 005 séquences de mobilité issues de l'EMD 2018 où, combinée à la mesure CED et un clustering hiérarchique, elle a permis la découverte et la compréhension de schémas de comportements de mobilité riches et intelligibles. Pour autant, la méthodologie étant générique et adaptative, nous prévoyons d'appliquer une étude similaire intégrant de nouveaux indicateurs et une mesure pour la comparaison de séquences tenant compte des aspects temporels telle que FTH. Par ailleurs, mentionnons ici que la méthodologie a été également mise à l'épreuve pour la découverte et l'interprétation de comportements issus de l'exploration de bases de données [166] où elle a offert une meilleure compréhension des différents profils utilisateurs et niveaux d'expertise pour le requêtage des bases de données. De telles connaissances sont précieuses pour la mise au point d'outils d'aide à la navigation dans les données, la fouille d'intentions, la détection de comportements frauduleux ou encore pour mieux comprendre les mécanismes liés à la psychologie humaine quant à l'intérêt utilisateur au cours d'une tâche d'exploration de données.

Dans le but de simplifier la transmission des connaissances issues du processus d'apprentissage et de découverte aux experts humains et d'accroître la collaboration humain-machine, nous avons concrétisé notre processus méthodologique dans un prototype d'application web, SIMBA, qui forme une preuve de concept pour le développement d'un outil plus complet dédié à l'exploration, la fouille et l'analyse interactive de séquences sémantiques complexes. SIMBA incorpore bon nombre d'indicateurs éprouvés lors du cas d'étude des EMD et a été mis à l'épreuve dans l'analyse et la découverte de comportements portant sur le jeu de données de séquences d'ensembles d'éléments sémantiques multidimensionnels complexes du projet MOBI'KIDS. Des résultats au potentiel prometteur doivent encore être analysés en collaboration conjointe des experts SHS (sociologues, psychologues, géographes) afin de corroborer les différentes hypothèses. Nonobstant, SIMBA a d'ores et déjà prouvé son utilité pour l'aide à l'exploration de séquences de mobilité sémantique auprès d'experts métiers. De nouvelles perspectives de développement avec l'incorporation d'indicateurs pour

la qualification temporelle des séquences sont prévues ainsi que de nouveaux tests utilisateurs dans le but de développer des scénarios d'analyse plus complexes et à dessein comparatif (e.g., comparer les comportements garçon vs fille / primaire vs collège).

Une perspective stimulante serait d'appliquer une étude similaire aux logs d'exploration de base de données de [166] aux logs récoltés suite à ces futurs tests utilisateurs de SIMBA. En outre, une telle étude nous permettrait de comprendre et mieux cerner les différents besoins et attentes au sein de notre outil et formerait un cadre unificateur à l'ensemble de nos travaux.

Enfin, nous sommes convaincus que la clé pour une meilleure découverte de connaissances, tant par les aspects qualité des résultats, d'interprétation, ou encore de créativité et de recherche réside dans la collaboration entre disciplines. Par ces contributions nous espérons oeuvrer, à notre échelle, au rapprochement entre experts techniques et métiers, humain-machine et sciences humaines, cognitives et formelles dans un but commun d'amélioration du savoir.

Chapitre 9
Conclusions et perspectives

Nous concluons cette thèse en récapitulant les différents aspects abordés, problématiques soulevées, contributions et perspectives possibles suite à nos travaux.

Pour rappel, cette thèse s'inscrit dans le domaine de la science des données et de l'IA. Elle se positionne à l'interface de deux projets, SMARTLOIRE et MOBI'KIDS, qui étudient respectivement la mobilité touristique en région Centre-Val de Loire, et la mobilité quotidienne d'enfants et de leurs parents dans un contexte urbain / péri-urbain. L'objectif principal de ces projets porte sur l'extraction de comportements intelligibles depuis un ensemble de séquences de mobilité sémantique. En outre, nous avons évolué en collaboration avec de nombreux acteurs métiers (décideurs politiques, sociologues, psychologue, urbanistes, géographes), avec des données riches, complexes, au potentiel sensible (par leurs aspects de vie privée) et dans un cadre d'apprentissage non supervisé. De fait, et suite à l'état de l'art établi, nous avons identifié deux manques majeurs : (i) une mesure permettant la comparaison de séquences sémantiques tenant compte des spécificités liées aux comportements humains et (ii) une méthodologie comblant les manques analytiques et éthiques, liés à la transparence et l'interprétation contextuelle des données, pour l'extraction de comportements depuis un ensemble de séquences sémantiques.

Pour lever le premier verrou scientifique liée à l'élaboration d'une mesure pour la comparaison de séquences de mobilité sémantique, nous nous sommes appuyés sur un corpus d'études de la mobilité et de la psychologie socio-comportementale. Cette revue de l'état de l'art nous a permis la formalisation de spécificités liées au comportement humain dans les séquences sémantiques correspondant à des exigences liées au temps (résistance aux décalages, permutations) et à la sémantique (homogénéité, répétitions d'activités). En première approche, nous avons répondu à l'ensemble de ces spécificités en alliant la distance d'édition à la logique floue et créé la mesure nommée Contextual Edit Distance (CED). Ayant prouvé sa généricité et son utilité lors de tâches d'extraction portant sur des séquences d'actions humaines telles que la mobilité ou l'exploration de bases de données, CED a été présentée dans différentes publications internationales [162, 164, 165, 166, 167].

Toutefois, CED demeure limité par sa complexité importante en temps de calcul et du fait qu'elle ne tienne pas compte la dimension temporelle autrement que par la notion

de précédence. En conséquence, nous avons ré-appliqué les concepts flous développés lors de la création de CED dans un cadre où le temps et les durées des activités sont modélisés de façon continue. Afin de faire chuter la complexité de calcul, nous nous sommes ré-appropriés la distance de Hamming dont la complexité est originellement linéaire mais dont la contrainte principale est d'être applicable seulement sur des séquences d'une même taille fixe. Ainsi, la mesure que nous avons proposée pour la comparaison de séquences sémantiques-temporelles, Fuzzy Temporal Hamming (FTH), tient compte de l'ensemble des spécificités soulevées avec une amélioration significative de la complexité temporelle comparée aux mesures existantes de l'état de l'art liées aux techniques d'Optimal Matchning. FTH a donné lieu à un article lors de la conférence Fuzz-IEEE 2021 [163] et a reçu le prix de Best Student Paper.

Enfin, nous avons proposé une approche pour le clustering de séquences d'ensembles d'éléments multidimensionnels permettant de tenir compte à la fois d'une potentielle richesse de description des entités constituant les séquences mais aussi des aspects liés à la simultanéité d'activités. En outre, notre approche se compose d'une mesure permettant la comparaison d'ensembles d'éléments multidimensionnels et utilise la technique de réduction de dimensionnalité UMAP afin de représenter les séquences complexes dans un espace 2D euclidien. Testée sur un cas d'étude liée à SMARTLOIRE faisant intervenir des entités touristiques issues de DATAtourisme et des séquences artificielles générées à partir d'un marcheur aléatoire markovien, notre approche s'est révélée très efficace pour regrouper des séquences adoptant des comportements similaires et a fait l'objet d'une publication lors de la conférence ACM SAC 2021 [162].

Le second verrou scientifique consistait en un apport méthodologique à la démarche de fouille et d'analyse de données. Notre objectif s'inscrit dans le cadre de l'XAI dans une approche explicable des processus d'extraction de connaissances. Suite aux constatations dressées par l'état de l'art mettant en lumière les manques liés à l'analyse contextuelle et compréhension préliminaire des données, nous avons proposé une méthodologie orientée sur l'analyse de données *pre* et *post-process*. Ces analyses sont développées pour l'étude de séquences sémantiques où un ensemble d'indicateurs statistiques et visuels sont proposés. Ces indicateurs sont en charge de la description et bonne interprétation des données et des clusters extraits et viennent ainsi expliquer les différents aspects des données. Cette phase d'explicabilité s'accomplit dans une perspective *human in the loop* où les experts métiers sont en charge à la fois de la contextualisation des résultats fournis et de leur pertinence. Ainsi, le processus de fouille est guidé et sa validation subordonnée à l'humain. Une telle approche permet d'améliorer à la fois la qualité des connaissances extraites et la compréhension des experts métiers. De fait, elle améliore la rapidité des échanges entre experts et pourrait se montrer utile pour l'évolution vers un cadre d'apprentissage semi-supervisé. Notre approche méthodologique a fait l'objet de plusieurs publications. Initialement développée dans l'article [165], elle a été mise en oeuvre pour l'extraction de comportements explicables d'exploration de base de données [166] lors d'un article à la conférence DOLAP @ EDBT/ICDT. Cet article a fait l'objet d'une extension pour le

Jeu de données	Volumétrie (nb. seq.)	Ontologie	Mesure	Algo. clustering	Axe contrib.	Article
-	-	-	CED	Hierarchical	T	[164]
EMD	10 005	Hetus	CED	Hierarchical	M, A	[165]
SMARTLOIRE	250	DATAtourisme	CED-multi	UMAP + k-means / Spectral	T, A	[162]
EMD	1 200	Hetus	FTH	Hierarchical	T	[163]
MOBI'KIDS	4 271	Hetus	CED-multi	Hierarchical	M	-
Artificial*	50	-	CED	Hierarchical	A	[167]
SQL share	2 809	-	CED	UMAP + DBSCAN	M, A	[166]

* D'autres jeux ont été éprouvés dans [167]

Table 9.1 – Table des contributions. Les lettres T, M et A font référence respectivement aux axes de contributions Théoriques, Méthodologiques et Applicatifs

journal Information Systems. Dans une visée plus technologique, notre méthodologie a été implémentée en une application web, SIMBA, permettant l'analyse et la fouille dynamique et interactive de séquences de mobilité sémantique. Notamment, SIMBA est en cours de déploiement auprès d'experts métiers (sociologues, psychologues, géographes) pour l'analyse des données issues du projet MOBI'KIDS. L'application a reçu un accueil très favorable, démontrant du même coup son efficacité et utilité pour la découverte de connaissances intelligibles et la collaboration homme-machine.

