Mesure de similarité pour les trajectoires sémantiques : prise en compte de trois niveaux de granularité

Cécile Cayèré [1], Christian Sallaberry [2], Cyril Faucher [1], Marie-Noëlle Bessagnet [2], Philippe Roose [2], Maxime Masson [2]

1. La Rochelle Université,
23 Avenue Albert Einstein,
17000 La Rochelle, France
cecile.cayere1@univ-lr.fr,cyril.faucher@univ-lr.fr

2. Université de Pau et des Pays de l'Adour,
Avenue de l'Université,
64000 Pau, France
christian.sallaberry@univ-pau.fr,marie-noelle.bessagnet@univ-pau.fr,philippe.roose@iutbayonne.univ-pau.fr,maxime.masson@univ-pau.fr

RÉSUMÉ. *Ce papier s'inscrit dans le cadre du projet DA3T en collaboration avec des géographes. L'objectif est de proposer des méthodes et outils ayant pour objectif de traiter des traces de mobilité. Ce travail est expérimenté dans le domaine du tourisme en vue d'améliorer l'analyse de mobilité de touristes et par conséquent l'aménagement et la valorisation du territoire. Dans le cadre de la conception d'un module de calcul de similarité entre des trajectoires sémantiques, nous présentons une nouvelle mesure de similarité en vue de comparer deux déplacements selon trois dimensions (c.-à-d. spatiale, temporelle et thématique) et leurs trois niveaux de granularité (c.-à-d. micro, méso et macro).*

ABSTRACT. *This paper is part of the DA3T project in collaboration with geographers. The objective is to propose methods and tools to process mobility traces in order to improve their analysis and consequently touristic territory planning and valorization. As part of the design of a module for computing similarity between semantic trajectories, we present a new similarity measure for comparing two trips on three dimensions (i.e. spatial, temporal and thematic) and three granularity levels (i.e. micro, meso and macro).*

MOTS-CLÉS : *mesure de similarité ; trajectoire sémantique ; trace de mobilité*

KEYWORDS: *similarity measure ; semantic trajectory ; mobility track*

1. Introduction

La traçabilité de la mobilité humaine est un phénomène qui prend beaucoup d'ampleur de par l'évolution des technologies GPS et l'augmentation des déplacements humains. Dans le projet régional Nouvelle-Aquitaine DA3T (c.-à-d. Dispositif d'Analyse des Traces numériques pour la valorisation des Territoires Touristiques), nous exploitons les traces laissées par des touristes afin d'aider les décideurs locaux dans la gestion et l'aménagement des territoires touristiques. Il s'agit d'un projet pluridisciplinaire réunissant informaticiens et géographes dans l'objectif de produire des outils et des méthodes d'analyse de traces de mobilité.
Dans le cadre de ce projet, nous avons développé une application mobile, nommée Geoluciole, permettant de capturer les déplacements de touristes volontaires, auxquels nous avons fait passer des entretiens semi-directifs afin d'obtenir plus d'informations sur ces déplacements (p. ex. activités touristiques pratiquées). Nous avons conçu un modèle générique de description de trajectoires sémantiques et une plateforme modulaire permettant la conception et l'exécution de chaînes de traitement dédiées aux traces de mobilité. Ainsi, cette plateforme permet de paramétrer et d'enchaîner des modules de traitement de bas niveau en vue de répondre à un questionnement de plus haut niveau sur un jeu de traces de mobilité. Nous avons expérimenté ces propositions sur différents questionnements et jeux de données dédiés au tourisme, à la migration de colonies d'oiseaux ou encore aux activités d'observation menées par des naturalistes.
Dans cet article, nous nous intéressons à un module particulier de la plateforme. Il s'agit du module de calcul de similarité de deux trajectoires sémantiques conçu pour la comparaison de déplacements touristiques. Le travail présenté ici, considère les dimensions spatiale, temporelle et thématique de deux trajectoires sémantiques pour établir leur degré de similarité. Nous nous positionnons dans les domaines informatique et géomatique pour traiter des données de capteurs enrichies. Le verrou scientifique relève de la recherche d'information géographique (RIG) : il s'agit de proposer une nouvelle métrique de comparaison de trajectoires combinant trois dimensions et trois niveaux de granularité pour chaque dimension. L'originalité tient dans l'hypothèse de travail qui consiste à observer chaque couple de trajectoires selon trois dimensions à un niveau micro, méso et macro successivement.
L'article est organisé comme suit. La section 2 présente quelques définitions relatives aux trajectoires sémantiques et illustre nos motivations grâce à un scénario utilisant le jeu de données touristiques relatif à l'été 2020 à La Rochelle. La section 3 fait l'état de l'art des mesures de similarité (et de distance) dédiées aux trajectoires sémantiques. La section 4 rappelle les besoins des géographes, détaille notre hypothèse de travail et présente notre nouvelle métrique dédiée au calcul de similarité de trajectoires sémantiques. La section 5 évalue cette mesure au travers d'une expérimentation. Pour finir, la section 6 conclut cet article et propose quelques perspectives.

