Apprentissage automatique pour l'analyse de trajectoires spatiales : extraction conjointe de caractéristiques démographiques et comportementales

Hippolyte Dubois

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ
ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique de l'Information et de la Communication
Spécialité : *INFO*

Par
Hippolyte DUBOIS

Apprentissage automatique pour l'analyse de trajectoires spatiales

Extraction conjointe de caractéristiques démographiques et comportementales

En vue de la soutenance de Thèse à Polytech Nantes, le 10/03/2023
Unité de recherche : LS2N - UMR CNRS 6004

Rapporteurs avant soutenance :

Christophe Claramunt	Professeur des Universités, École Navale
Jonathan Weber	Maitre de conférences, Université de Haute-Alsace

Composition du Jury :

Présidente :	Luce Morin	Professeure des Universités, INSA Rennes
Examinateurs :	Valérie Gyselinck	Directrice de recherche, Université Gustave Eiffel
	Frédéric Precioso	Professeur des Universités, Université Côte d'Azur - Sophia Antipolis
	Giuseppe Valenzise	Chargé de recherche, CNRS - Centrale SupElec - Université Paris Saclay
Dir. de thèse :	Patrick Le Callet	Professeur des Universités, Nantes Université
Co-dir. de thèse :	Antoine Coutrot	Chargé de recherche, CNRS - INSA - Université de Lyon

Remerciements

Je tiens à remercier Antoine, pour m'avoir donné la chance de prendre part à ce projet, et pour m'avoir conseillé et accompagné tout du long. Patrick également, pour son énergie et son implication. Giuseppe et Frédéric également, pour avoir su me comprendre et me guider.
J'adresse également un grand merci à l'équipe IPI, et tout particulièrement à ses doctorants, sa terrasse, sa machine à café, et tous les moments que cet environnement a permis, sans qui cette thèse n'aurait surement jamais aboutie. Merci également à Sophie et Neslihan, pour leur travail qui fait tenir ces murs.
Enfin, merci à ma famille, mes amis, ma colocataire.

Sommaire

TABLE DES FIGURES

Liste des tableaux

Organisation de la thèse

Cette thèse est organisée en sept chapitres (figure 1) :

- Le premier introduit le contexte dans lequel elle s'opère, en présentant les motivations à l'analyse de trajectoires et l'importance de la compréhension des mécanismes sous-jacents au sens de l'orientation. Il nous permet de lier cette thèse en informatique aux autres disciplines scientifiques qui s'intéressent à ce sujet.
- Le second dresse un portrait de l'état de l'art en se concentrant sur les questions de mesure de trajectoire et de leur groupement. Il nous permet de nous recentrer sur le sujet de cette thèse, à savoir le développement de nouvelles méthodes d'analyses de données.
- Les chapitres 3, 4, 5 et 6 présentent les contributions. Le chapitre 3 détaille les prétraitements utilisés pour pouvoir traiter les données. Le quatrième introduit le concept de réseau de neurones à signal composites, et en propose une implémentation. Le cinquième présente une nouvelle métrique de trajectoire tirant profit de l'utilisation de données massives pour apprendre à identifier ce qui fait une trajectoire normale. Enfin, le chapitre 6 propose une méthode de groupement conjoint pour données hétérogènes, en utilisant une architecture de modèle à mélange de distributions. Dans chacun de ces chapitres, la solution proposée est évaluée sur le même ensemble de données et une comparaison entre ces différentes approches est faite au fur et à mesure.
- Pour conclure, le chapitre 7 revient sur l'ensemble des contributions en les remettant en perspective par rapport aux problèmes énoncés dans les premiers chapitres. Il présente ensuite un ensemble de perspectives pour pousser plus loin les idées ébauchées dans cette thèse.

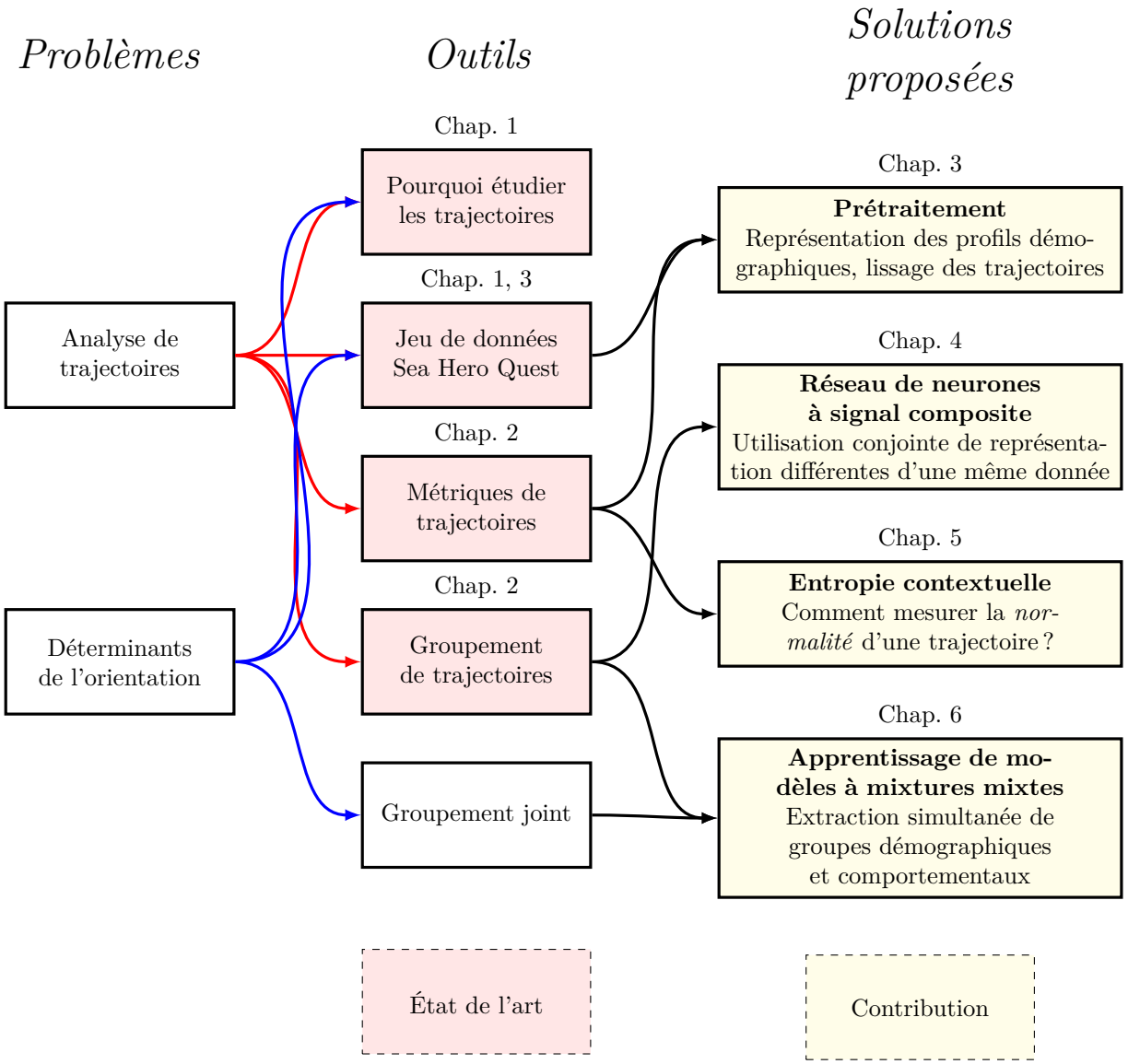


FIGURE 1 – Structure de la thèse. Nous partons de deux problèmes, et, à partir des outils présents dans la littérature, nous développons plusieurs solutions pour essayer d'y répondre. Une flèche traduit ce lien : par exemple, pour analyser des trajectoires, l'état de l'art propose un ensemble de métriques pour mesurer les caractéristiques d'une trajectoire, nous proposons donc d'utiliser l'entropie contextuelle pour mesurer sa normalité.

Disponibilité du code

Les contributions proposées dans ce manuscript sont regroupées sous la forme d'une boîte à outil Python disponible sur Github au lien suivant : SHQ-NN-Toolbox.

Chapitre 1

Introduction

1.1 Pourquoi étudier les trajectoires

La notion de comportement spatial est transverse à de nombreuses disciplines. Par exemple, M. Bonnes et G. Carrus[1] définissent :

> Le concept de comportement spatial concerne la manière dont les individus régulent et utilisent (en termes d'appropriation et de défense) leurs environnements spatiaux à différents niveaux : personnel, interpersonnel et collectif. La psychologie environnementale qui s'intéresse au comportement spatial se concentre sur le rôle des propriétés spatiales de l'environnement dans le façonnement et la régulation de l'interaction sociale dans les situations quotidiennes.

D. Evans et D. Herbert[2] la définissent ainsi :

> La géographie comportementale se concentre sur la prise de décision spatiale des individus avant d'agir et tient compte à la fois de la distance et de la direction. La recherche s'est concentrée à la fois sur le comportement spatial et le comportement dans l'espace. Elle se concentre donc sur le comportement et la prise de décision des individus.

K. Stanley, chercheur en informatique, en donne la définition suivante[3] :

> Le comportement spatial est un paramètre fondamental qui sous-tend de nombreux phénomènes anthropologiques, sociologiques et géographiques. L'endroit où se trouvent les gens, la façon dont ils utilisent l'espace, la façon dont l'endroit façonne d'autres contextes et la façon dont le contexte spatial façonne la perception.

Lorsque nous parlerons de comportement spatial, c'est dans le cadre de cette dernière définition que nous le faisons. Plus simplement, on peut résumer la notion de comportement spatial telle qu'utilisée dans cette thèse comme ceci : comment une personne va se déplacer et interagir avec son environnement.

La compréhension et modélisation du comportement spatial, puisqu'il est influencé par de nombreux facteurs, est importante dans de nombreux champs de la recherche, tels

que l'urbanisme[4], la prédiction du trafic routier[5], l'écologie[6], l'étude des foules[7], ou encore pour les neurosciences comportementales et cliniques[8].

Dans tous ces domaines d'application, on peut utiliser la trajectoire comme représentation du comportement spatial, et espérer, en analysant la trajectoire, en apprendre un peu plus sur les mécanismes sous-jacents.

D'une manière générale, la trajectoire est traitée comme une *trace*, un *signal* s produit par un *agent* a dans un *environnement* e, collectée par un *observeur* g. Cette trace porte donc potentiellement en elle une partie de l'information sur cet agent et cet environnement qui l'ont produite :

$$s = g(a, e)$$

L'apparition et le développement de technologies de collecte de données de localisation ont permis la constitution de jeux de données de trajectoires sur un vaste nombre d'applications, qui font de l'analyse de trajectoires un enjeu de la recherche contemporaine dans de nombreux domaines.

1.1.1 Étudier la ville : exemple des taxis

Parmi les nombreux jeux de données de trajectoires, les taxis sont une source récurrente[9-15]. Ils fournissent une information sur la ville, ses usages et son trafic routier.

FIGURE 1.1 – Visualisation de trajectoires de taxis dans la ville de Chengdu, Chine[16]. Chaque couleur représente un type de sous-trajectoire extrait par le modèle.

On peut identifier trois catégories d'analyses effectuées sur ces jeux de données[17] :

– Dynamiques sociales : identification des nœuds de transports, des motifs de déplacement, des liens entre régions et quartiers

- Trafic routier : détection des bouchons, des trajets alternatifs, des pics de pollution
- Dynamiques opérationnelles : classement des chauffeurs, analyse des stratégies de collecte de passagers

On voit bien à travers cet exemple comment une donnée d'apparence simple peut permettre d'analyser des dynamiques et processus complexes à modéliser dans leur globalité.

1.2 Comprendre le sens de l'orientation

Dans le cadre de l'étude des comportements humains et des fonctions cognitives, l'étude du sens de l'orientation bénéficie particulièrement de l'utilisation de données de trajectoires.