La table 9.1 reprend les différentes contributions de la thèse et détaille le jeu de données utilisé et sa volumétrie en nombre de séquences, l'ontologie associée, la mesure employée, l'algorithme de clustering retenu et l'axe de contribution (Théorique (T), Méthodologique (M) ou Applicatif (A)). Les mesures et analyses réalisées ont donc été validées sur un plusieurs jeux de données démontrant la généricité et l'intérêt de nos approches.

Perspectives et travaux futurs

Grâce aux nombreuses réalisations produites durant la thèse, de multiples perspectives et nouvelles applications peuvent être soulevées.

Au sujet de CED, nous avons relevé la problématique concernant le paramétrage de la fonction d'encodage du vecteur temporel. Une étude des propriétés formelles selon un ensemble de fonctions pré-définies pourrait être une piste souhaitable. Pour une visée encore plus fondamentale, la récupération de l'axiome de l'inégalité triangulaire dans CED (et FTH) nous apparaîtrait comme une avancée à la fois d'un point de vue théorique et pratique. Dans une perspective plus pragmatique et due à sa complexité de calcul importante, des améliorations en termes d'optimisation sont possibles pour

rendre CED opérationnelle sur des volumétries de données importantes. Ces réalisions sont de l'ordre du moyen terme, voire long terme.

Concernant FTH, la mesure étant une des contributions les plus récentes de la thèse, celle-ci n'a été testée que sur un ensemble restreint de données. Une perspective à court terme est donc la validation de FTH sur de nouveaux jeux de données, en particulier sur l'ensemble des données de l'EMD 2018 et de FTH-multi sur MOBI'KIDS. Pour ce faire, de nouveaux indicateurs descriptifs du temps et de la notion de durée doivent être incorporés à notre méthodologie d'analyse. Ces différents indicateurs permettront également de comprendre avec plus de précision le fonctionnement des différentes variantes de la fonction de coût – γ et Δ. Plus généralement, une comparaison plus détaillée des différentes mesures produites avec celles de l'état de l'art doit être réalisée. Notamment, une comparaison entre mesures spatiales et sémantiques nous semble pertinente afin d'étudier la complémentarité de ces deux dimensions, voire une possible corrélation entre les données spatiales et sémantiques à partir des clusters obtenus selon nos mesures. Un premier essai issu du jeu MOBI'KIDS a été réalisé par Laurent Etienne dont les figures et détails d'analyse sont présentés en Annexe A, figures A.1 – A.8. Pour l'heure, une étude visuelle préliminaire n'a montré, *a priori*, aucun lien immédiat et trivial entre ces deux dimensions et met en évidence la difficulté de la tâche ainsi que la complémentarité de ces deux types d'approches.

Ainsi, l'incorporation de la dimension spatiale est un enjeu futur stimulant. Une piste pour une telle réalisation pourrait être menée en incorporant une dimension sémantico-spatiale au sein de la modélisation des séquences d'ensembles d'éléments sémantiques multidimensionnels (voir chapitre 7). La composante spatiale serait abstraite en un label sémantique, par exemple "Blois" suffirait pour exprimer le fait d'être à l'instant considéré dans la ville de Blois. Cette perspective nous semble intéressante car elle permettrait d'embrasser du même coup les dimensions spatiale et sémantique en utilisant une ontologie/taxonomie administrative des zones géographique (e.g., Blois est dans le Loir-et-Cher). D'autres approches du type centre-ville vs banlieue ou proche d'une station de transport en commun vs loin restent possible grâce à ce type d'approche. De plus, cette vision permettrait de concilier approche spatiale et respect de la vie privée de l'utilisateur. D'autres pistes sont également possibles afin d'associer la dimension spatiale à notre approche comme par exemple le couplage à d'autres mesures spatiales selon une approche similaire à Furtado et al. [79]. Une telle perspective pourrait être menée à moyen ou long terme.

Enfin, à propos de la méthodologie d'analyse et de découverte de comportements, de nombreuses réflexions et pistes d'amélioration peuvent être suggérées. Pour commencer, nous proposons d'incorporer les suggestions du psychologue T. Miller [160] (voir section 4.3.2) pour mieux prendre en compte le ressenti utilisateur et améliorer les aspects de communication de l'information pour le résumé des comportements au sein des clusters. Nous avons déjà proposé en chapitre 8 l'ajout de nouveaux indicateurs d'éléments prototypiques plus représentatifs et robustes. De même, l'ajout d'éléments contre-factuels permettrait d'éclairer selon une vision constrastive les comportements

	CED	FTH	Methodo.	Spatial
Court		Application au jeu MOBI'KIDS	Ajout d'indicateurs temporels (durée) ; Résumés linguistiques flous	
Moyen	Étude du vecteur temporel ; Optimisations	Étude des variantes γ et Δ	Indicateurs de résumé de cluster (nuage de mots, prototype)	Ajout de labels sémantico-spatiaux (quartier, ville, etc.) ; Spatialisation des clusters
Long	Retrouver l'inégalité triangulaire		Adaptation et amélioration de SIMBA ; Visualisation et résumé en haute dimensionnalité ; Application pour l'aide à la prédiction / recommandation	Étude des corrélations entre dimensions spatiale et sémantique

Table 9.2 – Table des perspectives à court, moyen et long terme

extraits. Toutefois, comme nous l'avons soulevé en section 8.3.3 la mise au point de ces indicateurs peut se révéler être un challenge dans le cadre de séquences sémantiques complexes et notamment dans le cas d'ensemble d'éléments sémantiques multidimensionnels. D'autres indicateurs de résumé tels que les nuages et ciels de mots sont encourageants mais réclament une conception et étude ergonomique approfondie du point de vue de la visualisation d'informations afin d'être pleinement efficace. Nous pensons que ces perspectives pourraient être réalisées à moyen et long terme tout en donnant lieu à de nouvelles pistes de recherche originales et prometteuses. Elles constituent un défi à la fois d'un point de vue des sciences de la donnée mais aussi cognitives. Dans une vision à court et moyen terme, et afin d'automatiser l'extraction des comportements, l'ajout de résumés linguistiques flous permettrait l'assimilation des connaissances extraites depuis nos processus de clustering par un plus large public, sans pré-requis scientifique sur la lecture des graphiques liés aux indicateurs. Dans cette lignée, une perspective à long terme serait l'adaptation de SIMBA au traitement de tous types de séquences sémantiques avec l'espoir d'en faire un outil complet pour l'aide à l'analyse et la découverte de comportements.

D'autres pistes peuvent également être mentionnées comme la ré-utilisabilité de nos

approches à d'autres domaines et contextes métiers. Notamment pour la détection de comportements aberrants ou des aspects liés à la prédiction et la recommandation d'items tels que des musiques, produits ou l'aide à la navigation dans les données. La table 9.2 reprend et résume nos différentes propositions de perspectives de recherche selon la visée à terme et la contribution.

Pour finir, nous aimerions souligner les nombreuses questions et problématiques éthiques liées à nos applications et notre démarche. Bien qu'ayant été une de nos pré-occupations majeures, à quel point notre approche est-elle respectueuse de l'individu et de sa singularité ? Quelles doivent être les limites de notre approche ? Quels sont ses biais ? Peut-on et doit-on essayer d'abolir le hasard, la contingence et la surprise dans le comportement d'un individu ? La recherche d'explications, parfois, *ad-hoc*, ne poussent-elles pas à des biais de sur-interprétation et sur-apprentissage ? Etc. Ces interrogations, d'ordre philosophique et psychologique, dépassent notre champ de compétences. Aussi, suivons la maxime formulée par Wittgenstein en conclusion du Tractatus logico-philosophicus : "*Sur ce dont on ne peut parler, il faut garder le silence.*"

Bibliographie

[1] A. Abbott : *Time matters : On theory and method*. University of Chicago Press, 2001.

[2] A. Adadi et M. Berrada : Peeking inside the black-box : a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[3] P. Agarwal : Ontological considerations in giscience. *International Journal of Geographical Information Science*, 19(5):501–536, 2005.

[4] S. Aghabozorgi, A. S. Shirkhorshidi et T. Y. Wah : Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[5] R. Agrawal, C. Faloutsos et A. Swami : Efficient similarity search in sequence databases. *In International conference on foundations of data organization and algorithms*, p. 69–84. Springer, 1993.