2. Trajectoires sémantiques touristiques

L'objet central de nos recherches est la trace de mobilité touristique, elle représente le déplacement d'un objet mobile (p. ex. un touriste) à travers une suite de positions géolocalisées et horodatées. Nous construisons des trajectoires brutes à partir de ces traces selon les besoins de l'analyse. En effet, le concept de trajectoire représente la sous-partie de la trace qui a un intérêt pour une application donnée (Parent *et al.*, 2013). Dans nos travaux, les trajectoires sont construites sur des critères spatiaux et/ou temporels (p. ex. la trace d'une semaine d'un touriste pourrait résulter en un ensemble de trajectoires journalières ; on s'intéresse à l'activité d'un touriste durant une journée). Les trajectoires brutes peuvent ensuite être enrichies avec des données externes et deviennent des trajectoires sémantiques.
Ces données d'enrichissement peuvent être de simples labels (p. ex. "Tour de la Lanterne") ou des objets complexes (p. ex. nom : "Tour de la Lanterne", type : "tour", localisation : [46.15579, -1.15712], etc.) et sont liées à la trajectoire : entière, à un segment ou à une position de celle-ci. Nous enrichissons les trajectoires avec des objets complexes pouvant représenter n'importe quel phénomène du monde réel, appelés aspects (Mello *et al.*, 2019). Dans notre modèle, un aspect est lié à la trajectoire par l'intermédiaire d'un ou plusieurs épisodes (c.-à-d. un intervalle temporel) qui définissent la ou les parties de la trajectoire enrichies par l'aspect. Les aspects d'un même type sont liés à une même séquence d'épisodes représentant un axe thématique particulier, appelée interprétation de la trajectoire (p. ex. séquence d'épisodes météorologiques).
Une trajectoire sémantique a une dimension temporelle (c.-à-d. les *timestamps*), une dimension spatiale (c.-à-d. les coordonnées spatiales) et un ensemble de dimensions sémantiques (c.-à-d. les interprétations).
Deux trajectoires sémantiques sont plus ou moins similaires sur une ou plusieurs de leurs dimensions. Par exemple, deux touristes peuvent suivre un même itinéraire sans pour autant pratiquer les mêmes activités ou se déplacer sur une même temporalité. Nous souhaitons comparer deux trajectoires sémantiques en tenant compte de toutes ces dimensions afin d'identifier si deux touristes ont des comportements similaires.

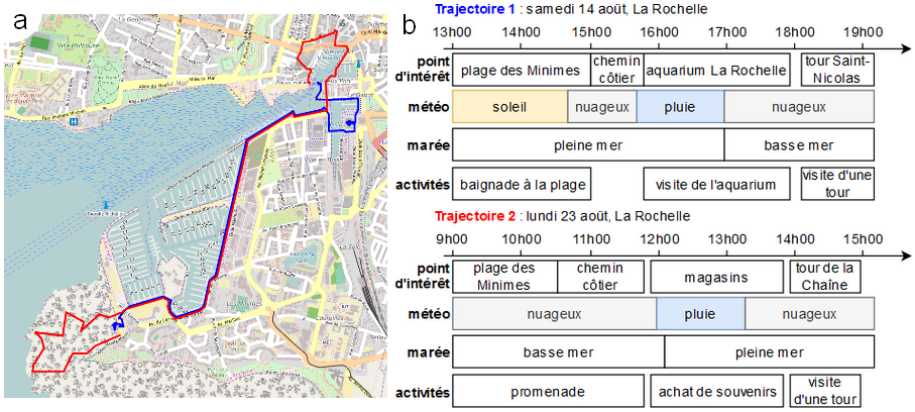


FIGURE 1. *Deux trajectoires sémantiques appartenant à deux touristes différents*

Prenons un exemple : comparons manuellement deux trajectoires sémantiques afin d'identifier les éléments essentiels de comparaison. La figure 1 montre deux trajectoires sémantiques construites à partir de traces collectées. La partie a de la figure illustre cartographiquement la dimension spatiale des trajectoires et la partie b met en évidence la dimension temporelle à travers un axe du temps ainsi que la dimension thématique grâce à la représentation des différentes interprétations des trajectoires. Les types d'aspects considérés ici sont les points d'intérêt, la météo, la marée et les activités touristiques (issues des entretiens).

Dimension spatiale (c.f. figure 1, a) : Des similitudes spatiales sont clairement visibles entre les trajectoires 1 et 2 (respectivement, bleue et rouge). Les deux se situent au centre-ville de La Rochelle et ont un point de départ et d'arrivée plus ou moins similaires (dans les mêmes zones). Les deux trajectoires comportent deux phases plutôt stationnaires ou de visite (c.-à-d. les zones de départ et d'arrivée où les positions sont plus proches les unes des autres) séparées par une phase de déplacement (c.-à-d. les longues lignes sans détour où les positions sont plus éloignées les unes des autres). De plus, il est à noter que les deux touristes traversent la ville en suivant le même chemin. Cependant, les arrêts (c.-à-d. amas de points aux mêmes endroits) que nous pouvons identifier à l'oeil nu ne sont pas les mêmes. Pour améliorer la comparaison, il faudrait pouvoir comparer les deux séquences de coordonnées géographiques. Nous pouvons déduire de toutes ces analyses que, malgré quelques légères différences, les deux trajectoires sont très similaires sur le plan spatial.
Dimension temporelle (c.f. figure 1, b) : Côté dimension temporelle, les deux se passent au mois d'août (c.-à-d. l'été). Ensuite, nous pouvons remarquer que l'une des trajectoires se passe le week-end (c.-à-d. un samedi) et l'autre en semaine (c.-à-d. un lundi). La durée des deux déplacements est de 7 heures environ mais ils ne se déroulent pas aux mêmes moments de la journée (l'un se déroule l'après-midi de 13h00 à 19h00, l'autre de 9h00 à 15h00). Ainsi, nous pouvons conclure que mis-à-part la saison et le mois, les trajectoires sont plutôt différentes sur le plan temporel.
Dimension thématique (c.f. figure 1, b) : Pour finir, concernant la dimension thématique, quatre interprétations enrichissent la trajectoire (à savoir, les points d'intérêt traversés par le touriste, la météo, la marée et les activités touristiques mentionnées dans l'entretien). Chaque épisode d'une interprétation (p. ex. "plage des Minimes") est un aspect décrit par un ensemble d'attributs non représenté ici pour ne pas surcharger la figure. La plus longue sous-séquence partagée par deux trajectoires est appelée plus longue séquence commune (Vlachos *et al*., 2002). En considérant uniquement les points d'intérêt traversés par les touristes, la plus longue séquence commune aux deux trajectoires est ⟨plage des Minimes, chemin côtier⟩. Nous pouvons aller plus loin grâce aux types (c.-à-d. l'attribut "type") des points d'intérêt, ce qui donne la plus longue séquence commune ⟨plage, chemin, tour⟩. En nous intéressant à toutes les interprétations en même temps, nous obtenons la plus longue séquence commune suivante : ⟨plage des Minimes, (plage des Minimes, nuageux),(chemin côtier, nuageux), chemin côtier, pluie, (pleine mer, pluie), nuageux, (visite dune tour, nuageux)⟩. En observant cette séquence, on se rend compte que les trajectoires sont plutôt similaires sur le plan thématique.