1.2.1 Processus cognitifs de l'orientation

Le sens de l'orientation est une compétence définie comme la capacité d'un individu à définir un chemin et à naviguer d'un point d'origine à une destination qui n'est pas visible[18]. Elle se décompose en quatre étapes :

Localisation l'individu essaie de déterminer sa position relative dans son environnement,

Planification il définit une stratégie, un chemin, pour se rendre à sa destination,

Actualisation pendant qu'il exécute sa stratégie, il s'assure qu'elle est correcte, en actualisant sa perception de sa position,

Identification une fois arrivé à destination, il la reconnait.

C'est un processus multi-échelle : quand il doit effectuer un déplacement complexe, l'individu va modulariser l'environnement, et répéter ces quatre étapes en continu, en définissant des destinations intermédiaires à partir de sa connaissance de l'environnement et des repères qu'il a précédemment identifiés.

Ce processus cognitif est opéré principalement par l'activation de neurones spécialisés localisés dans l'hippocampe appelés *cellules de lieu*, qui composent une carte cognitive[19].

Le développement de cette carte cognitive est le fruit de l'interaction entre l'individu, son environnement, et son cadre social[20, 21], on peut donc considérer le processus d'orientation comme source d'information pour étudier l'individu, l'environnement, la société, et les liens entre eux.

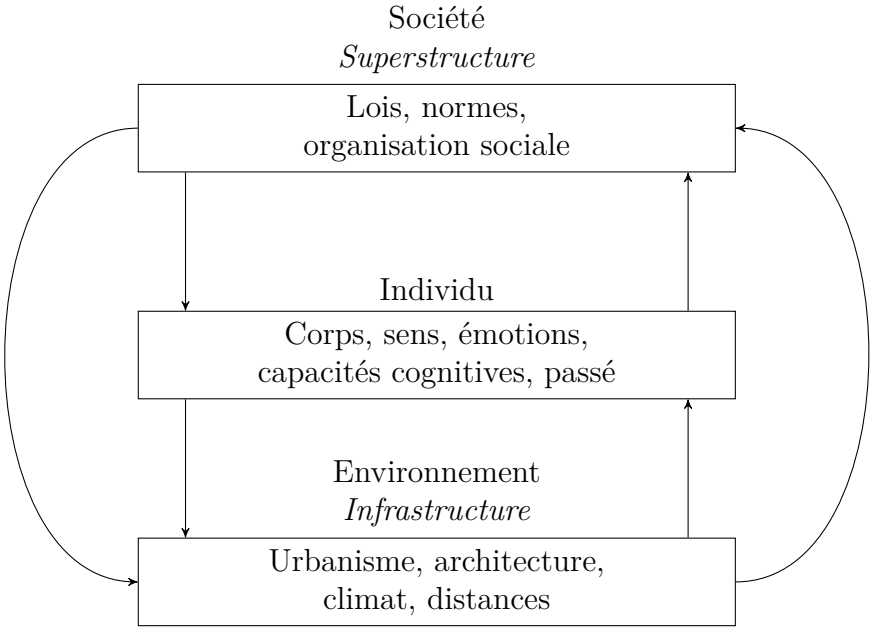


FIGURE 1.2 – Modèle des facteurs influençant le développement du processus cognitif d'orientation. Les flèches représentent l'influence qu'une structure a sur une autre.

1.2.2 Facteurs démographiques de l'orientation

L'orientation étant un processus cognitif complexe, influencé par de nombreux facteurs génétiques et environnementaux, elle présente une forte variabilité au sein de la population générale. La compréhension de ces facteurs et de leur influence est importante pour permettre de mieux penser les environnements[22], et mesurer l'impact de structures sociales sur les individus jusque dans leur cognition.

Impact du vieillissement sur l'orientation

Puisque l'orientation est une fonction cognitive gérée principalement par l'hippocampe, elle est impactée par les afflictions qui touchent cette zone du cerveau. Il a été montré que le vieillissement touche particulièrement l'hippocampe[23], et donc par extension les capacités d'orientation[24]. Mais l'on sait aussi que ce vieillissement n'est pas un processus sur lequel nous ne pouvons pas agir, et parmi les méthodes explorées, la stimulation des fonctions cognitives d'orientation joue un rôle significatif[25, 26].

Si l'on sait que l'age impacte l'orientation, il est encore difficile de caractériser cet impact. Le mesurer permettrait, par exemple, de mieux penser l'organisation des services de soin et des bâtiments destinés à accueillir cette partie de la population.

L'orientation, un marqueur des discriminations liées au genre

S'il est connu depuis longtemps qu'on peut attribuer une partie des différences dans les capacités d'orientation au sexe[27], il a récemment été identifié que ces différences seraient probablement plutôt dûes aux différences de vécu entre hommes et femmes[28], et que l'indice d'écart entre les genres[29] est corrélé à l'écart moyen entre les genres de performance sur une tâche d'orientation[30]. Cela suggère que cette différence serait plus acquise qu'innée, et que l'influence du genre sur l'expérience d'un individu impacterait le développement de ses fonctions cognitives.

Autres facteurs environnementaux

Si les travaux en neurosciences sur l'orientation ont trouvé un écho particulier dans la recherche en urbanisme et architecture[20], c'est parce que la conception d'environnements dans lesquels des individus vont être amenés à se déplacer nécessite une compréhension des mécanismes cognitifs qui régissent cette fonction. Des travaux récents ont mis en lumière l'impact que la ville et son organisation ont sur le développement de nos capacités d'orientation[31, 32].

De la même manière, les populations des îles de Polynésie ont développé des méthodes d'orientation et de navigation en mer spécifiques, transmises par voie orale, qui leur ont permis d'explorer les îles d'une zone de plus de 2 millions km^2, réalisant des voyages de plusieurs milliers de kilomètres il y'a plus de 3000 ans[33].

Le PIB d'un pays, ou la présence de course d'orientation au programme des cours de sport sont également corrélés à la performance moyenne sur des tâches d'orientation[30].

1.2.3 Impacts de la démence

Comme nous l'avons dit plus tôt, l'orientation est impactée par les afflictions de l'hippocampe. À ce titre, la dégradation des capacités d'orientation peut être un marqueur des maladies neuro-dégénératives touchant cette zone du cerveau.

Contrairement aux autres symptômes de la maladie d'Alzheimer, notamment les pertes de mémoire qui sont le facteur de diagnostic le plus utilisé aujourd'hui, la désorientation dans un espace connu y est plus spécifique, et permet notamment de la différencier de la dégénérescence fronto-temporale (voir la revue de littérature de Coughlan et ses collègues parue en 2018[8]).

1.3 La maladie d'Alzheimer

La maladie d'Alzheimer (AD) est une maladie neuro-dégénérative entrainant un trouble neurologique progressif, aussi appelée *démence*. Cette maladie, pour laquelle il n'existe aujourd'hui aucun traitement curatif et dont les mécanismes sont encore mal connus, touche approximativement 3.9% de la population mondiale âgée de plus de 60 ans[34], avec un très fort impact de l'age (voir figure 1.3).

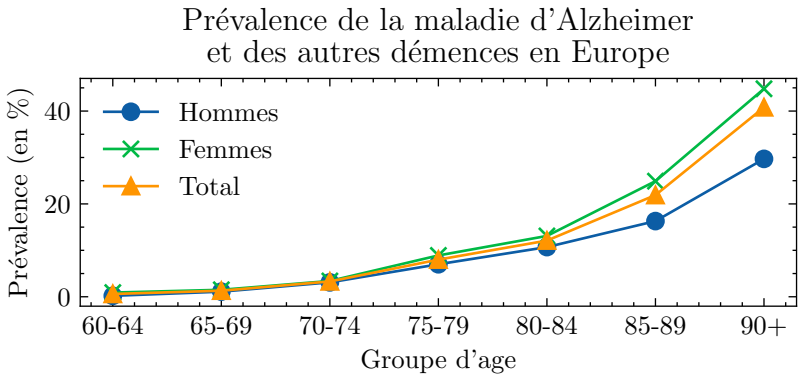


FIGURE 1.3 – Prévalence de la démence en Europe selon [35]

1.3.1 Impact social de la maladie d'Alzheimer

La démence causée par la maladie d'Alzheimer impacte globalement les fonctions cognitives du sujet, et se manifeste principalement par des difficultés de mémorisation, des troubles du langage, et des changements comportementaux (désinhibition, etc.). Ces symptômes apparaissent généralement tardivement dans la maladie, et ont un impact sur la vie du sujet, qui perd en autonomie, et sur celles de ses proches. Mais la maladie a aussi un impact sur la santé physique des sujets, en augmentant le risque de maladies cardio-vasculaires et de troubles musculo-squelettiques, ce qui occasionne également une perte d'autonomie physique[36].

Du fait de cette perte d'autonomie, la maladie d'Alzheimer a un impact social et économique majeur. En 2005, il était estimé que le coût global occasionné par les maladies de démence approchait les 300 milliards d'euros[37]. L'espérance de vie augmentant dans le monde, tandis que l'espérance de vie en bonne santé stagne, on peut imaginer que ce coût va augmenter, la part de personnes en perte d'autonomie augmentant.

Cependant, si on ne sait pas guérir la maladie d'Alzheimer, on peut en ralentir la progression, en agissant sur les facteurs physiologiques et psycho-sociaux connus[38, 39].

1.3.2 Diagnostic de la maladie d'Alzheimer

Le diagnostic de la maladie d'Alzheimer passe aujourd'hui avant tout par une analyse des symptômes perçus par le patient, ou identifiés par son entourage. Mais la superposition des symptômes avec d'autres maladies et les effets secondaires de traitements fréquemment prescrits chez les personnes vieillissantes rend difficile un diagnostic définitif de la maladie d'Alzheimer, qui peut se faire en réalisant un scanner du cerveau (CT, MRI ou PET)[40], pour identifier visuellement l'impact de la dégénérescence des tissus cérébraux. Cette procédure est lourde et coûteuse et donc peu adaptée à un dépistage à grande échelle.

Le fait de devoir attendre l'apparition de symptômes suffisamment impactants pour qu'ils soient identifiés par le patient et son entourage limitent l'efficacité des approches de réduction des risques puisqu'elles interviennent nécessairement tardivement dans le développement de la maladie.

1.4 Le projet Sea Hero Quest

C'est ici qu'intervient le projet Sea Hero Quest (SHQ), né de la collaboration de l'Alzheimer's Research Trust avec l'université d'East Anglia, University College de Londres, et Deutsche Telekom. En analysant l'impact précoce de la maladie sur les capacités d'orientations et les comportements spatiaux, on espère pouvoir mettre au point un outil de diagnostic précoce de la maladie, qui permettrait une mise en place précoce des dispositifs de prévention de la perte d'autonomie et la préparation pour le sujet, son entourage, et la société au moment où elle arrivera et où il faudra donc la prendre en charge.

Une meilleure compréhension de la façon dont la démence impacte les fonctions d'orientation permettrait aussi d'ouvrir des perspectives sur une meilleure adaptation des environnements aux personnes en perte d'autonomie. En identifiant les situations qui entraînent une perte de repère, qui sont des situations angoissantes pour le sujet, on pourrait mieux concevoir les espaces destinés à les accueillir pour justement éviter ces situations d'angoisse.

1.4.1 Mesurer l'orientation

Pour mesurer la capacité d'orientation, nous nous sommes tournés vers l'analyse de trajectoire.

Une trajectoire est la trace du processus cognitif d'un individu confronté à une tâche d'orientation. C'est un signal riche et facile à collecter. Ces qualités ont justifié son utilisation.

Collecter des données

Pour pouvoir mesurer les capacités d'orientation à large échelle, il faut une grande quantité de données de trajectoire, issue d'une population la plus représentative possible de la population générale.

Pour constituer un jeu de données, un jeu vidéo, *Sea Hero Quest* (SHQ), a été développé par le studio Glitchers. Dans ce jeu sur téléphone et tablette, le joueur est amené à naviguer d'un point à un autre dans un labyrinthe d'îles, en conduisant un petit bateau en vue subjective (voir figure 1.4).

FIGURE 1.4 – Captures d'écran et cartes de niveaux tirées du jeu Sea Hero Quest

L'utilisation du format vidéoludique a permis de réaliser cette collecte de données à très large échelle. Les trajectoires des joueurs dans les niveaux étaient collectées et stockées, pour constituer la première partie du jeu de données SHQ.