[6] A. K. Ahmad, A. Jafar et K. Aljoumaa : Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):1–24, 2019.

[7] X. Aimé : *Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'Ingénierie des Ontologies.* Thèse de doctorat, Université de Nantes, 2011.

[8] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann et A. Baronchelli : Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491, 2018.

[9] J. Aligon, M. Golfarelli, P. Marcel, S. Rizzi et E. Turricchia : Similarity measures for olap sessions. *Knowledge and information systems*, 39(2):463–489, 2014.

[10] J. F. Allen : Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[11] L. Alvares, V. Bogorny, B. Kuijpers, J. de Macedo, B. Moelans et A. Vaisman : A model for enriching trajectories with semantic geographical information. *Proc. of the 15th annual ACM international symposium on Advances GIS*, (22):1–8, 2007.

[12] M. Ankerst, M. M. Breunig, H.-P. Kriegel et J. Sander : Optics : ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[13] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.* : Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[14] A. Artelt et B. Hammer : On the computation of counterfactual explanations–a survey. *arXiv preprint arXiv :1911.07749*, 2019.

[15] AUDIAR Rennes : Enquête ménages-déplacements en ille-et-vilaine 2018. Rap. tech., 2019.

[16] A.-L. Barabási : The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[17] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini et M. Tomasini : Human mobility : Models and applications. *Physics Reports*, 734:1 – 74, 2018.

[18] M. Batty : *The New Science of Cities*. The MIT Press, Cambridge, MA, 2013.

[19] D. Bawden et L. Robinson : The dark side of information : overload, anxiety and other paradoxes and pathologies. *Journal of information science*, 35(2):180–191, 2009.

[20] M. A. Beber : *Individual and group activity recognition in moving object trajectories*. Thèse de doctorat, Universidade federal de Santa Catarina, 2017.

[21] R. R. Behrens : Art, design and gestalt theory. *Leonardo*, 31(4):299–303, 1998.

[22] J.-P. Benzécri *et al.* : *L'analyse des données*, vol. 2. Dunod Paris, 1973.

[23] D. J. Berndt et J. Clifford : Using dynamic time warping to find patterns in time series. *ACM SIGKDD*, 10(16):359–370, 1994.

[24] T. Berners-Lee, J. Hendler et O. Lassila : The semantic web. *Scientific american*, 284(5):34–43, 2001.

[25] J. C. Bezdek : *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[26] D. Birant et A. Kut : St-dbscan : An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221, 2007.

[27] M. J. Bitner : Servicescapes : The impact of physical surroundings on customers and employees. *Journal of marketing*, 56(2):57–71, 1992.

[28] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira et L. O. Alvares : Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1):66–88, 2014.

[29] D. Bollegala, Y. Matsuo et M. Ishizuka : Measuring semantic similarity between words using web search engines. *In Proceedings of the 16th International Conference on World Wide Web*, p. 757–766, New York, NY, USA, 2007.

[30] B. Bouchon-Meunier : *La logique floue*. QUE SAIS-JE ? PUF, 1993.

[31] M. Boulakbech, N. Messai, Y. Sam, T. Devogele et L. Etienne : Smartloire : A web mashup based tool for personalized touristic plans construction. *In WETICE*, p. 259–260, 2016.

[32] P. Bourdieu : *La distinction : critique sociale du jugement*. Minuit, 1974.

[33] D. Boyd et K. Crawford : Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[34] D. Brockmann, L. Hufnagel et T. Geisel : The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[35] L. Cao : In-depth behavior understanding and use : the behavior informatics approach. *Information Sciences*, 180(17):3067–3085, 2010.

[36] CERTU : Guide méthodologique des enquêtes ménages déplacement. Rap. tech., 2008.

[37] S. Chardonnel et M. Stock : Time-geography, 2005.

[38] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang et L. Feng : Trajcompressor : An online map-matching-based trajectory compression framework leveraging vehicle heading direction and change. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2012–2028, 2019.

[39] L. Chen et R. Ng : On the marriage of lp-norms and edit distance. *In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, p. 792–803, 2004.

[40] L. Chen, M. T. Özsu et V. Oria : Robust and fast similarity search for moving object trajectories. *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, p. 491–502, 2005.

[41] S. Chen, B. Ma et K. Zhang : On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24):2365–2376, 2009.

[42] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini et A. Vespignani : The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.

[43] E. Cho, S. A. Myers et J. Leskovec : Friendship and mobility : user movement in location-based social networks. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 1082–1090, 2011.

[44] S.-S. Choi, S.-H. Cha et C. C. Tappert : A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48, 2010.

[45] R. L. Cilibrasi et P. M. Vitanyi : The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3):370–383, 2007.

[46] C. Claramunt : Extending ladkin's algebra on non-convex intervals towards an algebra on union-of regions. *In Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, p. 9–14, 2000.

[47] C. Claramunt, C. Parent et M. Thériault : Design patterns for spatio-temporal processes. *In Data Mining and Reverse Engineering*, p. 455–475. Springer, 1998.

[48] K. D. Cole, M. R. Yavari et P. K. Rao : Computational heat transfer with spectral graph theory : Quantitative verification. *International Journal of Thermal Sciences*, 153:106383, 2020.

[49] A. M. Colman : *A dictionary of psychology*. Oxford quick reference, 2015.

[50] H. Cramér : *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton university press, 2016.

[51] V. Cross, X. Yu et X. Hu : Unifying ontological similarity measures : A theoretical and empirical investigation. *International Journal of Approximate Reasoning*, 54(7):861–875, 2013.

[52] F. J. Damerau : A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[53] M. L. Damiani, A. Acquaviva, F. Hachem et M. Rossini : Learning behavioral representations of human mobility. *In Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, p. 367–376, 2020.

[54] D. L. Davies et D. W. Bouldin : A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[55] C. de Runz : *Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique*. Thèse de doctorat, Université de Reims-Champagne Ardenne, 2008.

[56] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer et R. Harshman : Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[57] M. Detyniecki : *Mathematical aggregation operators and video querying their applicationto video querying*. Thèse de doctorat, Université Paris 6, France, nov 2000.

[58] A. K. Dey : Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[59] M. M. Deza et E. Deza : *Encyclopedia of distances*. Springer, 2016.

[60] M. Djedaini, K. Drushku, N. Labroche, P. Marcel, V. Peralta et W. Verdeaux : Automatic assessment of interactive olap explorations. *Information Systems*, 82:148–163, 2019.

[61] D. H. Douglas et T. K. Peucker : Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica : the international journal for geographic information and geovisualization*, 10(2): 112–122, 1973.

[62] C. du Mouza et P. Rigaux : Mobility patterns. *GeoInformatica*, 9(4):297–319, 2005.

[63] J. C. Dunn : A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[64] S. Duroudier, S. Chardonnel, B. Mericskay, I. André-Poyaud, O. Bedel, S. Depeau, T. Devogele, L. Etienne, A. Lepetit, C. Moreau *et al.* : Données hétérogènes de mobilités quotidiennes : protocole de diagnostic qualité et d'apurement à partir de la base mobi'kids. *In Spatial Analysis and GEOmatics*, 2019.

[65] C. d'Amato, S. Staab et N. Fanizzi : On the influence of description logics ontologies on conceptual similarity. *In International Conference on Knowledge Engineering and Knowledge Management*, p. 48–63. Springer, 2008.

[66] F. El Outa, M. Francia, P. Marcel, V. Peralta et P. Vassiliadis : Towards a conceptual model for data narratives. *In International Conference on Conceptual Modeling*, p. 261–270. Springer, 2020.

[67] C. H. Elzinga et M. Studer : Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, 44(1):3–47, 2015.

[68] N. Emery, N. Markosian et M. Sullivan : Time. *In* E. N. Zalta, éd. : *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 édn, 2020.

[69] P. Esling et C. Agon : Time-series data mining. *ACM Computing Surveys*, 45(1):1–34, 2012.

[70] C. G. Esteban Ortiz-Ospina et M. Roser : Time use. *Our World in Data*, 2020. https ://ourworldindata.org/time-use.

[71] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.* : A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd*, vol. 96, p. 226–231, 1996.

[72] Eurostat. : Harmonised european time use surveys. Rap. tech., 2019.

[73] C. Faloutsos, M. Ranganathan et Y. Manolopoulos : Fast subsequence matching in time-series databases. *Acm Sigmod Record*, 23(2):419–429, 1994.

[74] R. M. Fano : Transmission of information : A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

[75] T. Fawcett et F. Provost : Activity monitoring : Noticing interesting changes in behavior. *In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 53–62, 1999.

[76] C. Ferrero, L. Alvares et V. Bogorny : Multiple aspect trajectory data analysis : Research challenges and opportunities. *GeoInformatica*, 17:56–67, 2016.

[77] C. A. Ferrero, L. M. Petry, L. O. Alvares, C. L. da Silva, W. Zalewski et V. Bogorny : Mastermovelets : discovering heterogeneous movelets for multiple aspect trajectory classification. *Data Mining and Knowledge Discovery*, 34(3): 652–680, 2020.

[78] R. Fileto, C. May, C. Renso, N. Pelekis, D. Klein et Y. Theodoridis : The baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98:104–122, 2015.