La similarité (inverse de distance) entre deux trajectoires peut être évaluée grâce à une fonction de similarité (ou de distance) permettant d'attribuer un score qui varie selon leur ressemblance ou leur différence. Un score de similarité est élevé lorsque les trajectoires se ressemblent et faible lorsqu'elles diffèrent ; inversement, un score de différence est élevé lorsque les trajectoires diffèrent l'une de l'autre et faible lorsqu'elles se ressemblent. De nombreuses fonctions de calcul de similarité se basent sur une ou plusieurs dimensions des trajectoires, quelques unes sont présentées dans la section suivante.

3. Travaux connexes

Cette section a pour but de faire le tour des mesures de similarité (et de distance) existantes permettant d'évaluer la ressemblance de deux trajectoires sur une ou plusieurs dimensions. Il existe déjà plusieurs travaux s'intéressant à comparer et classifier les mesures permettant de comparer des trajectoires (Wang *et al*., 2013; Magdy *et al*., 2015; Cleasby *et al*., 2019; Su *et al*., 2020; Tao *et al*., 2021), cependant la dimension thématique est souvent omise, n'étant pas la dimension centrale de description du déplacement d'un objet mobile. Dans un premier temps, nous présentons les mesures de similarité spatiale. Dans un second temps, nous présentons les mesures de similarité temporelle. Dans un troisième temps, nous présentons les mesures de similarité thématique. Enfin, nous abordons le cas particulier des mesures s'intéressant aux séries temporelles qui peuvent être utilisées pour comparer les différentes dimensions des trajectoires.

La dimension spatiale d'une trajectoire GPS est une suite de coordonnées GPS, c.-à-d. une suite de paires $(longitude, latitude)$ qui représente plus ou moins fidèlement l'itinéraire emprunté par l'objet mobile. Pour calculer la similarité spatiale entre deux trajectoires, nous pouvons les considérer comme des suites de points, comme des suites de segments ou nous pouvons les simplifier à leurs enveloppes englobantes. Cela revient à calculer la similarité entre des points, entre des lignes ou entre des polygones.
Pour calculer la distance entre deux points d'une trajectoire, il est possible d'utiliser la distance euclidienne (c.-à-d. L_2_Norm) ou la distance de Manhattan (c.-à-d. L_1_Norm). La distance euclidienne (ou ED) (Faloutsos *et al*., 1994) peut être appliquée sur les points d'un espace euclidien à une dimension (p. ex. sur des éléments de séries temporelles) ou plusieurs dimensions (p. ex. sur des points de trajectoires). Pour mesurer la distance entre deux trajectoires dans un espace euclidien, il est possible d'utiliser la distance euclidienne entre les points correspondants des deux trajectoires (distance entre le i-ème point d'une trajectoire avec le i-ème point de l'autre trajectoire) puis d'additionner toutes les distances calculées. C'est la distance euclidienne à étapes bloquées (*lock-step euclidean distance*) (Tao *et al*., 2021). Pour calculer la distance entre deux points GPS (sans conversion vers un espace euclidien), il existe la formule de Haversine utilisée pour la première fois en 1805 par James Andrew (An-

drew, 1805).
D'autres mesures de similarité spatiale se basent sur la division des trajectoires en segments ou en sous-trajectoires qu'elles comparent deux à deux comme la mesure SpADe (*Spatial Assembling Distance*) (Y. Chen *et al*., 2007), la distance de Hausdorff (Alt, 2009), AMSS (*Angular metric for shape similarity*) (Nakamura *et al*., 2013) et TRACLUS (*TRAjectory CLUStering*) (Lee *et al*., 2007). Nous réutilisons TRACLUS qui compare deux segments sur trois éléments importants (c.-à-d. parallélisme, distance et angle).
Nous souhaitons comparer les polygones englobants les trajectoires. Pour cela, nous pouvons utiliser le système de raisonnement RCC-8 (*Region Connection Calculus*) qui étend les relations entre intervalles temporels d'Allen aux polygones spatiaux (Aiello, 2002)(Sallaberry, 2013) (p. ex. deux polygones sont déconnectés, se superposent, etc.) ou les 9-intersections (Egenhofer, 1997) qui décrivent les relations topologiques pouvant s'appliquer à des polygones, des lignes et des points. La mesure de similarité appliquée à la recherche d'information spatiale présentée dans Le Parc-Lacayrelle *et al*. (2007) s'appuie sur l'intersection de deux polygones pour évaluer leur similarité, avec un score nul lorsqu'il n'y a pas d'intersection. Nous réutilisons cette dernière mesure car elle permet de comparer deux trajectoires sur un gros grain de détail en utilisant leurs boîtes englobantes.

La dimension temporelle d'une trajectoire GPS est une suite de marqueurs temporels (*timestamps*). Chaque marqueur est lié à un point de la trajectoire ; le tout représente le déplacement de l'objet mobile observé. La similarité temporelle entre deux trajectoires est souvent calculée de pair avec la dimension spatiale. Cependant, nous nous intéressons dans cette section à la dimension temporelle uniquement. Pour calculer cette similarité, nous pouvons considérer les trajectoires comme des suites de marqueurs temporels ou des intervalles temporels.
Les relations d'Allen (Allen, 1983) sont un ensemble de 13 relations entre intervalles temporels (p. ex. les intervalles sont égaux, se rencontrent, etc.). Pour une paire d'intervalles donnés, ces relations renvoient des résultats booléens.
La mesure de similarité appliquée à la recherche d'information temporelle présentée dans Le Parc-Lacayrelle *et al*. (2007) s'appuie sur l'intersection entre deux intervalles temporels pour évaluer leur similarité, avec un score nul lorsqu'il n'y a pas d'intersection. Nous réutilisons cette mesure car elle permet de comparer deux trajectoires sur un gros grain de détail en utilisant les intervalles de temps englobants des trajectoires.