Pour pouvoir évaluer l'influence des facteurs démographiques sur l'orientation, il était demandé aux joueurs qui le voulaient de renseigner plusieurs informations les concernant : leur pays, leur âge, leur genre, s'ils sont droitiers ou gauchers, dans quel type d'environnement ils habitent (urbain, rural, etc.), leur niveau scolaire, combien d'heures ils dorment par nuit en moyenne, le temps qu'ils passent quotidiennement dans les transports en moyenne, et une auto-évaluation de leur sens de l'orientation. Ces données forment la deuxième partie du jeu de données SHQ.

En tout, plus de quatre millions de personnes du monde entier ont joué à ce jeu et fourni des informations les concernant.

Méthodes d'analyses

Il est possible d'utiliser les données de Sea Hero Quest pour mesurer les capacités d'orientation, puisqu'il a été montré que l'orientation dans un espace virtuel est prédictive des compétences dans un environnement réel[41]. Jusqu'ici, ces données ont été analysées au moyen d'outils "simples". Pour mesurer la performance d'un joueur, la durée de ses essais et la longueur de ses trajectoires ont été utilisées comme métrique[30, 31, 42, 43].

L'une de ces métrique est le score *OpCorr*, calculé pour un joueur à partir de la longueur de ses trajectoires sur un sous ensemble de niveaux. Ce score a été pensé pour prendre en compte le biais joueur vs non-joueur dans l'approche du jeu, en normalisant la longueur des trajectoires des niveaux utilisés pour le calculer par la longueur des trajectoires du joueur sur les deux niveaux d'entrainement.

En résumé :

✓ L'analyse de trajectoires nous permet d'étudier aussi bien l'environnement que l'individu.
✓ Le sens de l'orientation est le produit d'une multitude de facteurs qui interagissent entre eux, il peut donc être un bon indicateur de l'état cognitif d'une personne.
✓ Mais, pour pouvoir utiliser le sens de l'orientation comme outil d'analyse clinique, il faut d'abord en analyser les tendances au sein de la population générale.

Chapitre 2

Analyse de trajectoires : Applications et méthodes

2.1 Applications

Comme on l'a vu plus tôt, l'analyse de données de trajectoires se généralise à de plus en plus de disciplines, ce qui nécessite le développement de techniques de traitement appropriées.

On retrouve dans la littérature une variété de problèmes applicatifs liés à l'analyse de trajectoire, que l'on peut réunir en trois classes, comme montrées sur la figure 2.1.

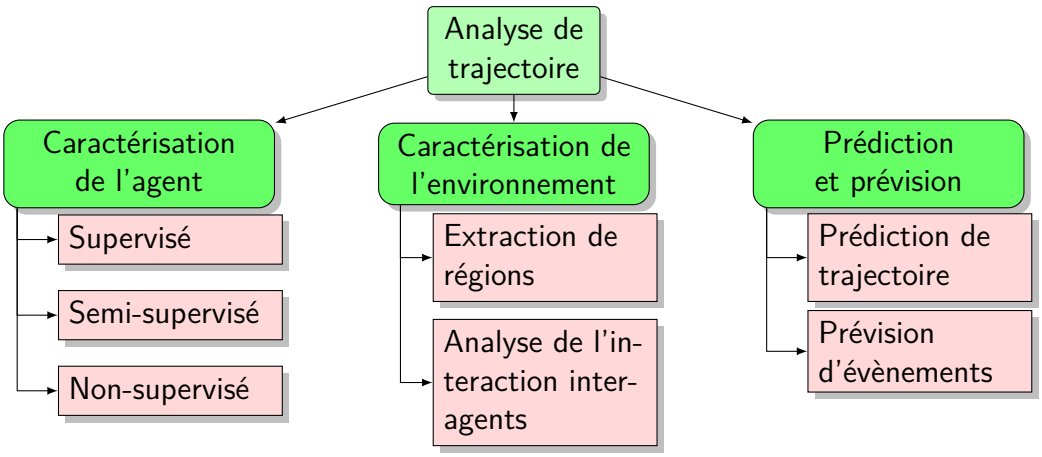


Figure 2.1 – Typologie des problèmes d'analyse de trajectoires

2.1.1 Caractérisation de l'agent

Apprentissage supervisé

Ici, on essaye de caractériser l'agent à partir de sa trace. Dans les faits, on ne peut pas représenter de manière complète un agent a, on en propose donc une représentation

partielle $\hat{a}$, qui est un ensemble de caractéristiques mesurées sur a. On va donc chercher à trouver une fonction f tel que $f(s) = \hat{a}$.

On part ici du principe que toutes les caractéristiques de la représentation $\hat{a}$ influent sur la trace s. Il est donc important de sélectionner un ensemble de caractéristiques jugées pertinentes à-priori.

Par exemple, si l'on s'intéresse au comportement d'oiseaux, on peut essayer de prédire à partir d'une trajectoire si l'oiseau l'ayant générée était en train de voler, marcher, ou s'il ne bougeait pas[44]. Essayer de prédire la couleur de son bec n'aurait ici pas fait sens, et il aurait été compliqué d'évaluer la pertinence d'un modèle pour réaliser cette tâche.

Apprentissage semi ou non-supervisé

Dans certains cas, on ne peut pas mesurer de caractéristiques sur a, ou alors on n'en a pas assez, ou pas d'assez bonne qualité sans trop de connaissance à-priori sur leur relation à s.

Dans ce genre de problèmes, on va chercher à caractériser la trajectoire s par une représentation z, et utiliser cette représentation comme proxy pour représenter a.

2.1.2 Caractérisation de l'environnement

Ici on va non pas s'intéresser aux agents individuellement, mais utiliser l'ensemble de leurs traces conjointement pour analyser l'environnement commun dans lequel ils ont défini ces trajectoires.

Extraction de régions

Il s'agit d'identifier des *régions d'intérêt* (ROIs). Il peut s'agir de dresser une typologie de lieux[13, 14], ou de trouver la localisation de points spécifiques[12, 45]. Les modèles sémantiques se prètent particulièrement à ce type de tâches[46-48].

Analyse de l'interaction inter-agents

On s'intéresse ici aux interactions inter-agents. Cette approche nécessite que les données collectées pour plusieurs utilisateurs l'aient été simultanément, et que les utilisateurs aient pu interagir.

2.1.3 Prédiction et prévision

Il existe dans la littérature de nombreux exemples d'algorithmes de prévision (*forecasting* en anglais) sur des données de trajectoires. On ne détaillera pas ici ces exemples, mais les mentionnons quand même dans notre typologie par soucis d'exhaustivité.

2.2 Techniques et méthodes

On retrouve dans l'état de l'art deux types d'approches à la résolution des problèmes impliquant des trajectoires.

La première, que l'on appellera *partielle*, cherche à produire à partir d'une trajectoire une représentation pouvant être traitée par un algorithme d'apprentissage classique. Par exemple, si on veut effectuer de la classification de trajectoire, on peut utiliser un classifieur linéaire h, qui nécessite que les données qui lui sont présentées soient dans un espace réel de dimension d :

$$f(x) = h(g(x)) \tag{2.1}$$

$$g : \mathcal{T} \mapsto \mathbb{R}^d \tag{2.2}$$

La seconde, que l'on appellera *complète*, ne se base cette fois plus sur un algorithme classique, mais propose un algorithme spécifique, adapté pour les données de trajectoires.

Les trajectoires étant des séries temporelles particulières, on peut y appliquer des approches pensées pour les séries temporelles.

2.2.1 Cas général des séries temporelles

Une série temporelle est un signal composé d'une série d'observations ordonnées s_t effectuées dans un ensemble $\mathbb{S}$. t est ici l'indice temporel. Sauf précision du contraire, on considèrera que les observations sont effectuées à intervalle régulier, on définit donc $t \in \mathbb{Z}$. Une série temporelle finie a une longueur l, c'est-à-dire le nombre d'observations qui la composent. On a donc $\{s_t \in \mathbb{S}\}_{t=0}^{l}$. Généralement, $\mathbb{S} = \mathbb{R}^d$.

Transformée ondelettes

La transformée ondelette est une généralisation de la transformée de Fourier. Une ondelette est une fonction ψ carré-intégrable. On définit une fonction ψ_{jk} paramétrisée par deux entiers $j, k \in \mathbb{Z}$:

$$\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$$

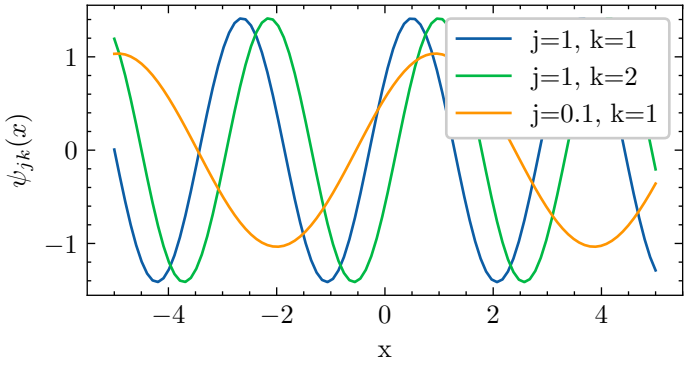


FIGURE 2.2 – Exemple d'ondelettes utilisant la fonction cosinus

On peut donc définir une famille d'ondelettes paramétrisées par différentes valeurs de j et k.

On considère qu'une série temporelle s est la somme des ondelettes de la famille :

$$s_t = \sum c_{jk}\psi_{jk}(t)$$

On peut donc représenter la série temporelle par les coefficients c_{jk} associés à chacune des ondelettes :

$$g(s) = c_{jk}$$

Ces coefficients peuvent être utilisés comme représentation pour une approche partielle[49-53]. Ces méthodes sont particulièrement adaptées aux signaux périodiques, ce qui est rarement le cas des trajectoires. De plus, il n'existe pas à notre connaissance d'exemples de transformées ondelettes multi-dimensionnelles : dans l'état de l'art, lorsque la série temporelle traitée est multidimensionnelle, une famille d'ondelettes est définie pour chacune des dimensions.

Modèles auto-encodeurs

Pour répondre à ce problème, une solution est d'utilisée des modèles à apprentissage profond avec une structure d'auto-encodeur[11, 54-56].

Un auto-encodeur est un réseau neuronal profond composé de deux blocs : un encodeur h, et un décodeur g. Le rôle de l'encodeur est de produire une représentation $z \in \mathbb{R}^d$ d'un signal x, qui sera ensuite utilisée par le décodeur qui essaye de reconstruire x à partir de z.

$$x' = g(h(x))$$

Si le modèle y arrive, on sait alors que la représentation z contient assez d'information sur x pour pouvoir le reconstruire. On peut donc espérer qu'elle sera pertinente du point de vue de notre problème. En introduisant une tâche prétexte (la reconstruction), on peut utiliser des méthodes d'apprentissage profond sur des problèmes non-supervisés comme le groupement, pour lesquels on n'a pas de labels de classe.

Un autre avantage de ces méthodes est qu'elles requièrent peu de connaissances à-priori sur les séries temporelles que l'on analyse. Contrairement aux approches descripteur classique[57], ici les descripteurs sont appris automatiquement à partir des données.

Il existe trois architectures de réseau de neurones qui peuvent être utilisées pour réaliser un auto-encodeur appliqué à des données de série temporelle :

MLP Le perceptron multicouche est un réseau de neurones à propagation avant composé d'un enchaînement de couches entièrement connectées. C'est la structure la plus simple, elle ne tient pas compte de la structure de la série-temporelle. Pour un signal $s \in \mathbb{R}^{d \times l}$, la première couche du réseau a $d \times l$ nœuds.

CNN Le réseau de neurones convolutionnel est un réseau de neurones à propagation avant composé de couches convolutives et de couches entièrement connexes en sortie. Une couche convolutive est une couche composée d'un ensemble de noyaux de convolutions appliqués par fenêtre glissante sur le signal en entrée. Cette structure permet de garder la structure temporelle du signal. Pour un signal $s \in \mathbb{R}^{d \times l}$, les convolutions de la première couche seront de la forme $f : \mathbb{R}^d \mapsto \mathbb{R}$.