[79] A. S. Furtado, D. Kopanaki, L. O. Alvares et V. Bogorny : Multidimensional similarity measuring for semantic trajectories. *Trans. in GIS*, 20(2):280–298, 2016.

[80] A. Gabadinho, G. Ritschard, N. S. Mueller et M. Studer : Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4):1–37, 2011.

[81] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal et A. Taly : Explainable ai in industry. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 3203–3204, 2019.

[82] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao et P. S. Yu : A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(3):1–34, 2019.

[83] A. Giacometti, B. Markhoff et A. Soulet : Comparison table generation from knowledge bases. *In European Semantic Web Conference*, p. 179–194. Springer, 2021.

[84] F. Giannotti, M. Nanni, F. Pinelli et D. Pedreschi : Trajectory pattern mining. *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, p. 330–339, New York, NY, USA, 2007. Association for Computing Machinery.

[85] M. C. González, C. A. Hidalgo et A.-L. Barabási : Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[86] T. R. Gruber : A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[87] T. R. Gruber : Toward principles for the design of ontologies used for knowledge sharing ? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

[88] N. Guarino, D. Oberle et S. Staab : *What Is an Ontology ?*, p. 1–17. 2009.

[89] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti et D. Pedreschi : A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[90] R. H. Güting, F. Valdés et M. L. Damiani : Symbolic trajectories. *ACM Transactions on Spatial Algorithms and Systems*, 1(2):1–51, 2015.

[91] S. J. Haberman : The analysis of residuals in cross-classified tables. *Biometrics*, 29(1):205–220, 1973.

[92] T. Hägerstraand : What about people in regional science ? *Papers in regional science*, 24(1):7–24, 1970.

[93] M. Hajiaghayi, S. Seddighin et X. Sun : Massively parallel approximation algorithms for edit distance and longest common subsequence. *In Proceedings of the Thirtieth Annual ACM Symposium on Discrete Algorithms*, p. 1654–1672. SIAM, 2019.

[94] M. Halkidi, B. Nguyen, I. Varlamis et M. Vazirgiannis : Thesus : Organizing web document collections based on link semantics. *The VLDB Journal*, 12:320–332, 2003.

[95] R. W. Hamming : Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[96] S. Harispe, S. Ranwez, S. Janaqi et J. Montmain : Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254, 2015.

[97] S. Hasan, X. Zhan et S. V. Ukkusuri : Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, p. 1–8, 2013.

[98] R. J. Hathaway et J. C. Bezdek : Nerf c-means : Non-euclidean relational fuzzy clustering. *Pattern recognition*, 27(3):429–437, 1994.

[99] F. Hausdorff : *Grundzüge der mengenlehre*, vol. 7. von Veit, 1914.

[100] K. Høffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann et A.-C. Ngonga Ngomo : Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920, 2017.

[101] D. Holten : Hierarchical edge bundles : Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics*, 12(5):741–748, 2006.

[102] K. Hornsby et M. J. Egenhofer : Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1):177–194, 2002.

[103] A. Howard, C. Zhang et E. Horvitz : Addressing bias in machine learning algorithms : A pilot study on emotion recognition for intelligent systems. *In 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, p. 1–7. IEEE, 2017.

[104] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher et E. Adar : A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics*, 19(12):2406–2415, 2013.

[105] D. P. Huttenlocher, G. A. Klanderman et W. J. Rucklidge : Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[106] D. P. Huttenlocher, W. J. Rucklidge et G. A. Klanderman : Comparing images using the hausdorff distance under translation. *In Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 654–656, 1992.

[107] P. Jaccard : Distribution comparée de la flore alpine dans quelques régions des alpes occidentales et orientales. *Bulletin de la Société Vaudoise de Sciences Naturelle*, (31):81–92, 1902.

[108] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan et C. Shahabi : Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

[109] A. K. Jain, M. N. Murty et P. J. Flynn : Data clustering : a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[110] S. Jain, D. Moritz, D. Halperin, B. Howe et E. Lazowska : Sqlshare : Results from a multi-year sql-as-a-service experiment. *In Proceedings of the 2016 International Conference on Management of Data*, p. 281–293, 2016.

[111] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin et K. Lindén : Automatic language identification in texts : A survey. *Journal of Artificial Intelligence Research*, 65:675–782, 2019.

[112] T. S. Jepsen, C. S. Jensen et T. D. Nielsen : Graph convolutional networks for road networks. *In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 460–463, 2019.

[113] S. Jiang, J. Ferreira et M. C. González : Clustering daily patterns of human activities in the city. *DMKD*, 25(3):478–510, 2012.

[114] S. Jiang, J. Ferreira et M. C. González : Activity-based human mobility patterns inferred from mobile phone data : A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–219, 2017.

[115] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli et M. C. González : A review of urban computing for mobile phone traces : current methods, challenges and opportunities. *In Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, p. 1–9, 2013.

[116] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale et M. C. González : The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

[117] S. C. Johnson : Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[118] J. Kacprzyk et R. R. Yager : Linguistic summaries of data using fuzzy logic. *International Journal of General System*, 30(2):133–154, 2001.

[119] S. Kaplan : The restorative benefits of nature : Toward an integrative framework. *Journal of environmental psychology*, 15(3):169–182, 1995.

[120] L. Kaufman et P. Rousseeuw : *Finding Groups in Data : An Introduction to Cluster Analysis*. 2009.

[121] B. Kim, O. Koyejo, R. Khanna *et al.* : Examples are not enough, learn to criticize ! criticism for interpretability. *In NIPS*, p. 2280–2288, 2016.

[122] D. E. Knuth : Son of seminumerical algorithms. *ACM SIGSAM Bulletin*, 9(4):10–11, 1975.

[123] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov et A. J. Wyner : Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

[124] A. Krishna : An integrative review of sensory marketing : Engaging the senses to affect perception, judgment and behavior. *Journal of consumer psychology*, 22(3):332–351, 2012.

[125] W. Kuhn : Geospatial semantics : why, of what, and how ? *In Journal on data semantics III*, p. 1–24. Springer, 2005.

[126] P. B. Ladkin : Time representation : A taxonomy of internal relations. *In AAAI*, p. 360–366, 1986.

[127] L. Lamport : Time, clocks, and the ordering of events in a distributed system. *In Communications of the ACM*, p. 179–196, 1978.

[128] G. N. Lance et W. T. Williams : A general theory of classificatory sorting strategies : 1. hierarchical systems. *The computer journal*, 9(4):373–380, 1967.

[129] T. K. Landauer et S. T. Dumais : A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[130] A. Laurent : Generating fuzzy summaries from fuzzy multidimensional databases. *In International Symposium on Intelligent Data Analysis*, p. 24–33. Springer, 2001.

[131] D. Lazer, R. Kennedy, G. King et A. Vespignani : The parable of google flu : traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[132] C. Leacock et M. Chodorow : Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, 49(2): 265–283, 1998.

[133] A. L. Lehmann, L. O. Alvares et V. Bogorny : SMSM : a similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science*, 33(9):1847–1872, 2019.

[134] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.* : Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[135] L. Lesnard : Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419, 2010.

[136] M.-J. Lesot et R. Kruse : Data summarisation by typicality-based clustering for vectorial and non vectorial data. *In 2006 IEEE International Conference on Fuzzy Systems*, p. 547–554. IEEE, 2006.

[137] M.-J. Lesot et R. Kruse : Typicality degrees and fuzzy prototypes for clustering. *In Advances in Data Analysis*, p. 107–114. Springer, 2007.

[138] V. I. Levenshtein : Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.

[139] M. Li, X. Chen, X. Li, B. Ma et P. M. Vitányi : The similarity metric. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004.

[140] S. Li et D.-H. Lee : Learning daily activity patterns with probabilistic grammars. *Transportation*, 44(1):49–68, 2017.

[141] Y. Li, Z. A. Bandar et D. McLean : An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882, 2003.

[142] D. Lin *et al.* : An information-theoretic definition of similarity. *In International Conference on Machine Learning*, vol. 98, p. 296–304, 1998.

[143] F. Lordon : *Capitalisme, désir et servitude : Marx et Spinoza*. La fabrique éditions, 2010.

[144] J. MacQueen : Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, p. 281–297, 1967.

[145] J.-E. Mai : *Looking for information : A survey of research on information seeking, needs, and behavior*, chap. Information Overload and Anxiety, p. 103–108. Emerald Group Publishing, 2016.

[146] S. Mallat : Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

[147] E. Margolis et S. Laurence : Concepts. *In* E. N. Zalta, éd. : *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 édn, 2021.

[148] L. McInnes, J. Healy et J. Melville : Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.

[149] B. D. McKay et A. Piperno : Practical graph isomorphism, ii. *Journal of Symbolic Computation*, 60:94 – 112, 2014.

[150] M. G. McNally et C. R. Rindt : *The activity-based approach*. Emerald Group Publishing Limited, 2007.

[151] R. d. S. Mello, V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner et C. Renso : Master : A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822, 2019.

[152] A. Menin : *eSTIMe : a visualization framework for assisting a multi-perspective analysis of daily mobility data*. Thèse de doctorat, Université Grenoble Alpes [2020-....], 2020.

[153] A. Menin, S. Chardonnel, P.-A. Davoine et L. Nedel : estime : Towards an all-in-one geovisualization environment for daily mobility analysis. *In 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, p. 39–46. IEEE, 2019.