La dimension thématique d'une trajectoire sémantique est un ensemble d'interprétations, c.-à-d. un ensemble de séquences d'épisodes temporels liés à des aspects d'un certain type (p. ex. météo, points d'intérêt, etc.). Ainsi, chaque position correspond à un certains nombres d'épisodes appartenant à différentes interprétations dont il faut tenir compte dans le calcul de similarité. Dans les travaux connexes, afin d'évaluer la similarité thématique des trajectoires sémantiques, ces dernières sont considérées comme des suites de données d'enrichissement simples (p. ex. label) ou complexes (p. ex. aspects) (chacune correspondant à une position), des suites d'épisodes sémantiques liés à des données d'enrichissement simples ou complexes ou des labels prin-

cipaux résumant des interprétations spécifiques des trajectoires. Il existe des mesures spécifiquement destinées à comparer des trajectoires multi-aspects comme la mesure TRAFOS (Varlamis *et al*., 2021) ou la mesure MUITAS (*MUltIple-aspect TrAjectory Similarity*) (May Petry *et al*., 2019). Nous réutilisons MUITAS car elle compare des trajectoires multi-aspect telles que nos trajectoires, des seuils sont appliqués pour comparer chaque attribut de chaque aspect et des pondérations régulent l'importance de chaque type d'aspect dans le calcul du score global. De plus, contrairement à TRAFOS, MUITAS ne nécessite pas l'utilisation de toutes les trajectoires du jeu de données pour calculer la similarité entre deux trajectoires. Certaines mesures considèrent les trajectoires comme des suites d'épisodes sémantiques comme la mesure LBS-Alignment (Lu, Tseng, 2009) ou la distance d'édition enrichie (Moreau *et al*., 2018) où toute donnée d'enrichissement comparée doit être considérée au sein d'une ontologie ou hiérarchie de concepts pour être comparée.

Les mesures LCSS/LCS (*Longest Common Subsequence*) (Vlachos *et al*., 2002), EDR (*Edit Distance on Real sequence*) (L. Chen *et al*., 2005), ERP (*Edit distance with Real Penalty*) (L. Chen, Ng, 2004) et DTW (*Dynamic Time Warping*) (Keogh, Ratanamahatana, 2005) sont des mesures permettant de comparer des séries temporelles qui peuvent être utilisées dans notre contexte. Elles consistent à étudier la proximité des éléments des séries, deux à deux, pour choisir les meilleures correspondances et calculer un score de similarité (ou de distance) final. Plus les éléments mis en correspondance sont éloignés, plus grande sera la pénalité ajoutée au score final. Ces mesures prennent en compte l'ordre entre les éléments mais pas l'écart temporel qui les séparent. Elles peuvent être utilisées pour calculer la similarité des différentes dimensions de la trajectoire. Par exemple, nous réutilisons DTW pour la similarité spatiale qui est est très adaptée au calcul de la similarité spatiale car elle utilise directement la distance entre les points sans seuil (dont la valeur peut dépendre de la taille de la trajectoire) et EDR pour la similarité des séquences thématiques car la correspondance des aspects peut être contrôlée avec un seuil.

4. Mesure de similarité DA3T dédiée aux trajectoires sémantiques

La mesure de similarité DA3T est dédiée à la comparaison de trajectoires sémantiques de mobilité. Nous commençons par rappeler les besoins exprimés par les partenaires du projet. Nous détaillons ensuite l'hypothèse sur laquelle repose cette nouvelle métrique. Enfin, nous présentons la mesure DA3T.

4.1. Rappel des besoins

Dans le projet DA3T, les géographes veulent comparer des trajectoires deux à deux (p. ex. comparaison de trajectoires représentatives appartenant à deux catégories de visiteurs différentes, comparaison d'une trajectoire de touriste avec un parcours type de l'Office du tourisme, etc.). Ils souhaitent une mesure paramétrable (pour les dimensions spatiale, temporelle et thématique) et calculée automatiquement car ce

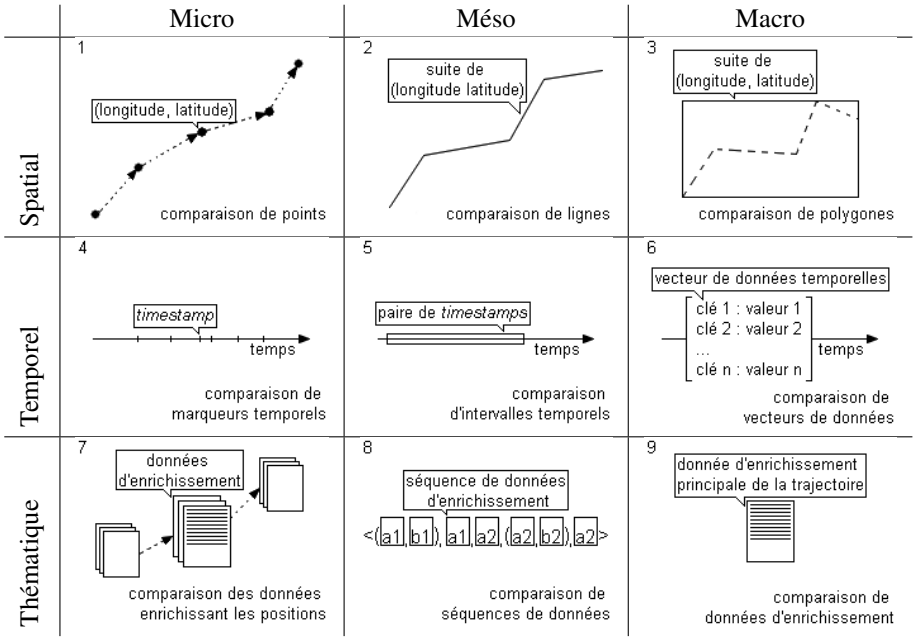
type de comparaison est long et, à ce jour, uniquement réalisé par des experts en géographie du tourisme.