RNN Le réseau de neurones récurrents est un réseau de neurones à propagation avant avec un mécanisme de mémorisation, qui permet de construire une représentation au fur et à mesure du parcours séquentiel du signal. Là aussi, il permet de garder la structure temporelle du signal.

Les structures d'auto-encodeur classique ne garantissent pas la structure de l'espace latent de représentation, ce qui peut poser problème si l'on souhaite utiliser les représentations apprises pour une tâche de groupement par exemple. Pour répondre à ce problème, on peut utiliser un auto-encodeur variationnel[56] : l'encodeur ne produit plus ici une variable d'encodage, mais les paramètres μ et Σ d'une distribution gaussienne dans l'espace latent. z, la représentation utilisée par le décodeur, est tirée aléatoirement depuis cette distribution.

Autres caractéristiques

Il existe un nombre important de mesures de caractérisation des séries temporelles dans l'état de l'art[58]. L'ensemble *hctsa*[59] par exemple propose plusieurs milliers de caractéristiques. Ce nombre conséquent rend difficile leur utilisation, puisqu'il devient computationnellement inefficient de les utiliser. Un sous ensemble de 22 caractéristiques a été proposé[57], obtenu en mesurant leur performance sur 93 tâches différentes et en gardant les meilleures non redondantes.

Nom de la caractéristique hctsa	Description
Distribution	
DN_HistogramMode_5	Mode de la distribution centrée-réduite (histogramme 5 classes)
DN_HistogramMode_10	Mode de la distribution centrée-réduite (histogramme 10 classes)
Statistiques temporelles simples	
SB_BinaryStats_mean_longstretch1	Plus longue période de valeurs successives au dessus de la moyenne
DN_OutlierInclude_p_001_mdrmd Time	Intervalle entre les évènements extrêmes successifs supérieurs à la moyenne
DN_OutlierInclude_n_001_mdrmd Time	Intervalle entre les évènements extrêmes successifs inférieurs à la moyenne
Autocorrélation linéaire	
CO_f1ecac	Premier $1/e$ crossing de la fonction d'autocorrélation
CO_FirstMin_ac	Premier minima de la fonction d'autocorrélation
SP_Summaries_welch_rect_area_5_1	Puissance du cinquième inférieur des fréquences de la densité spectrale de puissance de Fourier
SP_Summaries_welch_rect_centroid	Centroïde de la densité spectrale de puissance de Fourier
FC_LocalSimple_mean3_stderr	Erreur moyenne d'une prédiction par moyenne sur fenêtre glissante de 3 échantillons
Autocorrélation non-linéaire	
CO_trev_1_num	Mesure de la réversibilité temporelle, $\langle (x_{t+1} - x_t)^3 \rangle_t$
CO_HistogramAMI_even_2_5	Information automutuelle, $m = 2$, $\tau = 5$
IN_AutoMutualInfoStats_40_gaussian_fmmi	Premier minima de la fonction d'information automutuelle
Différenciations successives	
MD_hrv_classic_pnn40	Proportion des différenciations successives $> 0.04\sigma$
SB_BinaryStats_diff_longstretch0	Plus longue périodes de diminutions successives
SB_MotifThree_quantile_hh	Entropie de Shannon de deux signes successifs d'un alphabet de 3 signes équiprobables
FC_LocalSimple_mean1_tauresrat	Différence de la longueur de corrélation après différenciations successives
CO_Embed2_Dist_tau_d_expfit_meandiff	Exponential fit des distances successives dans un espace de plongement 2d
Analyse des fluctuations	
SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	Proportion de fluctuations à échelle plus lente avec DFA (50% sampling)
SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	Proportion de fluctuations à échelle plus lente avec ajustements linéaires d'intervalles.
Autres	
SB_TransitionMatrix_3ac_sumdiagcov	Trace de la matrice de covariance de transition entre les symboles d'un alphabet de 3 signes
PD_PeriodicityWang_th0_01	Mesure de périodicité

TABLE 2.1 – Les 22 caractéristiques de l'ensemble *catch22*

2.2.2 Groupement de trajectoires

Le groupement de trajectoires est un des sujets d'analyse de trajectoires les plus populaires. Le problème du groupement de trajectoire est double : Comment représenter les trajectoires, et comment définir les groupes ? Pour la représentation, si les méthodes classiques de traitement des séries temporelles sont applicables, il existe des représentations propres aux données de trajectoire.

Algorithmes de groupement

On peut grouper les algorithmes de groupement classiques en quatre types :

Type	Distance	Avantages	Inconvénients
Connectivité	Point à point	Peut être appliqué sur une matrice de distance	Complexité en $\mathbb{O}(n^3)$ ou $\mathbb{O}(2^n)$
Centroïdes	Point à centroïde	Particulièrement adapté aux gros jeux de données	Nécessite de spécifier le nombre de groupes, assume que les groupes sont sphériques
Densité	Point à point	Permet les groupements non sphériques	Sensible au bruit et à l'initialisation des hyperparamètres
Modèle	Point à prototype	Capture la corrélation entre les dimensions, approche probabiliste	Nécessite de définir un modèle approprié aux données, enclin au surapprentissage

TABLE 2.2 – Typologie des algorithmes de groupement

Groupement souple et fort Les algorithmes de groupement basés modèle, et certains basés centroïdes, permettent de faire du groupement souple : contrairement au groupement fort, qui assigne à chaque point un label, le partitionnement souple permet d'associer à chaque point un score d'appartenance à chaque groupe. Cela permet de gérer à la fois l'incertitude et l'appartenance d'un point à plusieurs groupes.

2.2.3 Représentations spécifiques

Mesures intrinsèques

Cette première façon de représenter une trajectoire est dans la lignée des autres caractéristiques comme celles de l'ensemble Catch22 présentées précédemment. L'idée est de développer de manière théorique des mesures de trajectoire issues de l'état de l'art du champ de recherche d'origine[60, 61].

Avantages Parce qu'elles sont issues d'une connaissance des données analysées, et qu'elles sont conçues pour un problème, ces métriques sont facilement interprétables, et permettent de biaiser le groupement dans une direction souhaitée.

Inconvénients Mais pour ces mêmes raisons, il est compliqué de les utiliser sur d'autres données ou d'autres problèmes. Pour ce faire, il serait nécessaire de construire un ensemble de métriques issues de problèmes et de jeux de données variés, et de filtrer ou pondérer cet ensemble lorsque l'on souhaite les appliquer à de nouvelles données ou un nouveau problème. De plus, elles ne permettent pas d'identifier de nouvelles caractéristiques à partir des données.

Représentations inspirées du traitement du langage

L'analyse de trajectoire s'est beaucoup nourrie des techniques développées pour l'analyse de données textuelles, en procédant à une analogie évènement - mot. Une trajectoire est représentée par un ensemble de symboles-évènements, soit sous la forme d'une séquence[62] ou d'un *sac-de-mot,DeWall2022* La définition des évènements peut être empirique ou théorique.

Avantages La représentation d'une trajectoire par un ensemble d'évènements peut permettre d'introduire une composante sémantique à l'analyse. Le traitement du langage est un champ de recherche prolifique qui propose de nombreux descripteurs et algorithmes qui permettent de conserver la dimension temporelle de la trajectoire.

Inconvénients Les algorithmes issus du traitement du langage sont souvent plus lourds computationnellement, leur application sur de larges jeux de données peut s'avérer compliquée. La discrétisation des trajectoires en les représentant par un ensemble d'évènements

entraîne une perte d'information dans le signal.

Similarité inter-trajectoires

Plutôt que de projeter les trajectoires dans un espace de représentation pour y mesurer leurs similarités, on retrouve dans la littérature de nombreuses approches qui proposent de mesurer cette similarité directement entre les trajectoires elles mêmes[5, 63].

Avantages Ces approches permettent de conserver toute l'information des signaux pour les comparer : il n'y a pas de perte dûe à la transformation de la trajectoire pour la représenter. Elles permettent aussi de faire du groupement en continu : le groupement se fait au fur et à mesure de la trajectoire, qui peut changer de groupe entre son début et sa fin.

Inconvénients Les fonctions de distance sont particulièrement coûteuses en calcul, et le stockage d'une matrice de similarité coûteux en mémoire, ce qui rend ces méthodes peu adaptées aux gros jeux de données.

2.2.4 Apprentissage profond appliqué aux données de trajectoires

Comme on l'a vu plus tôt, l'intérêt des méthodes d'apprentissage profond est qu'elle permette une définition empirique des caractéristiques du signal appropriées à une tâche donnée. Elles requièrent généralement une annotation des données, sauf à utiliser des tâches prétextes[11, 54-56, 64, 65].

L'utilisation de modèles MLP est rare, puisque cette architecture ne tient pas compte de la structure du signal temporel. Les modèles convolutionnels ou récurrents permettent de prendre en compte cette structure, mais ont chacun des avantages et désavantages.

Réseaux convolutionnels

Avantages L'utilisation de réseaux convolutionnels permet d'extraire des caractéristiques locales de la trajectoire. Ils tendent à être plus facilement interprétables que les réseaux récurrents.

Inconvénients Les réseaux convolutionnels mesurent uniquement les caractéristiques locales à un instant t d'une trajectoire, sans mémoire des caractéristiques à $t-\tau$, $\forall\, k/2 < \tau < t$, avec k la taille du noyau de convolution. L'utilisation d'opérations de sous-échantillonnages entre les couches de convolution peut permettre de simuler cette prise en compte du contexte, mais complexifie l'interprétation des modèles.

Réseaux récurrents

Avantages Contrairement aux réseaux convolutionnels, les réseaux récurrents intègrent un mécanisme de mémorisation des états précédents.

Inconvénients Mais, en conséquence, ils sont plus difficilement interprétables.

Traitement du signal sur graphe

Puisqu'elles sont tirées de l'analyse de séries temporelles, les réseaux convolutionnels et récurrents ne tiennent pas compte de la structure spatiale des données de trajectoire. Cette dimension est perçue uniquement comme une caractéristique locale du signal, au même titre que la vitesse ou l'accélération par exemple. Ces modèles présentent aussi le problème de difficilement pouvoir traiter des trajectoires avec un nombre d'échantillons hétérogène, ce qui implique de devoir ré-échantillonner les données, traitement source de distorsion du signal.

Pour répondre à ces problèmes en s'inspirant de la représentation sémantique des données de trajectoire, il a été proposé de représenter une trajectoire sur un graphe[10, 66]. Ces méthodes cherchent à extraire des données de trajectoire des points d'intérêts à partir desquels elles définissent un graphe sur lequel projeter le signal des trajectoires. Le signal est ici analysé de l'angle spatial, l'information temporelle est perçue uniquement comme une caractéristique.

Le développement de techniques de traitement du signal sur graphe[67, 68] ont permis le développement de modèles d'apprentissage profond sur graphe pour l'analyse de trajectoires sur graphe[9, 69, 70].