[154] C. B. Mervis et E. Rosch : Categorization of natural objects. *Annual review of psychology*, 32(1):89–115, 1981.

[155] T. Mikolov, I. Sutskever, K. Chen, G. Corrado et J. Dean : Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv :1310.4546*, 2013.

[156] G. A. Miller et W. G. Charles : Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

[157] H. J. Miller : A measurement theory for time geography. *Geographical analysis*, 37(1):17–45, 2005.

[158] H. J. Miller : Time geography and space–time prism. *International encyclopedia of geography : People, the earth, environment and technology*, p. 1–19, 2016.

[159] H. J. Miller et J. Han : *Geographic data mining and knowledge discovery*. CRC press, 2009.

[160] T. Miller : Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[161] M. Mongeau et D. Sankoff : Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175, 1990.

[162] C. Moreau, A. Chanson, V. Peralta, T. Devogele et C. de Runz : Clustering sequences of multi-dimensional sets of semantic elements. *In Proceedings of the 36th Annual ACM Symposium on Applied Computing*, p. 384–391, 2021.

[163] C. Moreau, T. Devogele, C. de Runz, V. Peralta, E. Moreau et L. Etienne : A fuzzy generalisation of the hamming distance for temporal sequences. *In FUZZ-IEEE 2021*, p. 1–8. IEEE, 2021.

[164] C. Moreau, T. Devogele, V. Peralta et L. Etienne : Contextual edit distance for semantic trajectories. *In Proceedings of the 35th Annual ACM Symposium on Applied Computing*, p. 635–637, 2020.

[165] C. Moreau, T. Devogele, V. Peralta, L. Etienne et C. de Runz : Methodology for mining, discovering and analyzing semantic human mobility behaviors. *arXiv preprint arXiv :2012.04767*, 2020.

[166] C. Moreau et V. Peralta : Learning analysis behavior in sql workloads. *In DOLAP*, p. 61–70, 2021.

[167] C. Moreau, V. Peralta, P. Marcel, A. Chanson et T. Devogele : Learning analysis patterns using a contextual edit distance. *In DOLAP 2020, EDBT/ICDT*, vol. 2572, p. 46–55, 2020.

[168] D. Müllner : fastcluster : Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, 53(9):1–18, 2013.

[169] S. A. Murray, M. Kendall, K. Boyd et A. Sheikh : Illness trajectories and palliative care. *British Medical Journal*, 330(7498):1007–1011, 2005.

[170] J.-P. Nakache et J. Confais : *Approche pragmatique de la classification : arbres hiérarchiques, partitionnements*. Editions Technip, 2004.

[171] National Aeronautics and Space Administration : The Prediction Of Worldwide Energy Resources (POWER), 2021.

[172] A. Y. Ng, M. I. Jordan et Y. Weiss : On spectral clustering : Analysis and an algorithm. *In Advances in Neural Information Processing Systems*, p. 849–856, 2002.

[173] D. Nguyen, W. Luo, T. D. Nguyen, S. Venkatesh et D. Phung : Sqn2vec : Learning sequence representation via sequential patterns with a gap constraint. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 569–584. Springer, 2018.

[174] D. Noël, M. Villanova-Oliver, J. Gensel et P. Le Quéau : Modeling semantic trajectories including multiple viewpoints and explanatory factors : application to life trajectories. *In Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, p. 107–113, 2015.

[175] D. Noel, M. Villanova-Oliver, J. Gensel et P. Le Quéau : Design patterns for modelling life trajectories in the semantic web. *In International Symposium on Web and Wireless Geographical Information Systems*, p. 51–65. Springer, 2017.

[176] T. P. Nogueira, R. B. Braga, C. T. de Oliveira et H. Martin : Framestep : A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications*, 92:533–545, 2018.

[177] R. L. Oliver : *Satisfaction : A behavioral perspective on the consumer : A behavioral perspective on the consumer*. Routledge, 2014.

[178] J. A. Ouellette et W. Wood : Habit and intention in everyday life : The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, 124(1):54, 1998.

[179] P. O'Neil, E. O'Neil, X. Chen et S. Revilak : The star schema benchmark and augmented fact table indexing. *In Technology Conference on Performance Evaluation and Benchmarking*, p. 237–252. Springer, 2009.

[180] A. T. Palma, V. Bogorny, B. Kuijpers et L. O. Alvares : A clustering-based approach for discovering interesting places in trajectories. *In Proceedings of the 2008 ACM symposium on Applied computing*, p. 863–868, 2008.

[181] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti et A.-L. Barabási : Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):1–8, 2015.

[182] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. G. Divanis, J. Macedo, N. Pelekis, Y. Theodoridis et Z. Yan : Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4), 2013.

[183] J. Pearl : *Causality*. Cambridge university press, 2009.

[184] J. Pearl et D. Mackenzie : *The book of why : the new science of cause and effect*. Basic books, 2018.

[185] J. Pei, J. Han, B. Mortazavi-Asl et H. Zhu : Mining access patterns efficiently from web logs. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, p. 396–407. Springer, 2000.

[186] V. Peralta : *Data quality evaluation in data integration systems*. Thèse de doctorat, Université de Versailles-Saint Quentin en Yvelines ; Université de la . . . , 2006.

[187] D. J. Peuquet : It's about time : A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, 84(3):441–461, 1994.

[188] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire et Y. W. Choong : Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):1–37, 2010.

[189] A. Pred : The choreography of existence : comments on hägerstrand's time-geography and its usefulness. *Economic geography*, 53(2):207–221, 1977.

[190] M. A. Quddus, W. Y. Ochieng et R. B. Noland : Current map-matching algorithms for transport applications : State-of-the art and future research directions. *Transportation research part c : Emerging technologies*, 15(5):312–328, 2007.

[191] J. R. Quinlan : *C4.5 : programs for machine learning*. Elsevier, 2014.

[192] R. Rada, H. Mili, E. Bicknell et M. Blettner : Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[193] W. M. Rand : Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[194] M. Raubal, H. J. Miller et S. Bridwell : User-centred time geography for location-based services. *Geografiska Annaler : Series B, Human Geography*, 86(4):245–265, 2004.

[195] J. Reeve et W. Lee : A neuroscientific perspective on basic psychological needs. *Journal of personality*, 87(1):102–114, 2019.

[196] P. Resnik : Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[197] S. Rizzi et E. Gallinucci : Cubeload : a parametric generator of realistic olap workloads. *In International Conference on Advanced Information Systems Engineering*, p. 610–624. Springer, 2014.

[198] E. Rosch : Cognitive representations of semantic categories. *Journal of experimental psychology : General*, 104(3):192, 1975.

[199] P. Rousseeuw : Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[200] C. Rovelli : *The order of time*. Riverhead books, 2019.

[201] S. T. Roweis et L. K. Saul : Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[202] H. Rubenstein et J. B. Goodenough : Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[203] Rui Xu et D. Wunsch : Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[204] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding et C.-T. Lin : A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

[205] M. Schläpfer, L. Dong, K. O'Keeffe, P. Santi, M. Szell, H. Salat, S. Anklesaria, M. Vazifeh, C. Ratti et G. B. West : The universal visitation law of human mobility. *Nature*, 593(7860):522–527, 2021.

[206] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda et M. C. González : Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[207] S. Schockaert, M. De Cock et E. E. Kerre : Fuzzifying allen's temporal interval relations. *IEEE Transactions on Fuzzy Systems*, 16(2):517–533, 2008.

[208] E. Schubert et P. J. Rousseeuw : Faster k-medoids clustering : Improving the PAM, CLARA, and CLARANS algorithms. *ArXiv*, abs/1810.05691, 2018.

[209] A. D. Selbst et J. Powles : Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 12 2017.

[210] Z. Shan, W. Sun et B. Zheng : Extract human mobility patterns powered by city semantic diagram. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[211] C. E. Shannon : A mathematical theory of communication. *Bell System Technical Journal,*, 27(3):379–423, 1978.

[212] J. Shi et J. Malik : Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[213] N. Shoval, B. McKercher, E. Ng et A. Birenboim : Hotel location and tourist activity in cities. *Annals of tourism research*, 38(4):1594–1612, 2011.

[214] W. Siabato, C. Claramunt, S. Ilarri et M. Á. Manso-Callejo : A survey of modelling trends in temporal gis. *ACM Computing Surveys (CSUR)*, 51(2):1–41, 2018.

[215] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.* : Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[216] V. Singh, J. Gray, A. Thakar, A. S. Szalay, J. Raddick, B. Boroski, S. Lebedeva et B. Yanny : Skyserver traffic report-the first five years. *arXiv preprint cs/0701173*, 2007.

[217] B. F. Skinner : *Science and human behavior*. Num. 92904. Simon and Schuster, 1965.

[218] T. F. Smith, M. S. Waterman *et al.* : Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[219] C. Song, T. Koren, P. Wang et A.-L. Barabási : Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[220] C. Song, Z. Qu, N. Blumm et A.-L. Barabási : Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[221] J. F. Sowa : Semantic networks. *Encyclopedia of Artificial Intelligence*, 1987.

[222] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto et C. Vangenot : A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1):126–146, 2008.

[223] M. Studer et G. Ritschard : What matters in differences between life trajectories : A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 179:481–511, 02 2016.