4.2. Hypothèses de travail pour une nouvelle mesure

Nous faisons l'hypothèse que si chaque dimension, spatiale, temporelle et thématique est observée selon trois niveaux de granularité, respectivement micro, méso et macro, nous calculons un score de similarité entre deux trajectoires avec plus de précision. Le tableau 1 illustre ces niveaux de granularité.
Concernant la **dimension spatiale**, à l'échelle micro (c.f. tableau 1, 1), nous comparons deux trajectoires point à point. À l'échelle méso (c.f. tableau 1, 2), nous comparons des sous-parties (segments) de ces trajectoires afin de découvrir si elles ont la même tendance générale. Ainsi, par exemple, pour deux trajectoires de touristes ayant empruntés la même rue sur une partie de leurs déplacements, nous pouvons détecter une forte similarité au grain méso. Enfin, à l'échelle macro (c.f. tableau 1, 3), nous comparons deux trajectoires par rapport à la taille, la forme et le chevauchement de leurs boites englobantes (*bounding-box*). Cela permet d'identifier si deux objets mobiles ont globalement les mêmes comportements de déplacement.
Concernant la **dimension temporelle**, à l'échelle micro (c.f. tableau 1, 4), nous comparons les *horodatages* des positions des deux trajectoires. À l'échelle méso (c.f. tableau 1, 5), comme pour le spatial, nous comparons des segments/intervalles temporels correspondants à des couples de points horodatés. Enfin, à l'échelle macro (c.f. tableau 1, 6), nous comparons le contexte temporel relatif à chaque trajectoire pour savoir si elles se déroulent durant une même année, une même saison, un même mois, un même jour de semaine, etc. Ainsi, par exemple, une trajectoire journalière se passant un mardi du mois d'août 2020 et une autre s'étalant sur un week-end du mois d'août 2021 sont en partie similaires car elles se déroulent toutes les deux au mois d'août, mais pas la même année ni le même jour de la semaine.
Enfin, concernant la **dimension thématique**, à l'échelle micro (c.f. tableau 1, 7), nous comparons les données d'enrichissement associées aux positions des trajectoires. Par exemple, prenons un point d'une trajectoire enrichi avec les données suivantes : {nom : "Tour de la Lanterne", catégorie : "tour", ...} et {description : "soleil", température : 30, ...} et un point d'une seconde trajectoire enrichi avec : {nom : "Tour de la Lanterne", catégorie : "tour", ...} et {description : "soleil", température : 11, ...} Les deux points sont similaires sur plusieurs attributs : les touristes était proche du même point d'intérêt et il faisait soleil, mais il ne faisait pas la même température. À l'échelle méso (c.f. tableau 1, 8), nous comparons des séquences de données thématiques. Par exemple, deux trajectoires enrichies avec le découpage administratif et décrites respectivement par les séquences ⟨Les Minimes, Saint-Nicolas, Les Minimes⟩ et ⟨Les Minimes, Saint-Nicolas, Centre-ville⟩, se ressemblent sur les deux premiers éléments de la séquence mais ne sont pas identiques. Pour finir, à échelle macro (c.f. tableau 1, 9), nous comparons les thématiques dominantes. Prenons par exemple deux trajectoires de touristes enrichies avec des données de météo ; l'une peut être résumée

par le label "nuageux", l'autre par le label "ensoleillé". Ces deux trajectoires sont donc globalement différentes.

TABLEAU 1. *Tableau récapitulatif des dimensions et niveaux de granularité d'une trajectoire sémantique*



Ainsi, nous faisons l'hypothèse (H1) qu'introduire différents niveaux de granularité (c.-à-d. micro, méso et macro) pour chacune des dimensions (c.-à-d. spatiale, temporelle et thématique) permettra d'améliorer les mesures existantes.

4.3. Mesure de similarité DA3T

Pour valider l'hypothèse (H1), nous mettons en place une formule de calcul de score de similarité qui combine des sous-scores spatial, temporel et thématique pondérés par des coefficients. Chacun de ces sous-scores est à son tour la combinaison de trois scores de différents niveaux de granularité (c.f. tableau 1). Nous expérimentons ensuite la formule sur un jeu de données issues d'une campagne d'étude de mobilité touristique dans la ville de La Rochelle. Nous comparons les résultats obtenus avec ceux issus de formules de comparaisons existantes ainsi qu'avec l'avis d'experts en géographie.
Notre mesure est définie par l'équation 1.

$$S_{\text{glb}} = \alpha_{\text{spt}} * S_{\text{spt}} + \beta_{\text{tmp}} * S_{\text{tmp}} + \gamma_{\text{thm}} * S_{\text{thm}} \quad (1)$$

Dans cette formule, chaque sous-fonction de calcul de similarité relatif à une dimension spécifique peut être détaillée en trois niveaux de granularité, telle que :

$$S_{\text{spt}} = \alpha_{\text{spt-mic}} * S_{\text{spt-mic}} + \alpha_{\text{spt-mes}} * S_{\text{spt-mes}} + \alpha_{\text{spt-mac}} * S_{\text{spt-mac}} \quad (2)$$

$$S_{\text{tmp}} = \beta_{\text{tmp-mic}} * S_{\text{tmp-mic}} + \beta_{\text{tmp-mes}} * S_{\text{tmp-mes}} + \beta_{\text{tmp-mac}} * S_{\text{tmp-mac}} \quad (3)$$

$$S_{\text{thm}} = \gamma_{\text{thm-mic}} * S_{\text{thm-mic}} + \gamma_{\text{thm-mes}} * S_{\text{thm-mes}} + \gamma_{\text{thm-mac}} * S_{\text{thm-mac}} \quad (4)$$

La somme des coefficients de pondération d'un même niveau (c.-à-d. α_*, β_* et γ_*) est toujours égal à 1 telle que : $\alpha_{\text{spt}} + \beta_{\text{tmp}} + \gamma_{\text{thm}} = 1$, $\alpha_{\text{spt-mic}} + \alpha_{\text{spt-mes}} + \alpha_{\text{spt-mac}} = 1$, $\beta_{\text{tmp-mic}} + \beta_{\text{tmp-mes}} + \beta_{\text{tmp-mac}} = 1$ et $\gamma_{\text{thm-mic}} + \gamma_{\text{thm-mes}} + \gamma_{\text{thm-mac}} = 1$. De plus, toute mesure de similarité S est telle que : $0 \leq S \leq 1$. Plus le score de similarité est proche de 1, plus les trajectoires sont similaires ; plus le score de similarité est proche de 0, plus elles sont différentes.
Dans l'équation 1, nous ré-utilisons certaines mesures de similarité existantes. Premièrement pour les mesures de la dimension spatiale (c.f. équation 2), nous utilisons les suivantes :

– $S_{spt-mic}$ → DTW (Keogh, Ratanamahatana, 2005) et distance de Haversine (Andrew, 1805) : Pour comparer les trajectoires à l'échelle des points, nous utilisons la mesure DTW qui a pour avantage d'utiliser directement la distance spatiale entre les points pour calculer la distance total entre les trajectoires. Pour évaluer la distance entre les points, nous utilisons la distance de Haversine qui permet de mesurer la distance entre deux points GPS sur le plan terrestre. Ainsi, nous n'avons pas besoin de convertir nos données vers un espace euclidien.