2.3 Applicabilité des méthodes développées dans la littérature

Les données Sea Hero Quest sont particulières, et présentent de nombreuses caractéristiques qui rendent leur analyse par les méthodes déjà développées compliquée :

- On a, pour chaque niveau, plusieurs dizaines de milliers de trajectoires, ce qui rend l'utilisation d'algorithmes de mesure de similarité très coûteuse et compliquée à mettre en place,
- Tous les joueurs doivent exécuter la même tâche, ce qui implique une très grande similarité entre les trajectoires. Ce qui détermine principalement la distance entre deux trajectoires est la stratégie adoptée par les joueurs, qui n'est pas forcément l'information discriminante pour identifier l'influence des traits démographiques sur le comportement spatial. Et, quand bien même ce serait la stratégie qui serait discriminante, la faible complexité des environnements implique que les différences de stratégie n'ont pas d'impact massif sur la forme des trajectoires (si le joueur A contourne un objectif par la droite et le joueur B par la gauche, la différence est faible à l'échelle de la trajectoire toute entière),
- Ces trajectoires sont de longueurs variées : on souhaite, si possible, éviter d'avoir à les ré-échantillonner, ce qui introduirait une perte d'information dans un signal déjà distordu à la collecte,
- On ne dispose pas de vérité-terrain, le lien entre démographie et comportement spatial n'a pas été clairement établi, c'est justement ce que l'on cherche à faire ici. Il faut donc penser nos approches en tenant compte de l'incertitude inhérente à notre tâche. On ne peut pas assurer qu'il y a une information à trouver, et s'il y en a une, elle est vraisemblablement faible (voir la figure 2.3). On ne peut donc pas appliquer tels quels les méthodes d'apprentissage profond sur nos données, au risque de converger quasi-systématiquement vers des solutions dégénérées,
- La topologie des niveaux de Sea Hero Quest est dans leur majorité assez simple en comparaison aux réseaux routiers urbains analysés dans la littérature. Le nombre de routes et intersections est assez faible, il faut le prendre en compte pour identifier les points d'intérêts si l'on souhaite construire un graphe.
- On souhaite effectuer un groupement à la fois au niveau des données de trajectoire, qui représentent le comportement, et au niveau des données démographiques. Ces

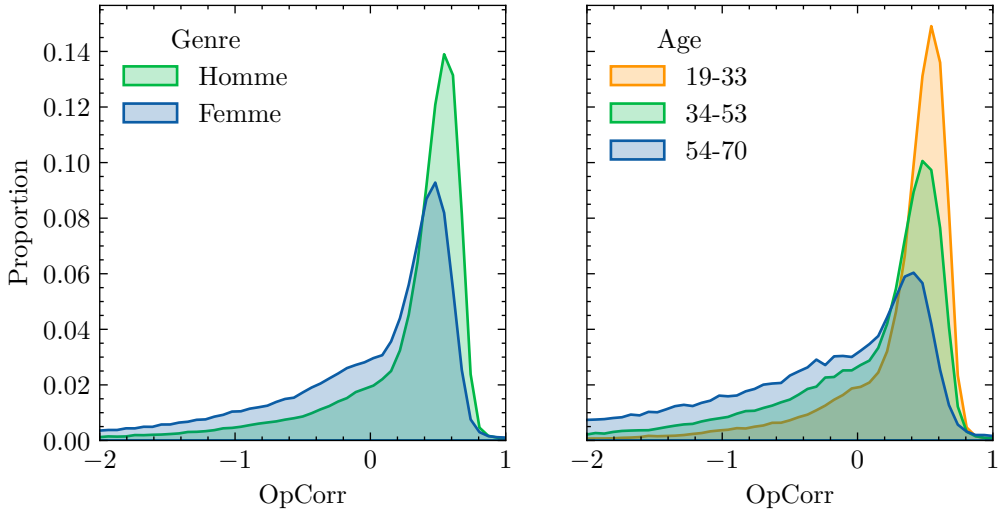


FIGURE 2.3 – Distributions du score OpCorr (voir section 1.4.1) en fonction du genre ou de l'âge. Les distributions se chevauchent fortement, ce qui montre la faiblesse de l'information présente dans les données.

deux ensembles de donnée ne sont pas du même type, il faut donc adapter notre approche à l'hétérogénéité dans la représentation des données.

Les méthodes et approches que nous avons essayées de développer au cours de cette thèse et que nous allons présenter ici essayent de répondre à ces problématiques.

En résumé :

✓ Il existe de nombreuses méthodes d'analyse de trajectoire issues de nombreux champs académiques différents.
✓ Les spécificités du jeu de donnée Sea Hero Quest rendent nécessaire le développement de nouvelles méthodes.
✓ Ces méthodes ne doivent pas être applicables qu'aux données Sea Hero Quest, mais doivent permettre d'analyser toutes données de trajectoires définies dans un environnement contraint et résultant d'une même tâche, ce qui induit une faible variété.
✓ Nous devons aussi proposer une méthode qui permet de lier l'analyse des trajectoires aux profils démographiques des joueurs.

Les méthodes proposées devront :

- Tenir compte de la nature spatio-temporelle des données de trajectoire
- Permettre d'évaluer l'impact des différentes caractéristiques démographiques sur le comportement spatial
- Permettre d'identifier des groupes homogènes d'un point de vue comportemental et démographique.

Chapitre 3

Données utilisées

3.1 Pré-traitement des données

3.1.1 Données de trajectoire

Par la suite nous allons utiliser les normes de notation suivantes :

s le signal d'une trajectoire

l la longueur d'une trajectoire

t un indice temporel

θ l'information de direction d'une trajectoire

$d\theta$ l'information de courbure d'une trajectoire

3.1.2 Données démographiques

Discrétisation de l'age

Dans le jeu de données brut, l'âge est représenté en années. Il présente un grand déséquilibre, et étant représenté par une variable continue, il se prête à une tâche de régression, ce qui complique la gestion de ce déséquilibre. Pour pouvoir mieux le gérer, et transformer la tâche de régression en tâche de classification, on transforme la variable âge $\in \mathbb{N}$ en un vecteur $\in [0;1]^k$. Chaque dimension de ce vecteur correspond à la probabilité d'appartenance à une tranche d'âge. Pour obtenir ces valeurs, on répartit uniformément k distributions gaussiennes univariées entre 18 et 99 (valeurs minimale et maximale possible pour l'âge).

En ne discrétisant pas strictement (c'est-à-dire en ne faisant pas un encodage *one-hot*), on permet de conserver une continuité dans la représentation de l'âge, et ainsi de ne pas imposer arbitrairement un effet de seuil sur les données. En effet, on ne sait pas à partir de quel âge exactement un effet de l'âge sur le comportement spatial est significatif.

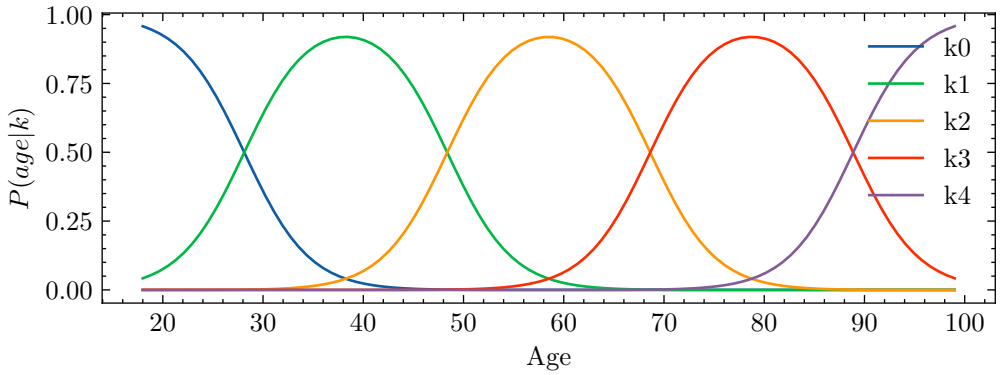


FIGURE 3.1 – Distributions utilisées pour l'encodage de l'age avec $k = 5$

Discrétisation du sommeil

Comme l'âge, le nombre d'heures de sommeil par nuit est une variable continue dans nos données. Pour faciliter son traitement par les algorithmes, on transforme cette variable continue en variable catégorique : on crée trois classes, peu de sommeil (Low Sleep), sommeil normal (Avg Sleep), et trop de sommeil (Hi Sleep), un sommeil normal étant un sommeil entre 6 et 10 heures inclus.

3.2 Sélection des données et métriques utilisées pour les expérimentations

3.2.1 Choix des niveaux

Le jeu de données SHQ a été collecté sur les soixante niveaux du jeu. Même si le nombre de joueurs ayant complété le niveau (et donc le nombre de trajectoires) décroit au fur et à mesure qu'on avance dans le jeu, cela représente une quantité impressionnante de données.

Pour des raisons logistiques (principalement d'espace disque), il a été décidé de n'utiliser qu'un sous ensemble des niveaux du jeu pour l'évaluation des méthodes développées dans cette thèse. De plus, pour assurer une base comparable entre les niveaux, seuls sont gardés les joueurs ayant donné toutes les informations démographiques, et ayant complété tous les niveaux du sous ensemble ont été gardés. Cela réduit grandement le jeu de données, ne laissant que 16 571 utilisateurs. Les niveaux qui ont été sélectionnés sont les niveaux 8, 32, 36, 56 et 67 (voir figure 3.2). Ils ont été choisis arbitrairement, l'intention

étant d'avoir un ensemble de niveaux avec des topologies assez diversifiées.

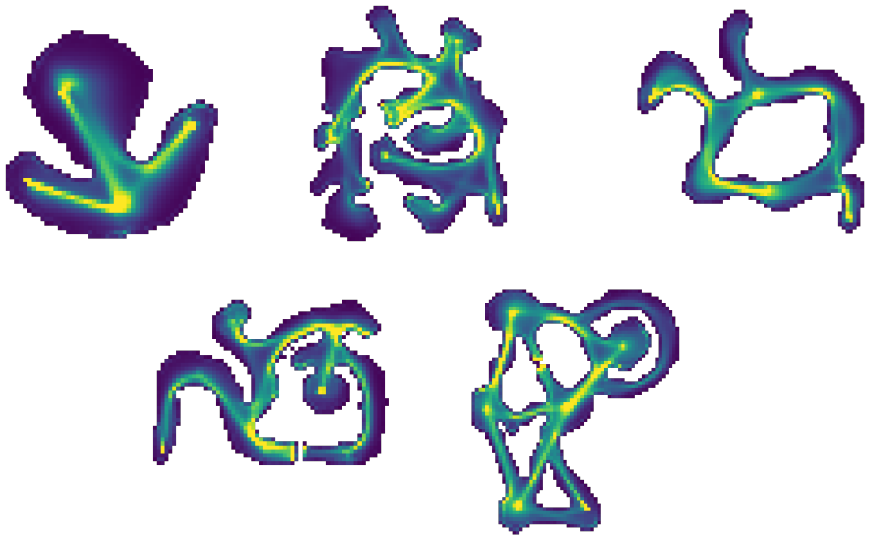


FIGURE 3.2 – Cartes des niveaux choisis

Les profils démographiques de ces 16 751 utilisateurs suivent une distribution similaire à celle des profils démographiques des 772 827 utilisateurs ayant renseigné toutes leurs informations démographiques (voir figure 3.3).

Parce qu'un même joueur peut rejouer plusieurs fois au même niveau, on a plus de trajectoires par niveau que d'utilisateurs. On a donc, respectivement, 36276, 29162, 23037, 23042, et 24922 trajectoires pour les niveaux du sous ensemble. Pour le niveau 8, environ 60% des joueurs n'ont joué qu'une fois. Pour les autres niveaux, ce chiffre monte à 80% (voir figure 3.4).