[224] R. Su, E. C. McBride et K. G. Goulias : Pattern recognition of daily activity patterns using human mobility motifs and sequence analysis. *Transportation Research Part C : Emerging Technologies*, 120:102796, 2020.

[225] D. d. C. Teixeira, A. C. Viana, M. S. Alvim et J. M. Almeida : Deciphering predictability limits in human mobility. *In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 52–61, 2019.

[226] J. B. Tenenbaum, V. De Silva et J. C. Langford : A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[227] H. Timmermans, T. Arentze et C.-H. Joh : Analysing space-time behaviour : new approaches to old problems. *Progress in human geography*, 26(2):175–190, 2002.

[228] Tripadvisor. : Tripbarometer 2016 – traveler trends & motivations global findings. Rap. tech., 2016.

[229] W. Tu, J. Cao, Y. Yue, S.-L. Shaw, M. Zhou, Z. Wang, X. Chang, Y. Xu et Q. Li : Coupling mobile phone and social media data : A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, 31(12):2331–2358, 2017.

[230] A. Tversky : Features of similarity. *Psychological review*, 84(4):327, 1977.

[231] A. Tversky et I. Gati : Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

[232] F. Valdés et R. H. Güting : A framework for efficient multi-attribute movement data analysis. *The VLDB Journal*, 28(4):427–449, 2019.

[233] S. L. Vargo et R. F. Lusch : Evolving to a new dominant logic for marketing. *Journal of marketing*, 68(1):1–17, 2004.

[234] M. Vlachos, G. Kollios et D. Gunopulos : Discovering similar multidimensional trajectories. *In Proceedings 18th International Conference on Data Engineering*, p. 673–684, 2002.

[235] G. A. Vouros, G. M. Santipantakis, C. Doulkeridis, A. Vlachou, G. Andrienko, N. Andrienko, G. Fuchs, J. M. C. Garcia et M. G. Martinez : The datacron ontology for the specification of semantic trajectories. *Journal on Data Semantics*, 8(4):235–262, 2019.

[236] G. A. Vouros, A. Vlachou, G. M. Santipantakis, C. Doulkeridis, N. Pelekis, H. V. Georgiou, Y. Theodoridis, K. Patroumpas, E. Alevizos, A. Artikis *et al.* : Big data analytics for time critical mobility forecasting : Recent progress and research challenges. *In EDBT*, p. 612–623, 2018.

[237] R. A. Wagner et M. J. Fischer : The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

[238] J. H. Ward Jr : Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[239] G. I. Webb, M. J. Pazzani et D. Billsus : Machine learning for user modeling. *User modeling and user-adapted interaction*, 11(1):19–29, 2001.

[240] W. Wood, J. M. Quinn et D. A. Kashy : Habits in everyday life : Thought, emotion, and action. *Journal of personality and social psychology*, 83(6):1281, 2002.

[241] W. Wood et D. Rünger : Psychology of habit. *Annual review of psychology*, 67:289–314, 2016.

[242] W. Wood, L. Tam et M. G. Witt : Changing circumstances, disrupting habits. *Journal of personality and social psychology*, 88(6):918, 2005.

[243] F. Wu, Z. Li, W.-C. Lee, H. Wang et Z. Huang : Semantic annotation of mobility data using social media. *In Proceedings of the 24th International Conference on World Wide Web*, p. 1253–1263, 2015.

[244] Z. Wu et M. Palmer : Verb semantics and lexical selection. *Association for Computational Linguistics*, p. 133–138, 1994.

[245] A. R. Wyler, M. Masuda et T. H. Holmes : Magnitude of life events and seriousness of illness. *Psychosomatic Medicine*, 1971.

[246] X. Yan, T. Ai, M. Yang et H. Yin : A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS journal of photogrammetry and remote sensing*, 150:259–273, 2019.

[247] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra et K. Aberer : Semitri : a framework for semantic annotation of heterogeneous trajectories. *In Proc. of EDBT*, p. 259–270, 2011.

[248] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra et K. Aberer : Semantic trajectories : Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–38, 2013.

[249] Z. Yan et S. Spaccapietra : Towards semantic trajectory data analysis : A conceptual and computational approach. *In VLDB PhD Workshop*, vol. 3, p. 23. Citeseer, 2009.

[250] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng et V. S. Tseng : Mining user similarity from semantic trajectories. *In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, p. 19–26, 2010.

[251] W. Youyou, M. Kosinski et D. Stillwell : Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[252] L. A. Zadeh : Fuzzy sets. *In Information and Control*, vol. 8, p. 338–353, 1965.

[253] E. Zgraggen, Z. Zhao, R. Zeleznik et T. Kraska : Investigating the effect of the multiple comparisons problem in visual analysis. *In Proceedings of the Conference on Human Factors in Computing Systems*, p. 1–12, 2018.

[254] C. Zhang, J. Han, L. Shou, J. Lu et T. Porta : Splitter : Mining fine-grained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 7:769–780, 05 2014.

[255] Y. Zheng, L. Zhang, X. Xie et W.-Y. Ma : Mining interesting locations and travel sequences from gps trajectories. *In Proceedings of the 18th international conference on World wide web*, p. 791–800, 2009.

[256] G. Zhu et C. A. Iglesias : Computing semantic similarity of concepts in knowledge graphs. *IEEE Trans. on Knowledge and Data Engineering*, 29(1):72–85, 2016.

Troisième partie

Annexes

Annexe A
Spatialisation des clusters MOBI'KIDS

Les figures suivantes ont été fournies par Laurent Etienne et présentent la spatialisation des clusters obtenus sur les deux premières années de recueil (sur les trois années au total) du jeu de données MOBI'KIDS. Les clusters sont ici centralisés autour de la ville de Rennes.

La mesure utilisée pour la comparaison des séquences est la mesure de CED avec une fonction d'encodage du vecteur temporel telle que décrite équation 7.5. La similarité utilisée pour comparer les symboles est la mesure de similarité entre ensembles d'éléments multidimensionnels décrite équation 7.3.

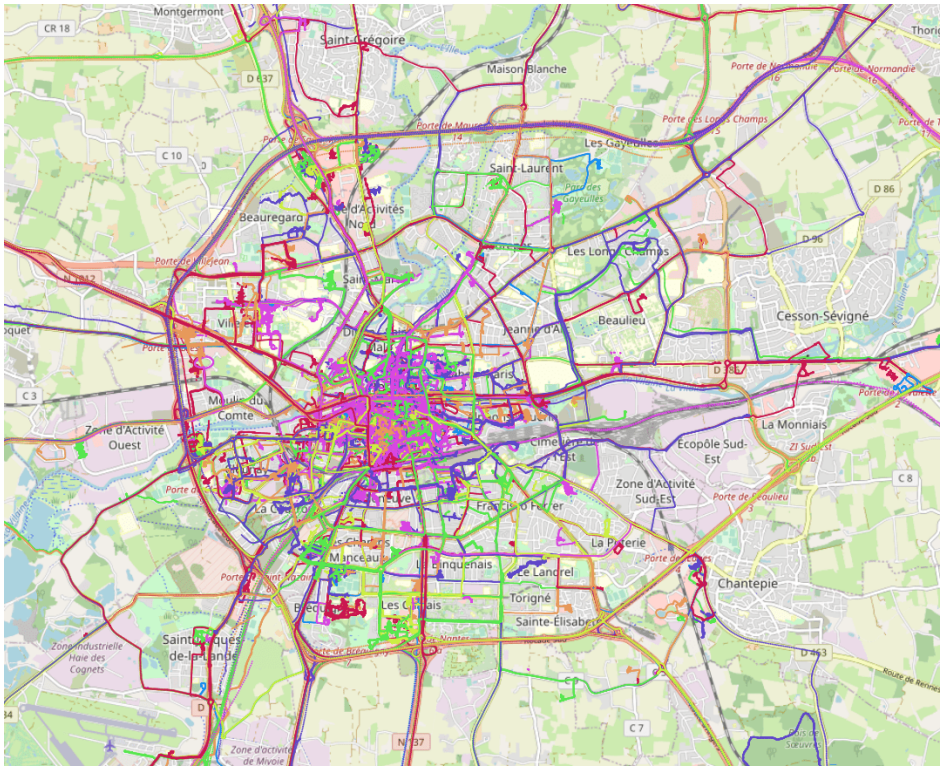


Figure A.1 – Ensemble des trajectoires obtenus sur l'agglomération rennaise. Les couleurs font référence au cluster affecté