– $S_{spt-mes}$ → TRACLUS (Lee *et al.*, 2007) : Pour comparer les trajectoires à l'échelle des segments, nous utilisons la mesure TRACLUS car elle prend en compte différentes caractéristiques des segments (c.-à-d. leur parallélisme, leur distance et leur angle) pour les comparer et chacun des ces types de comparaison peut également être pondéré.

– $S_{spt-mac}$ → Mesure de similarité appliquée à la RI spatiale (Le Parc-Lacayrelle *et al.*, 2007) : Pour comparer les boites englobantes des trajectoires, nous utilisons la mesure de similarité appliquée à la RI spatiale qui utilise l'intersection entre les polygones pour calculer leur distance.

Deuxièmement, pour les mesures de la dimension temporelle (c.f. équation 3), nous utilisons les suivantes :

– $S_{tmp-mic}$ → EDR (L. Chen *et al.*, 2005) appliqué à des séries de labels : Pour comparer les trajectoires à l'échelle des *timestamps*, nous attribuons une période de la journée (p. ex. matin, après-midi, soir, etc.) à chaque *timestamp* et nous utilisons EDR pour comparer deux suites de périodes associées aux trajectoires. Nous considérons qu'il y a correspondance entre deux périodes si elles sont exactement égales.

– $S_{tmp-mes}$ → Mesure de similarité appliquée à la RI temporelle (Le Parc-Lacayrelle *et al.*, 2007) : Pour comparer les trajectoires à l'échelle des intervalles temporels, nous ramenons les intervalles temporels des trajectoires à une échelle jour-

nalière. Nous appliquons ensuite la mesure de similarité appliquée à la RI temporelle qui utilise l'intersection entre les intervalles pour calculer leur distance.

– $S_{tmp-mac} \rightarrow$ Mesure de similarité de vecteurs de données : Pour comparer les trajectoires à l'échelle du contexte temporel, nous associons un vecteur de données temporelles à la trajectoire entière (p. ex. année : 2020, saison : automne, mois : 11, etc.) et nous comparons deux trajectoires sur leur vecteurs de données. Chaque éléments des vecteurs qui différent apportent une pénalité au score.

Enfin, troisièmement, pour les mesures de la dimension thématique (c.f. équation 4), nous utilisons les suivantes :

– $S_{thm-mic} \rightarrow$ MUITAS (May Petry *et al.*, 2019) : Pour comparer les trajectoires à l'échelle des données enrichissants les positions, nous utilisons la mesure MUITAS car elle permet de comparer des trajectoires multi-aspects en tenant compte de tous attributs de chaque aspect.

– $S_{thm-mes} \rightarrow$ EDR (L. Chen *et al.*, 2005) appliqué à des séries d'épisodes : Pour comparer les trajectoires à l'échelle des séquences d'épisodes sémantiques, nous utilisons l'attribut principal des aspects pour construire les séquences et nous exécutons EDR sur ces séquences. Lorsque nous travaillons avec des aspects lié à une ontologie, nous avons pour but d'utiliser la distance d'édition enrichie (Moreau *et al.*, 2018).

– $S_{thm-mac} \rightarrow$ LCSS (Vlachos *et al.*, 2002) : Pour comparer les trajectoires à l'échelle des thématiques générales, nous les résumons avec les valeurs majoritaires que prennent les attributs principaux de chaque type d'aspect et nous utilisons la mesure LCSS, qui est très adaptée pour comparer deux chaînes de caractères, pour comparer ces valeurs.

5. Expérimentation

Nous avons mis en place une expérimentation pour valider notre mesure. Elle a été conçue pour répondre à plusieurs questionnements :

1. Est-ce que les scores obtenus avec les mesures sont conformes à l'avis d'experts en géographie ?

2. Quels coefficients de pondération optimisent les résultats de la mesure ?

3. Dans le contexte de la mobilité touristique, est-il pertinent de considérer les trois dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) ?

4. Est-ce que notre hypothèse de départ (H1) est validée par cette expérimentation ?

5.1. Corpus de trajectoires

Le corpus utilisé pour cette expérimentation contient 23 paires de trajectoires journalières de touristes issues de la campagne de collecte Geoluciole. Ces paires représentent une variété de cas différents où les trajectoires peuvent être semblables

sur toutes, plusieurs, une ou aucune des dimensions présentées précédemment. Nous avons enrichi ces trajectoires avec des données de météo, de lever et de coucher de soleil issues d'OpenWeatherMap, des données concernant les points d'intérêt de La Rochelle issues de OpenStreetMap, des données concernant les quartiers, les espaces verts et les plages issues de l'Open Data de la Rochelle.

5.2. Protocole d'expérimentation

Le protocole de l'expérimentation est défini par les étapes suivantes : (1) collecter l'avis des experts sur la similarité ou la non-similarité de chaque paire de trajectoires de manière globale et selon chaque dimension ; (2) collecter les résultats issus de la mesure DA3T pour chaque paires de trajectoires de manière globale, selon chaque dimension ainsi que selon chaque niveau de granularité par dimension en ayant fixé les seuils (c.-à-d. valeur de la mesure au-delà de laquelle deux trajectoires sont considérées comme similaires) et coefficients ; (3) collecter les résultats issus des mesures DTW (Keogh, Ratanamahatana, 2005), de similarité de RI temporelle (Le Parc-Lacayrelle *et al*., 2007) et MUITAS (May Petry *et al*., 2019) correspondant respectivement aux mesures de référence spatiale, temporelle et thématique dans l'état de l'art ; (4) utiliser les métriques de précision (c.-à-d. nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par la mesure mais pas forcément par les experts), de rappel (c.-à-d. nombre de paires évaluées similaires par la mesure et par les experts rapporté au nombre de paires évaluées similaires par les experts mais par forcément par la mesure) et de F1-mesure (c.-à-d. mesure combinant la précision et le rappel) pour calculer la pertinence de mesure DA3T par rapport à l'avis des experts puis la comparer avec les mesures de l'état de l'art ; (5) réitérer les étapes (2) et (4) avec des seuils et coefficients différents afin d'optimiser ces valeurs. Présentons maintenant les résultats de l'expérimentation.