Représentation des données démographiques

Pour pouvoir comparer entre elles les différentes approches proposées par la suite, on propose ce format unifié pour les données démographiques :

- Genre
  - Homme • Femme
- Main
  - Droitier • Gaucher
- Environnement
  - Urbain • Mixe • Périphérique • Rural

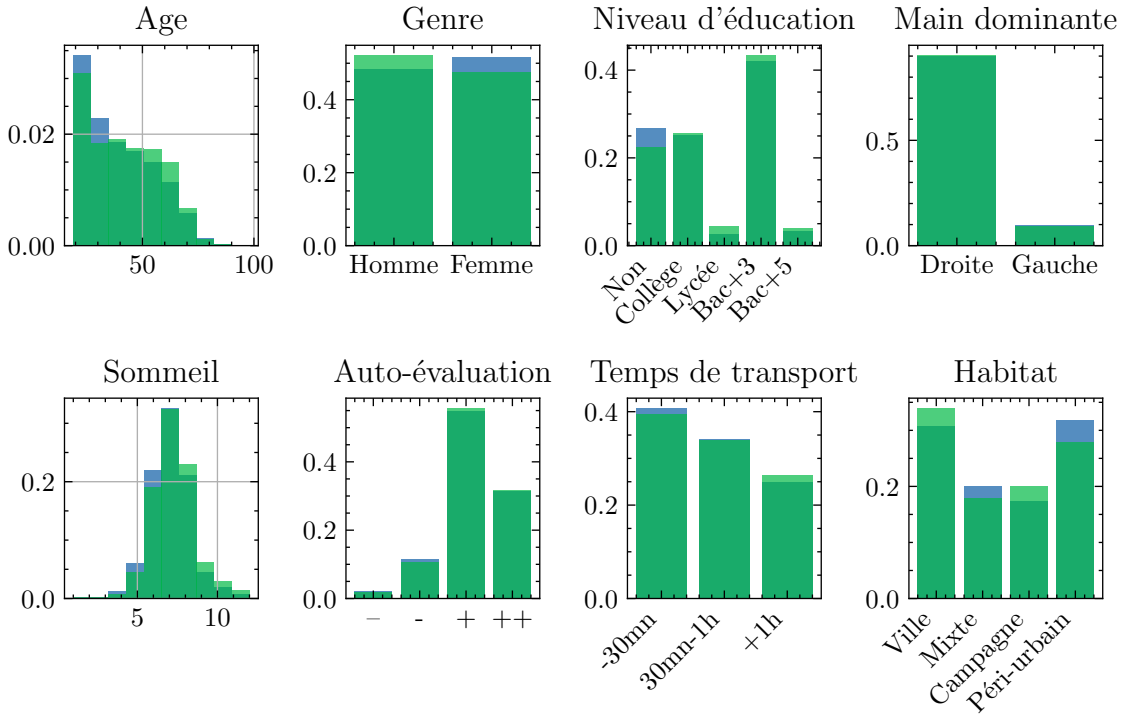


FIGURE 3.3 – Distributions des profils démographiques de l'ensemble des joueurs (en bleu) et du sous-ensemble sélectionné (en vert)

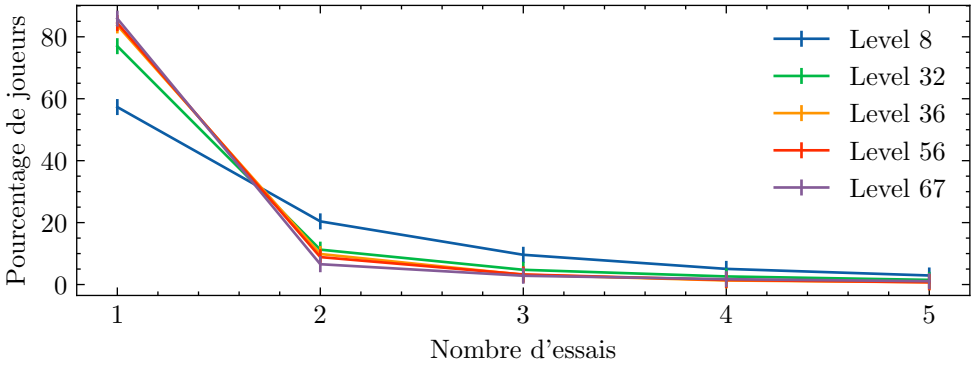


FIGURE 3.4 – Nombre d'essais par joueur pour chaque niveau du sous-ensemble.

- Éducation
  - Aucune • Collège • Lycée • Bac+3 • Bac+5
- Auto-évaluation
  - −− • − • + • ++
- Sommeil
  - <6h • 6 à 10h • >10h
- Transport
  - <30mn • 30mn à 1h • >1h
- Âge (voir section 3.1.2)
  - k1 • k2 • k3 • k4 • k5

Grâce à cette représentation, nous n'avons plus de caractéristiques démographiques représentées par une variable continue. Nous pouvons donc approcher toutes les tâches comme des problèmes de classification.

Du point de vue des modèles, ces données ne sont pas "structurées". Elles sont présentées sous la forme d'un vecteur de longueur 28 de valeurs comprises entre 0 et 1. Nous la spécifions ici, parce qu'à posteriori, lorsque nous évaluerons les performances des modèles, nous serons amenés à le faire par caractéristique démographique.

Précision Par la suite, lorsque nous parlerons d'une *caractéristique*, nous ferons référence à une information de niveau 1 dans la liste ci-dessus (Genre, Âge, etc.). Une caractéristique est composée de *dimensions* démographiques, ici des informations de niveau 2 dans la liste (par exemple, pour le genre, il y a deux dimensions, `Homme` et `Femme`).

Nettoyage des données de trajectoire

Au moment de la collecte dans le jeu, les données de trajectoire sont discrétisées de $\mathbb{R}^2$ dans $\mathbb{N}^2$. Une partie de l'information est donc perdue au passage. L'information de direction en particulier est dégradée : on ne peut pas extrapoler fiablement la direction du joueur à partir de ses coordonnées.

Pour essayer de retrouver cette information, on pose l'hypothèse suivante : un comportement naturel aura plus tendance à être lisse que chaotique. On sait aussi que la coordonnée réelle se trouve dans le carré de côté 1 centré sur la coordonnée discrétisée.

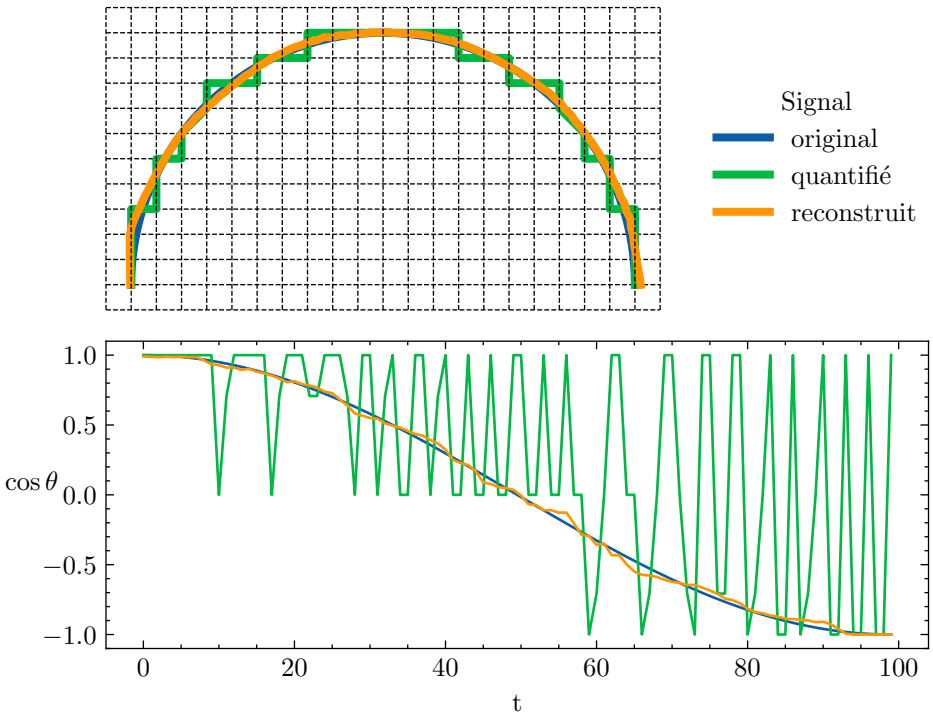


FIGURE 3.5 – Exemple de signal reconstruit

On va donc chercher pour chaque trajectoire s un vecteur $\epsilon \in [0,5;0,5]^{l\times 2}$, avec l le nombre d'échantillons de s. On définit $s' = s + \epsilon$ la trajectoire recomposée. On va calculer ϵ en minimisant l'équation suivante :

$$\min_{\epsilon} \sum \left|\left| \frac{dx, y}{ds'} \right|\right|^2$$

La figure 3.5 montre le résultat de cette méthode sur une trajectoire synthétique. La reconstruction permet de retrouver l'information de direction qui était autrement perdue.

Certains modèles nécessitent d'avoir une même longueur de trajectoire dans l'ensemble du jeu de donnée. Pour répondre à cette contrainte, nous utilisons une méthode de ré-échantillonnage régulier de la trajectoire, en fixant le nombre d'échantillons pour un même niveau. On le définit comme le troisième quartile de la distribution des longueurs des trajectoires du niveau.

Métriques

Évaluation des classifieurs Pour évaluer tous les modèles sur le même plan, nous allons utiliser le *Kappa de Cohen*[71] noté κ. C'est une métrique d'évaluation de classification, qui mesure l'accord entre deux annotateurs. On a $-1 \leq \kappa \leq 1$, où $\kappa = -1$ indique que les deux annotateurs sont en désaccord complet, $\kappa = 0$ que leur accord est complètement aléatoire, et $\kappa = 1$ qu'ils sont systématiquement d'accord. L'intérêt du Kappa de Cohen, et qu'il prend en compte la chance aléatoire d'être d'accord, et permet donc d'avoir une métrique normalisée qui peut être utilisée pour comparer des résultats obtenus sur des jeux de données avec un nombre de classes différent et un équilibre de classe différent.

On définit :

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Où $Pr(a)$ est la proportion d'accord mesuré entre les deux annotateurs, et $Pr(e)$ la probabilité d'un accord aléatoire.

Mesure de l'importance des caractéristiques démographiques Pour mesurer l'importance relative de chaque caractéristique démographique, sauf précision du contraire, nous utilisons l'importance de permutation[72] $\mathbf{I}$. Elle mesure la relation entre les dimensions de la variable indépendante x et la variable dépendante y.

On définit :

$$\mathbf{I}_c = L(f(\text{perm}(x, c), y)) - L(f(x, y))$$

Où L est une métrique d'évaluation de modèle, f un prédicteur, et $\text{perm}(x, c)$ une permutation des valeurs de la caractéristique c des données x. Si $\mathbf{I}_c = 0$, alors la caractéristique c n'est pas importante pour le modèle. Si $\mathbf{I}_c < 0$, alors sa permutation améliore les performances du modèle : son importance est négative. Au contraire, si $\mathbf{I}_c > 0$, alors

cette caractéristique est importante pour le modèle, puisque sa permutation dégrade les performances.

3.3 Résultats préalables

Comme on l'a vu dans l'introduction, il existe déjà des analyses de l'impact des différentes caractéristiques démographiques sur le comportement spatial développées à partir du jeu de données Sea Hero Quest. Celle qui essaie de faire le lien entre trajectoire et caractéristiques démographiques se base sur une mesure de la longueur des trajectoires d'un joueur, le score OpCorr (voir la section 1.4.1), qu'elle essaie de prédire à partir du profil démographique au moyen d'une régression linéaire.

Pour mesurer l'impact des différentes caractéristiques démographiques sur le comportement spatial, un test F est utilisé. La figure 3.6 montre les résultats de ce test : l'âge du joueur est le principal facteur, contribuant à plus de 70% de la prédiction, suivi par le genre. Les autres caractéristiques ont une importance marginale.

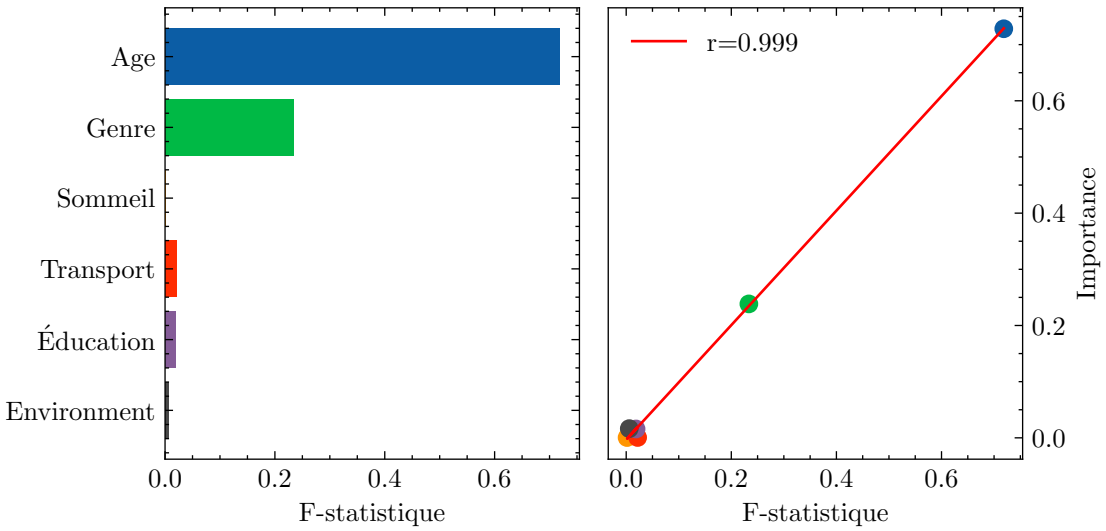


FIGURE 3.6 – Résultats de l'analyse de l'importance des caractéristiques démographiques dans la prédiction du score OpCorr par une régression linéaire. Les résultats du test F montrent une très forte corrélation avec l'importance mesurée par permutation. Il est à noter qu'il manque certaines des caractéristiques démographiques que nous utiliserons dans nos expérimentations.