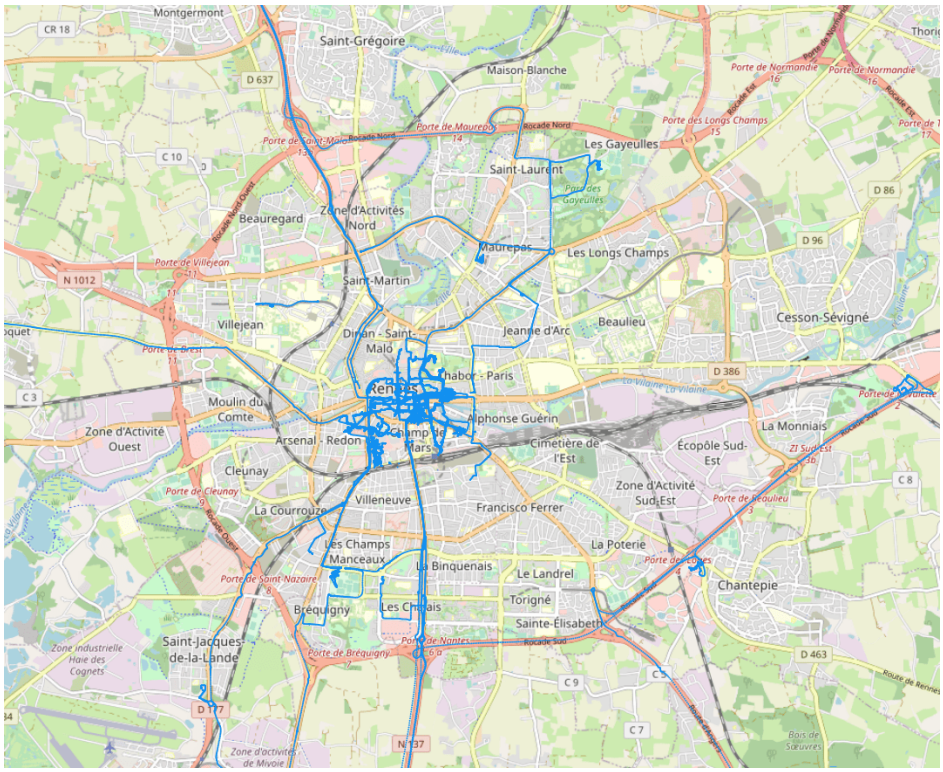


Figure A.2 – Cluster 1 – MOBI'KIDS Rennes

Les clusters ont été constitués par un algorithme de classification ascendante hiérarchique (algorithme AGNES).

Pour l'heure, nous ne constatons visuellement aucune corrélation immédiate ou triviale entre les dimensions spatiale et sémantique.

Il peut être noté toutefois certaines nuances de densité et d'étendue entre certains clusters (voir figures A.4 et A.5). Des analyses approfondies doivent être menées pour capter avec plus de précision et de méticulosité les différents lieux (privés et communs) et espaces traversés par ces traces (e.g., espaces verts, commerces, etc.). Toutefois, la gestion des lieux privés (e.g., domicile, travail) demande une gestion par individu qui peut se révéler coûteuse et difficile.