5.3. Résultats et discussion

Dans un premier temps, nous décrivons notre travail relatif à l'optimisation des seuils et coefficients de pondération de la mesure DA3T. Dans un second temps, nous procédons à l'évaluation de notre mesure et commentons les résultats obtenus. Le tableau 2 présente les résultats issus de l'optimisation de la mesure DA3T faite consécutivement à la collecte des avis des experts. Il présente les scores micro, méso et macro en terme de rappel, précision et F1-mesure pour chaque dimension.
Concernant la dimension spatiale, nous observons de légères disparités dans les résultats relatifs aux niveaux de granularité micro, méso et macro. Le niveau micro donne une F1-mesure de 0.889, légèrement supérieure aux deux autres. Le score spatial global prend en compte les trois niveaux de granularité de manière équivalente et nous constatons une amélioration des résultats par rapport à ceux des niveaux de granularité pris individuellement : F1-mesure à 0.914. Pour ce jeu de données, nous pouvons en conclure que les experts utilisent tous les grains dans leur observation d'une trajectoire touristique dans la ville.

TABLEAU 2. *Résultats issus de l'exécution de la mesure DA3T*

Score	α	β	γ	Seuil	Précision	Rappel	F1-mesure
$S_{spt-mic}$	–	–	–	0.9	1	0.8	0.889
$S_{spt-mes}$	–	–	–	0.892	0.762	0865	0.865
$S_{spt-mac}$	–	–	–	0.024	0.813	0.813	0.813
S_{spt}	0.33	0.33	0.34	0.616	1	0.842	0.914
$S_{tmp-mic}$	–	–	–	0.15	1	0.762	0.865
$S_{tmp-mes}$	–	–	–	0.22	0.813	0.929	0.867
$S_{tmp-mac}$	–	–	–	0.34	0.938	0.789	0.857
S_{tmp}	0.2	0.6	0.2	0.4	0.813	1	0.897
$S_{thm-mic}$	–	–	–	0.26	0.538	0778	0.636
$S_{thm-mes}$	–	–	–	0.05	1	0.565	0.722
$S_{thm-mac}$	–	–	–	0.2	1	0.867	0.929
S_{thm}	0.2	0.1	0.7	0.275	0.923	0.857	0.889
S_{glb}	0.4	0.2	0.4	0.4	1	0.857	0.923

Concernant la dimension temporelle, nous observons que, parmi les trois niveaux de granularité micro, méso, macro, aucune ne se démarque des autres, toutes ont une F1-mesure autour de 0,86. Nous constatons une amélioration de cette métrique dans le score temporel global : F1-mesure à 0.897. Par conséquent, ici également, il est intéressant de considérer les trois niveaux de granularité pour calculer la similarité temporelle de deux trajectoires sémantiques.
Concernant la dimension thématique, nous observons de fortes disparités dans les résultats relatifs aux niveaux de granularité micro, méso et macro. Le niveau macro donne une F1-mesure de 0.929 nettement supérieure aux deux autres. Tout d'abord, nous en concluons que les experts privilégient les aspects dominants de chaque thématique (p. ex. météo globalement ensoleillée, visite centrée autour du quartier du port, activité de restauration prédominante, etc.). Notons également ici que le score thématique global obtenu n'améliore pas le score thématique macro : F1-mesure 0.889 et de 0.929 respectivement. Dans ce cas particulier, nous préconisons une pondération des coefficients micro, méso et macro à 0, 0 et 1 respectivement.
Enfin, les résultats obtenus avec la mesure DA3T globale sont supérieurs à ceux obtenus avec les mesures dimensionnelles considérées séparément, ce qui nous permet d'affirmer que les experts utilisent toutes les dimensions des trajectoires sémantiques pour les comparer. Nous constatons, cependant, que le coefficient de pondération attribué à la dimension temporelle et optimisant les résultats est un peu plus faible que ceux des dimensions spatiale et thématique, ce qui laisse à penser que les experts s'intéressent un peu moins à la dimension temporelle lorsqu'ils comparent deux trajectoires touristiques.

Passons maintenant à l'évaluation de notre mesure par rapport à des mesures de référence existantes dans les différentes dimensions et globalement. Les mesures de référence choisies sont : DTW présentée dans Keogh, Ratanamahatana (2005) pour la dimension spatiale, la mesure de similarité de RI temporelle présentée dans Le Parc-

TABLEAU 3. *Comparaison de la mesure DA3T avec des mesures de référence grâce à la F1-mesure*

Dimension	Mesure de référence			Mesure DA3T
	DTW	RI temp.	MUITAS	
Spatiale	0.889			0.914
Temporelle		0.867		0.897
Thématique			0.636	0.889
Combinées	0.857			0.923

Lacayrelle *et al*. (2007) pour la dimension temporelle et enfin la mesure MUITAS présentée dans May Petry *et al*. (2019) pour la dimension thématique. Concernant la mesure de référence combinant les trois dimensions, nous avons utilisé la moyenne des trois mesures précédemment citées. Le tableau 3 montre que la mesure DA3T donne des résultats s'approchant plus de l'avis des experts que des mesures de référence choisies, et ce, dans toutes les dimensions et globalement.
Pour conclure, nous reprenons les quatre questions annoncées en début de section. Premièrement, pour chaque dimension, nous obtenons une F1-mesure autour de 0.90, ce qui nous permet de dire que notre mesure donne des résultats relativement proches de l'avis des experts. Deuxièmement, nous avons optimiser les coefficients de pondération par grain et par dimension. Notons que, selon la dimension, les grains privilégiés sont différents. Troisièmement, les résultats de la mesure globale montrent qu'il est très pertinent de considérer les trois dimensions des trajectoires sémantiques pour les comparer. Enfin, quatrièmement, notre hypothèse de départ, qui, rappelons le, propose d'observer chaque dimension selon trois niveaux de granularité, est validée. En effet, les dimensions spatiale et temporelle montrent une amélioration du résultat global par rapport aux résultats granulaires pris individuellement. D'autres part, le système de pondération permet d'éviter d'éventuelles pertes de performance.