Pour utiliser ces résultats comme base de comparaison pour les approches que nous allons développer à partir de maintenant, on mesure la correspondance entre les résultats

du test F et l'importance par permutation comme nous l'avons définie plus tôt. On peut voir que la corrélation entre ces deux mesures est très forte sur la figure 3.6.

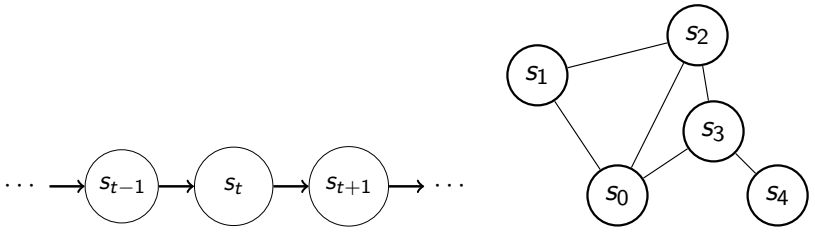
Chapitre 4

Multi-représentations de trajectoires

Une partie des travaux présentés dans ce chapitre ont fait l'objet d'une publication à EUSIPCO2020 [73]. Depuis la publication de cet article, la méthode a évolué sur certains aspects pour corriger les faiblesses identifiées, mais les principes explorés restent les mêmes.

4.1 Comment représenter une donnée spatio-temporelle ?

Les trajectoires de navigation sont des signaux riches, produit d'interactions complexes entre un environnement et un individu, façonné par un milieu social, avec son état cognitif, ses objectifs, etc. (voir sous-section 1.2.1). De ce fait, elles ont deux aspects indissociables, un aspect spatial, qui reflète l'impact de l'environnement, et un aspect temporel, qui reflète le processus cognitif de l'individu, les deux interagissants entre eux, ces aspects sont indissociables.



(a) Représentation temporelle (b) Représentation spatiale

La plupart des méthodes présentes dans la littérature choisissent de se placer du point de vue de l'un de ces aspects. Les approches sac-de-mot par exemple choisissent celui de l'espace, chaque évènement est la visite d'un lieu, la temporalité de ces visites est perdues. Au contraire, les approches utilisant des réseaux de neurones de type CNN ou RNN se

placent de l'angle temporel, et laissent le soin au modèle d'apprendre l'aspect spatial à partir des dimensions du signal temporel.

À nos yeux, cela ne suffit pas, surtout pour des données comme celles de Sea Hero Quest où l'environnement est simple et l'information faible. Une modélisation des données de trajectoire se doit donc d'analyser ce signal sous ces deux aspects, au risque de perdre une quantité cruciale d'information. Notre hypothèse de départ est aussi que l'utilisation d'une architecture tirant profit de plusieurs représentations différentes d'une même donnée permettrait de limiter la profondeur du modèle, permettant ainsi d'avoir un meilleur compromis performance/explicabilité.

Nous allons donc définir plusieurs modalités de représentation d'une trajectoire, chacune spécifique à un aspect de la donnée, avec une architecture de modèle à réseau de neurones propre.

4.2 Représentation temporelle de la trajectoire

Cette représentation est la représentation "par défaut" de la trajectoire, sous la forme d'une série temporelle.

Le signal originel est constitué de deux dimensions, les coordonnées x et y. Les trajectoires brutes étant de longueurs variables, on utilise des trajectoires lissées et réechantillonnées pour avoir le même nombre de points, comme détaillé dans section 3.2.1. À partir de ces deux dimensions, on calcule les dimensions additionnelles suivantes (voir figure 4.1) :

- la vitesse,
- l'accélération,
- la direction (Nord-Sud-Est-Ouest, NSEO) ;
  représentée par le couple $cos(\theta); sin(\theta)$,
- la courbure (Avant-Arrière-Droite-Gauche, AADG) ;
  représentée par le couple $\delta cos(\theta); \delta sin(\theta)$

4.2.1 Modèle ad-hoc

Pour apprendre les caractéristiques des trajectoires d'un point de vue temporel, on utilise un réseau de neurones convolutif (CNN) à une dimension, avec à chaque couche deux sous-couches parallèles, chacune traitant le signal à une échelle différente. La première

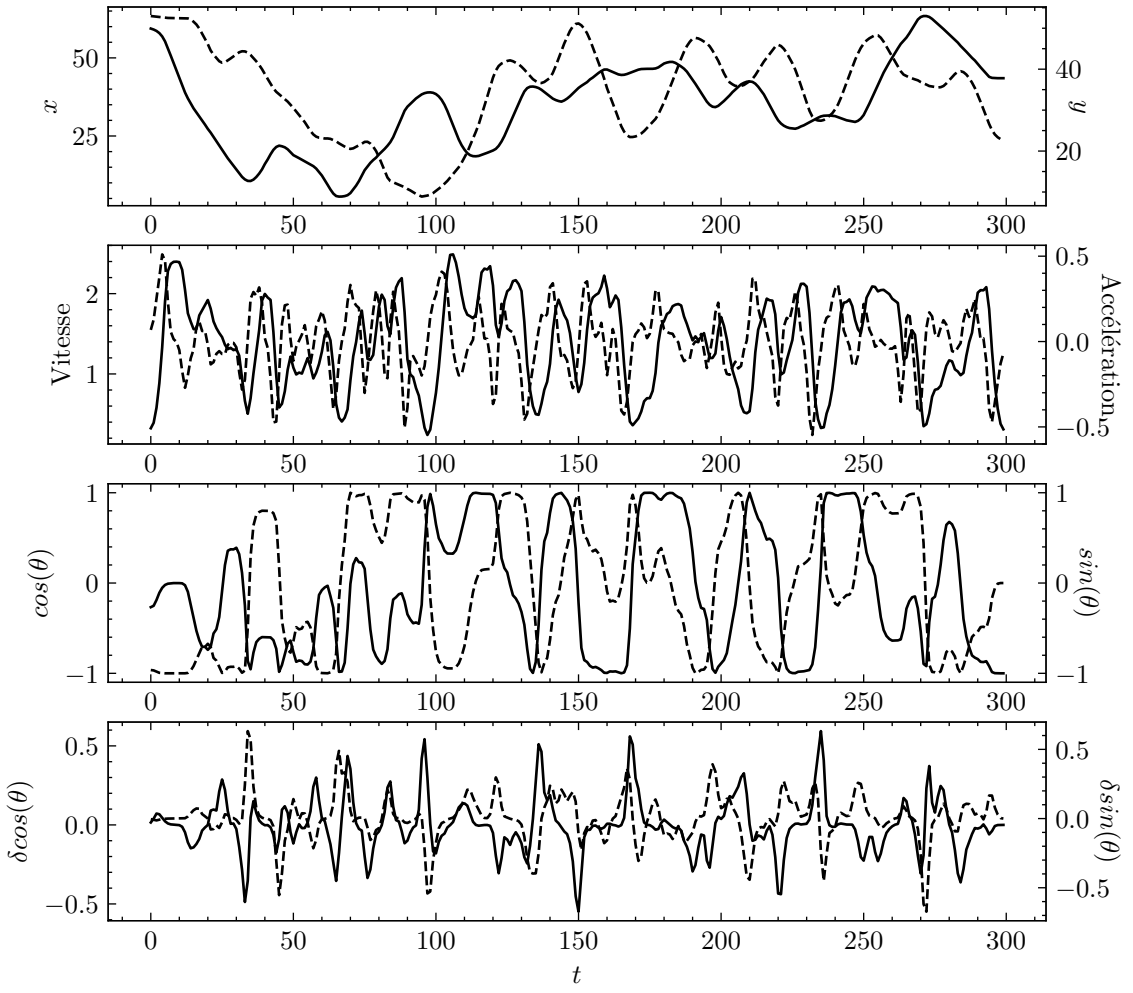


FIGURE 4.1 – Exemple de dimensions extraites d'une trajectoire. La dimension en pointillée correspond à la légende de droite.

analyse les caractéristiques basse-fréquence, qui correspondent plutôt à des marqueurs cognitifs de stratégie exploratoire, et l'autre analyse les caractéristiques haute-fréquence, qui expriment plutôt la réalisation immédiate de ces stratégies. Pour ce faire, on module deux paramètres des couches convolutives : la taille du noyau, c'est-à-dire la largeur de la fenêtre glissante, et le facteur de dilatation, qui élargit le noyau en "ignorant" des points entre ceux observés par la fenêtre. La couche traitant les caractéristiques basse-fréquence du signal a une taille de noyau k_{low} et un facteur de dilatation d_{low} grands, et celle traitant les caractéristiques haute-fréquence une taille de noyau k_{high} petite et pas de dilatation (voir figure 4.2). Les sorties de ces deux sous-couches sont concaténées et passées à la couche suivante.

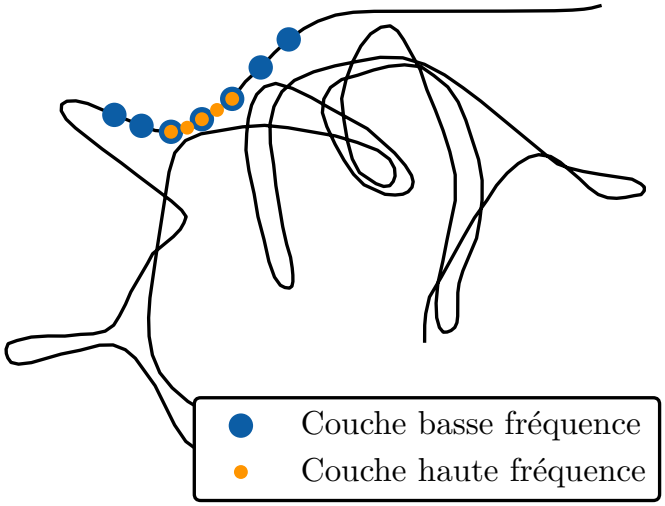


FIGURE 4.2 – Exemple des points vus par les deux sous-modules, avec $k_{low} = 7, d_{low} = 1, k_{high} = 5$.

4.3 Représenter la navigation spatiale avec des graphes

Pour saisir les caractéristiques spatiales de la trajectoire, on décide de la représenter sous la forme d'un signal défini sur un graphe $\mathcal{G}(N, E)$. Ceci permet plusieurs choses :

- Uniformiser la dimension de la donnée associée à une trajectoire sans avoir à la ré-échantillonner ;
- donner une représentation explicite à la stratégie du joueur, utilisable par un modèle.

Un graphe est généré à partir des données de chaque niveau. Chaque nœud du graphe est associé à une région de l'environnement dans lequel évolue le joueur. Le graphe représente donc l'environnement. La trajectoire, elle, est définie comme un signal sur les nœuds de ce graphe, qu'on extrait en la découpant en sous-trajectoires, chacune contenue dans la région associée au nœud. La figure 4.5 donne une illustration du lien entre d'un côté la carte et le graphe, et de l'autre la trajectoire et le signal.

Pour générer un graphe à partir d'un niveau, il faut définir une méthode adaptée aux espaces non euclidiens.

4.3.1 Définition du graphe

Une première approche pourrait être d'utiliser un algorithme de groupement sur tous les points de toutes les trajectoires (ou en tout cas un sous-ensemble) d'un niveau (figure 4.3).