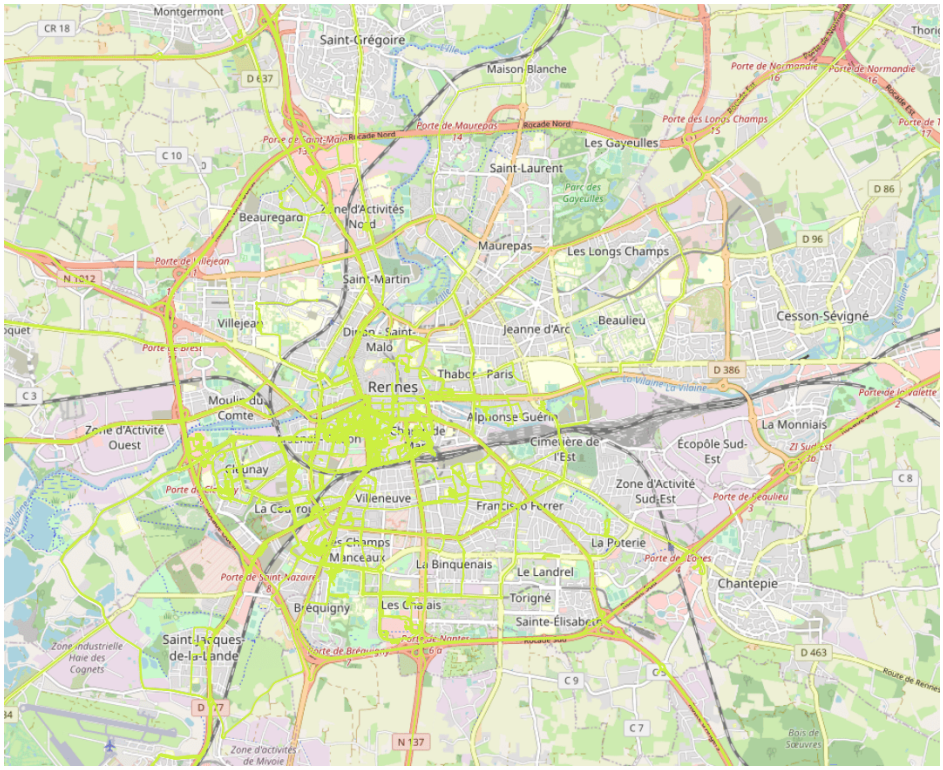


Figure A.3 – Cluster 2 – MOBI'KIDS Rennes

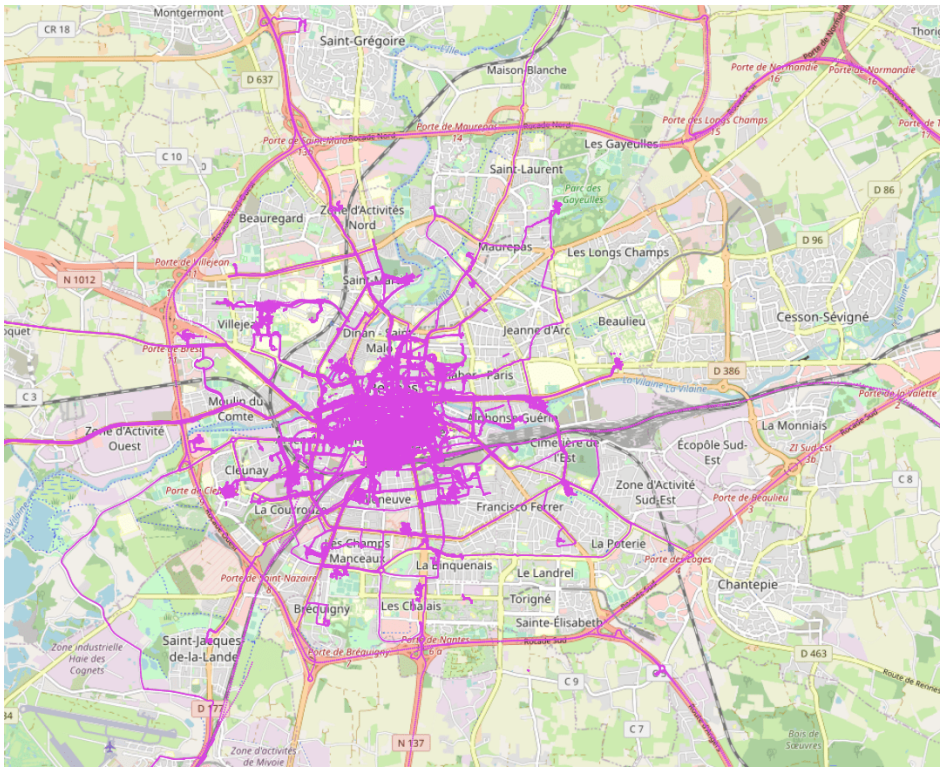


Figure A.4 – Cluster 3 – MOBI'KIDS Rennes

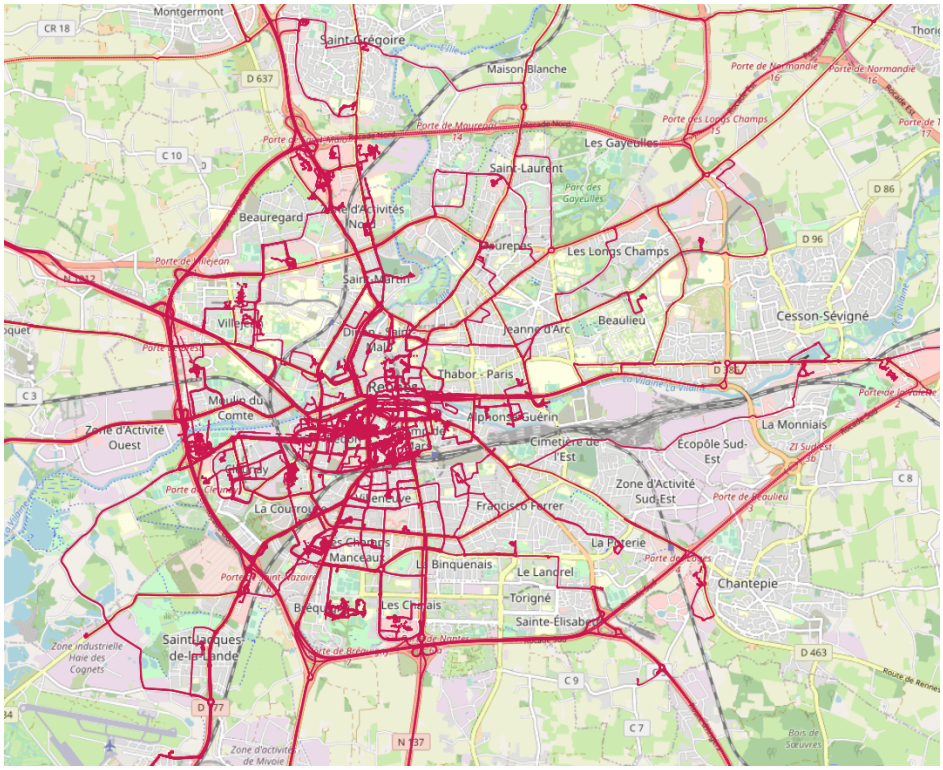


Figure A.5 – Cluster 4 – MOBI'KIDS Rennes

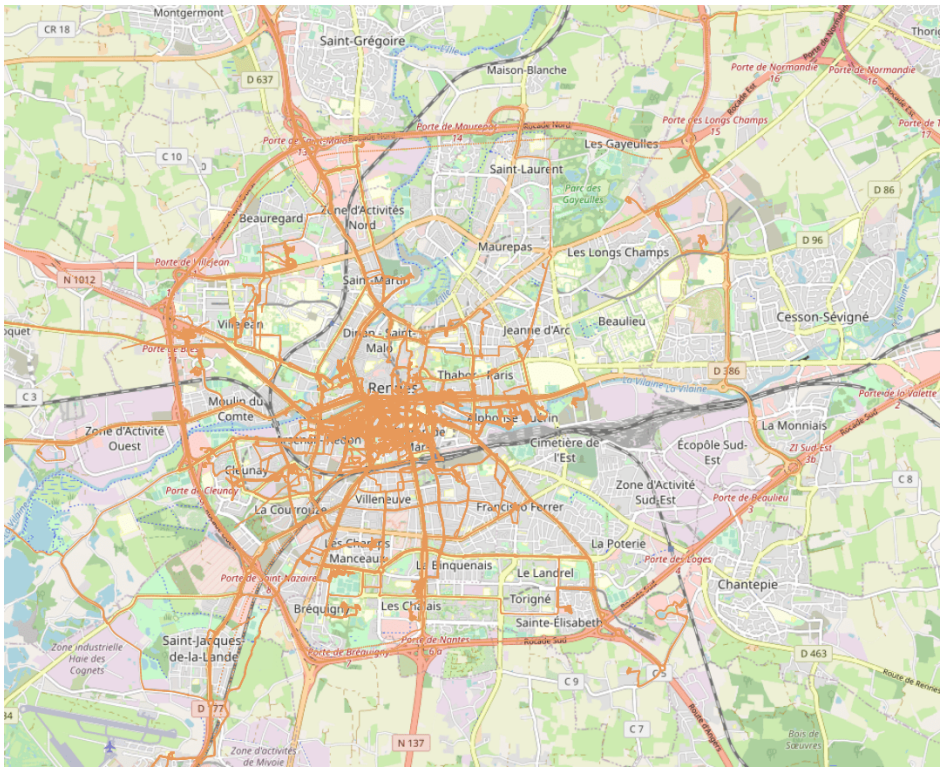


Figure A.6 – Cluster 5 – MOBI'KIDS Rennes

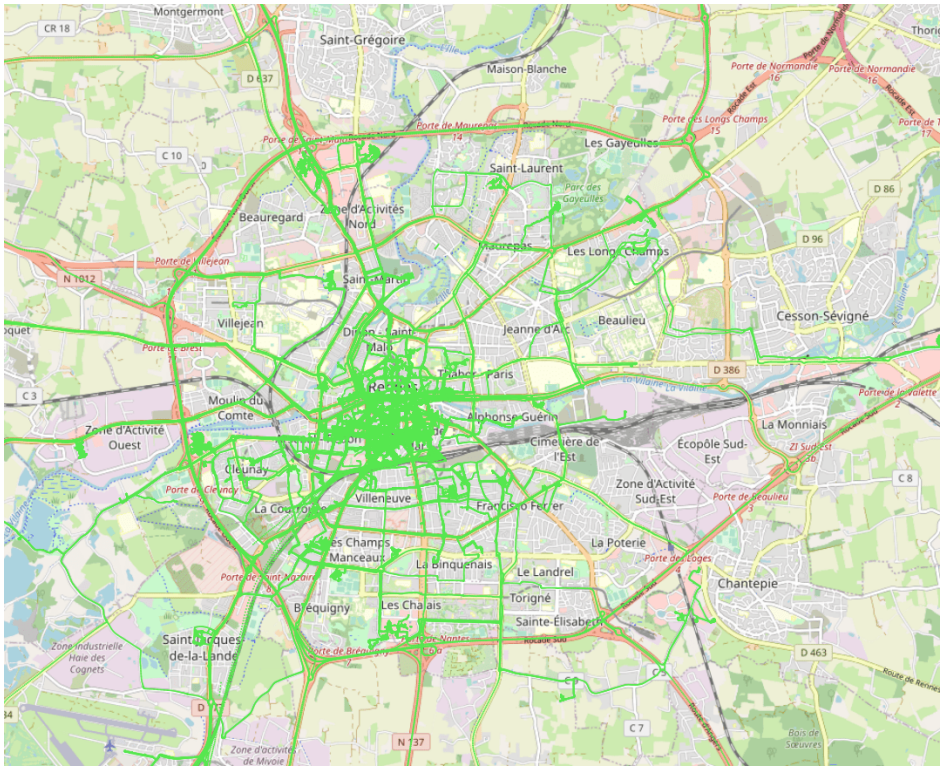


Figure A.7 – Cluster 6 – MOBI'KIDS Rennes

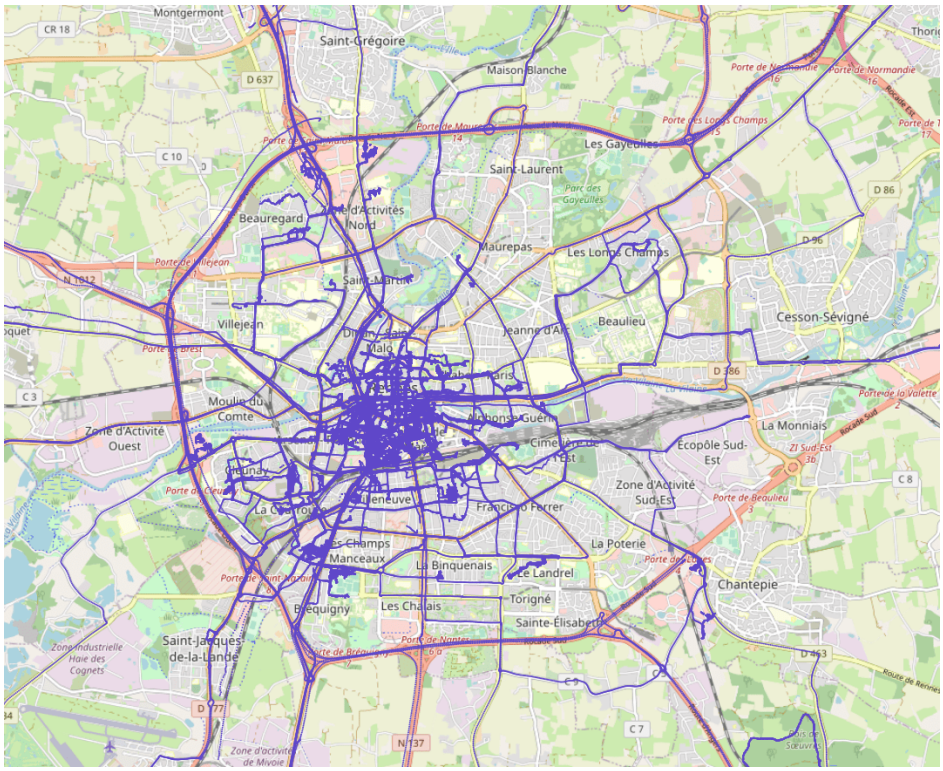


Figure A.8 – Cluster 7 – MOBI'KIDS Rennes

Annexe B
Liens projets GitHub

La page suivante liste les différents liens vers les repositories (code, expérimentation, données) des projets et articles réalisés au cours de la thèse. La page principale est

`https://github.com/Clement-Moreau-Info`

Listes des projets :

- SAC20 – *A Contextual Edit Distance for Semantic Trajectories*
  → `https://github.com/Clement-Moreau-Info/CED`
- Dolap20 – *Learning Analysis Patterns using a Contextual Edit Distance*
  → `https://github.com/Clement-Moreau-Info/DOLAP20`
- FuzzIEEE21 – *A fuzzy generalisation of the Hamming distance for temporal sequences*
  → `https://github.com/Clement-Moreau-Info/FTH`
- SAC21 – *Clustering Sequences of Multi-dimensional Semantic Elements*
  → `https://github.com/Clement-Moreau-Info/SAC2021`
- DMKD – *Methodology for mining, discovering and analyzing semantic human mobility behaviors*
  → `https://github.com/Clement-Moreau-Info/EMD2018_DMKD`
- SIMBA
  → `https://github.com/Clement-Moreau-Info/SIMBA`
- DOLAP21 – *Learning Analysis Behavior in SQL Workloads*
  → `https://github.com/Clement-Moreau-Info/DOLAP`
- Information Systems – *Mining SQL Workloads for Learning Analysis Behavior*
  → `https://github.com/Clement-Moreau-Info/IS-DOLAP21`

Annexe C
Curriculem Vitae