6. Conclusion

Cet article a permis de présenter une nouvelle fonction de calcul de similarité entre trajectoires sémantiques dans un cadre d'activités touristiques. Cette fonction prend en compte toutes les dimensions des trajectoires sémantiques (c.-à-d. spatiale, temporelle et thématique) et ce sur trois niveaux de granularité (c.-à-d. micro, méso et macro) offrant la possibilité d'analyses plus ou moins fines et précises selon les conditions et besoins exprimés. Chaque composante de notre mesure a un coefficient de pondération qui permet de paramétrer son importance dans le calcul final du score. Afin de valider l'efficacité de notre mesure, nous avons mis en place une expérimentation basée sur l'avis des experts qui s'est révélée concluante.
Dans de prochains travaux, nous souhaitons enrichir notre mesure en prenant en compte les aspects multi-dimensionnels des trajectoires (p. ex. la dimension spatio-temporelle pour tenir compte de la vitesse de déplacement). Nous souhaitons également mener une expérimentation de plus grande envergure avec plus d'experts et plus de couples

de trajectoires à évaluer. Actuellement générique, cette mesure peut également être améliorée afin de l'adapter à des contextes spécifiques (c.-à-d. déplacements humains touristiques ou professionnels, animaliers, etc.) et ainsi être pondérée de façon automatique selon ces contextes d'usage.

Remerciements

Cet article a été écrit dans le cadre du projet DA3T, financées par la région Nouvelle-Aquitaine et la société Berger-Levrault.

Bibliographie

Aiello M. (2002, janvier). A spatial similarity measure based on games: Theory and practice. *Logic Journal of IGPL*, vol. 10.

Allen J. F. (1983, novembre). Maintaining knowledge about temporal intervals. *Communications of the ACM*, vol. 26, n° 11, p. 832–843.

Alt H. (2009, septembre). The Computational Geometry of Comparing Shapes. In *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, p. 235–248. Berlin, Heidelberg, Springer-Verlag.

Andrew J. (1805). Astronomical and nautical tables.

Chen L., Ng R. (2004, janvier). On The Marriage of Lp-norms and Edit Distance. In, p. 792–803.

Chen L., Özsu M. T., Oria V. (2005, janvier). Robust and fast similarity search for moving object trajectories. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, p. 491–502.

Chen Y., Nascimento M. A., Ooi B. C., Tung A. K. H. (2007, avril). SpADe: On Shape-based Pattern Detection in Streaming Time Series. In *2007 IEEE 23rd International Conference on Data Engineering*, p. 786–795.

Cleasby I. R., Wakefield E. D., Morrissey B. J., Bodey T. W., Votier S. C., Bearhop S. *et al.* (2019, novembre). Using time-series similarity measures to compare animal movement trajectories in ecology. *Behavioral Ecology and Sociobiology*, vol. 73, n° 11, p. 151.

Egenhofer M. J. (1997, août). Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages & Computing*, vol. 8, n° 4, p. 403–424.

Faloutsos C., Ranganathan M., Manolopoulos Y. (1994, mai). Fast subsequence matching in time-series databases. *ACM SIGMOD Record*, vol. 23, n° 2, p. 419–429.

Keogh E., Ratanamahatana C. A. (2005, mars). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, vol. 7, n° 3, p. 358–386.

Lee J.-G., Han J., Whang K.-Y. (2007, juin). Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, p. 593–604. New York, NY, USA, Association for Computing Machinery.

Le Parc-Lacayrelle A., Gaio M., Sallaberry C. (2007). La composante temps dans l'information géographique textuelle. *Document Numérique*, vol. 10, n° 2, p. 129–148. (Publisher: Lavoisier)

Lu E. H.-C., Tseng V. S. (2009, mai). Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, p. 273–278.

Magdy N., Sakr M., Abdelkader T., Elbahnasy K. (2015, décembre). Review on trajectory similarity measures.

May Petry L., Ferrero C., Alvares L., Renso C., Bogorny V. (2019, juin). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, vol. 23.

Mello R. D., Bogorny V., Alvares L. O., Santana L. H. Z., Ferrero C. A., Frozza A. A. *et al*. (2019, mai). MASTER: A multiple aspect view on trajectories. *Transactions in GIS*, p. tgis.12526.

Moreau C., Devogele T., Etienne L. (2018, novembre). Extraction de motifs de trajectoires sémantiques similaires. In *Spatial Analysis and Geomatics*. Montpellier, France.

Nakamura T., Taki K., Nomiya H., Seki K., Uehara K. (2013, novembre). A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, vol. 16, n° 4, p. 535–548.

Parent C., Spaccapietra S., Renso C., Andrienko G. L., Andrienko N. V., Bogorny V. *et al*. (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, vol. 45, n° 4, p. 42:1–42:32.

Sallaberry C. (2013). *Geographical Information Retrieval in Textual Corpora*. Wiley-ISTE.

Su H., Liu S., Zheng B., Zhou X., Zheng K. (2020, janvier). A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, vol. 29, n° 1, p. 3–32.

Tao Y., Both A., Silveira R. I., Buchin K., Sijben S., Purves R. S. *et al*. (2021, juillet). A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing*, vol. 58, n° 5, p. 643–669.

Varlamis I., Sardianos C., Bogorny V., Alvares L. O., Carvalho J. T., Renso C. *et al*. (2021, mars). A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. New York, NY, USA, Association for Computing Machinery.

Vlachos M., Kollios G., Gunopulos D. (2002). Discovering similar multidimensional trajectories. In *Proceedings 18th International Conference on Data Engineering*, p. 673–684. San Jose, CA, USA, IEEE Comput. Soc.

Wang H., Su H., Zheng K., Sadiq S., Zhou X. (2013, janvier). An effectiveness study on trajectory similarity measures. In, p. 13–22.