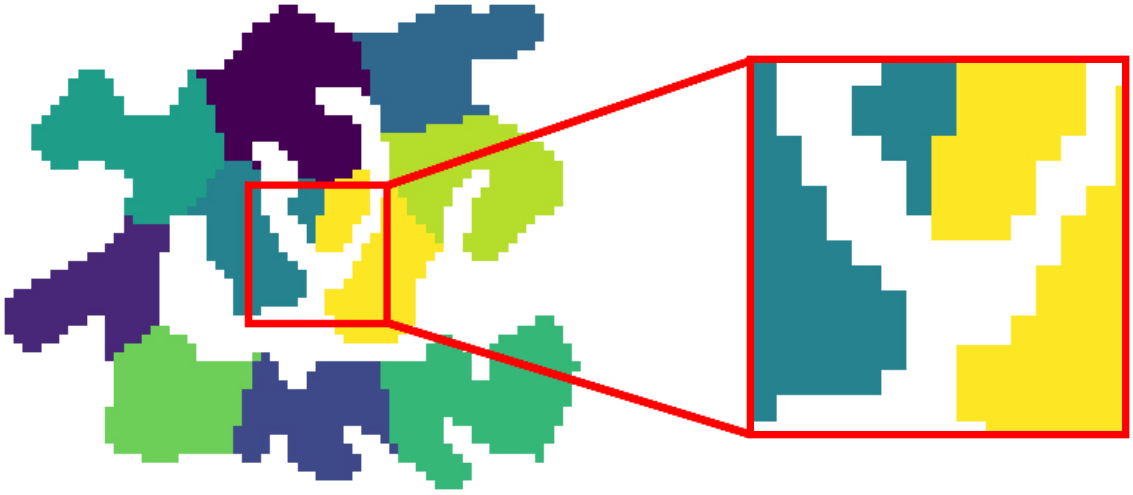


FIGURE 4.3 – Exemple de segmentation de la carte du niveau 32 en utilisant KMeans. À droite, exemple de résultat invraisemblable : une même zone traverse les murs.

Le problème de cette approche est que la majorité des algorithmes de groupement utilisent une distance euclidienne pour construire les groupes, ce qui donne des résultats invraisemblables lorsque appliqués dans un environnement non euclidien, ce que sont les niveaux de Sea Hero Quest (voir figure 4.3). Un autre problème est que cette approche identifie facilement les zones de passage, mais ne les sépare pas nécessairement des zones peu visitées, qui sont pourtant des zones particulièrement importantes dans notre tâche.

Pour capturer la singularité du signal, tout en maintenant le nombre de nœuds bas pour éviter la redondance et réduire la complexité du modèle, le graphe est défini à partir de la carte de chaleur du niveau, les nœuds en étant déterminés par segmentation. On utilise un algorithme de Watershed [74]. Les minimas locaux sont utilisés comme centres de "bassins" qui se remplissent petit à petit jusqu'à se rencontrer et former une segmentation.

Le graphe est construit à deux niveaux. Tout d'abord, en utilisant la carte de chaleur inverse (les zones les plus visitées sont les minimas locaux), on extrait des zones "larges" de la carte. Elles correspondent au comportement macroscopique : quelles zones le joueur visite-t-il ? Ensuite, dans chacune de ces zones larges, on utilise la carte de chaleur normale (les zones les plus visitées sont les maximas) pour y extraire les zones de "singularité", qui permettent d'identifier un comportement à une échelle plus microscopique (voir figure 4.4).

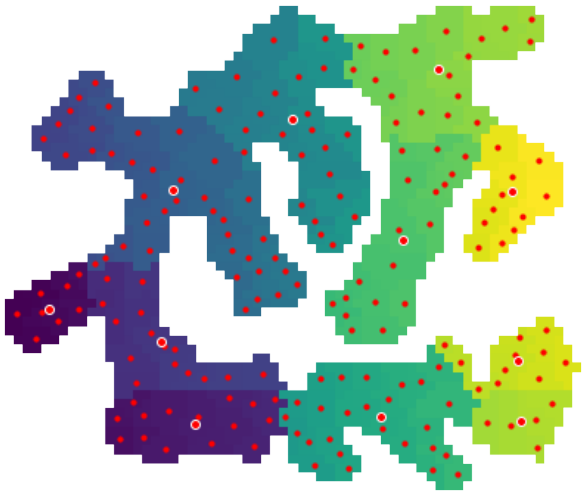


FIGURE 4.4 – Exemple de segmentation de la carte du niveau 32 en utilisant l'approche watershed. Les nœuds gros entourés de blanc correspondent aux zones macro, les petits aux zones micro

4.3.2 Définition du signal sur le graphe

Pour passer du signal de la série temporelle au signal défini sur le graphe, donc à un signal défini spatialement, il faut définir le signal sur chaque nœud de zone micro du graphe. Par défaut, un nœud non visité aura un signal nul. Un nœud visité a un signal qui peut être défini de différentes façons :

Signal de visite Dans cette configuration, le signal est simple : 1 si le nœud a été visité, 0 sinon (voir figure 4.5).

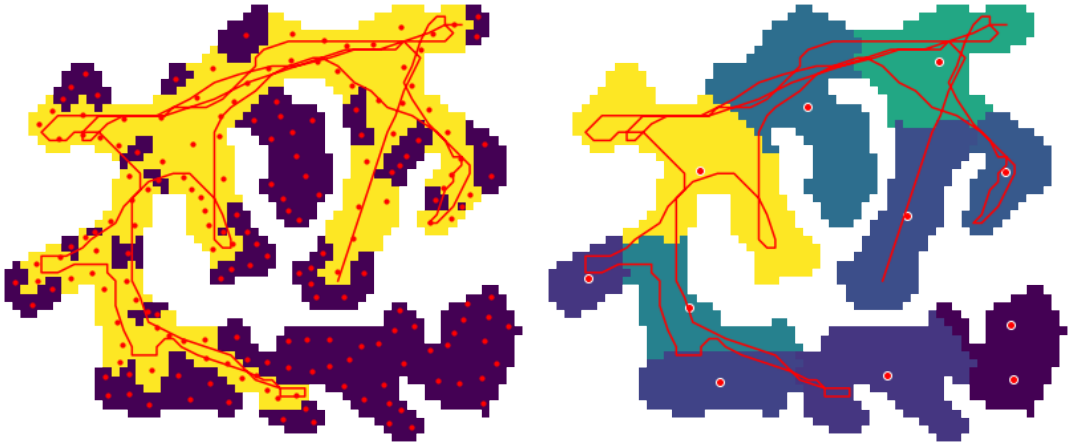


FIGURE 4.5 – Exemple de signal de visite associé à une trajectoire. A gauche le signal micro, à droite le signal macro. L'agrégation micro/macro se fait par une simple opération somme.

Signal de comportement Ici, le signal est multivarié. Il contient les dimensions suivantes (voir figure 4.1) :

- La vitesse moyenne (voir figure 4.6)
- L'accélération moyenne
- La direction moyenne (représentée par le couple cosinus/sinus de l'angle absolu θ)
- La courbure maximale (représentée par le couple cosinus/sinus de l'angle relatif $d\theta$)

Signal appris Le signal sur les nœuds du graphe est obtenu à partir des caractéristiques apprises par un CNN comme décrit précédemment. Le graphe agit ici comme une opération de *pooling* spatial sur un signal temporel. Cette définition permet de réaliser dans la modélisation le lien entre les caractères spatiaux et temporels de la trajectoire.

Ces caractéristiques sont calculées à partir du signal corrigé comme détaillé dans section 3.2.1. L'appartenance à un nœud se fait sur le signal brut, la segmentation du niveau se faisant dans $\mathbb{N}^2$.

4.3.3 Pooling sur graphe

Dans le cas des signaux de comportement et appris, qui sont définis dans l'espace temporel $\mathbb{T}$, leur projection dans l'espace du graphe $\mathbb{G}(\mathcal{N}, \mathcal{E})$ se fait en calculant le produit matriciel de la matrice d'assignation aux nœuds B avec la caractéristique f du signal $s_{\mathbb{T}}$:

$$\begin{array}{ccccc} B & \cdot & s_{\mathbb{T}}(f) & = & s_{\mathbb{G}}(f) \\ |\mathcal{N}| \times l & \cdot & l \times 1 & \mapsto & |\mathcal{N}| \times 1 \end{array}$$

Pour éviter qu'un nœud très visité ai un signal artificiellement plus fort que les autres, la matrice B est normalisée de telle manière à ce que la somme des lignes soit égale à 1.

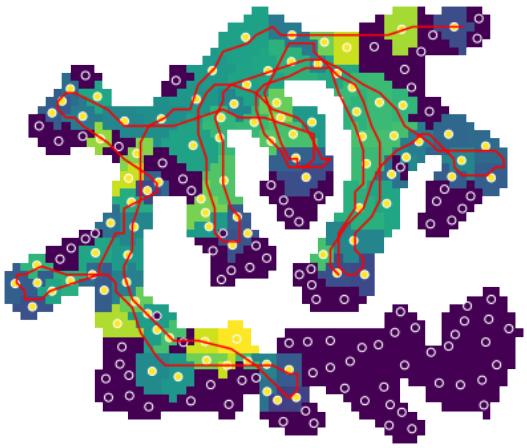


FIGURE 4.6 – Exemple de signal de comportement associé à une trajectoire. Ici on visualise uniquement la composante vitesse. Les nœuds visités sont représentés par un point jaune.

Nous appelons ce type d'opération transformant un signal temporel en signal spatial *Temp2Space*.

4.3.4 Modèle hiérarchique sur graphe

Pour traiter le signal défini sur le graphe, on tire profit de la structure hiérarchique du graphe : les caractéristiques sont évaluées d'abord à l'échelle micro-scopique, puis agrégées à l'échelle macro-scopique, un nœud macro traitant l'information venant des nœuds micros qui lui sont associés. En gardant cette structure d'arbre dans le graphe, on évite d'avoir besoin d'utiliser une approche de traitement de signal sur graphe qui nécessiterait d'avoir à localiser le signal et calculer son spectre pour y effectuer des opérations de convolution,

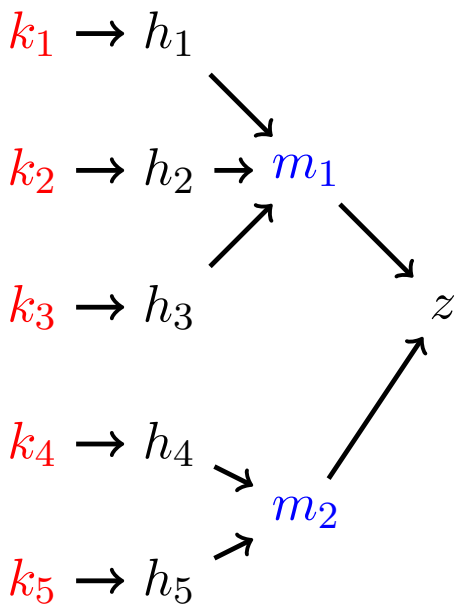


FIGURE 4.7 – Schéma du modèle hiérarchique sur graphe TreeNN, avec $\mathcal{N}_\mu = 5$ et $\mathcal{N}_M = 2$.

opérations plus coûteuses et moins explicables.

On définit le modèle comme un réseau de neurones composé de deux modules : un premier, f_μ, traitant le signal à échelle micro, puis un second, f_M, qui le traite à échelle macro. On l'exprime de la façon suivante :

$$\text{TreeNN}(s) = MLP(f_M(f_\mu(s)).\text{flatten}()) \qquad |\mathcal{N}_\mu| \times k \mapsto d \tag{4.1}$$

$$f_\mu(s) = MLP(s) \qquad |\mathcal{N}_\mu| \times k \mapsto |\mathcal{N}_\mu| \times h \tag{4.2}$$

$$f_M(s) = MLP(B \times s) \qquad |\mathcal{N}_\mu| \times h \mapsto |\mathcal{N}_M| \times m \tag{4.3}$$

Avec B la matrice d'adjacence du graphe biparti reliant les nœuds micros $\mathcal{N}_\mu$ aux nœuds macros $\mathcal{N}_M$, k le nombre de caractéristiques sur les nœuds micros en entrée du modèle, h et m les nombres de caractéristiques cachées sur les nœuds micros et macros respectivement, et d le nombre de dimensions démographiques à prédire.

4.4 Traiter la trajectoire comme un signal spatio-temporel : le réseau neuronal composite

Maintenant que nous avons défini des outils pour analyser les composantes temporelles de la trajectoire avec les réseaux CNN, et les composantes spatiales avec les réseaux sur graphe, il faut trouver une façon de combiner les deux.

Pour ce faire, nous proposons l'utilisation d'un modèle de type réseau neuronal avec une architecture que nous appellerons *composite* :

Les premières couches du modèle sont des modèles spécifiques, adaptés à une représentation des données, sans échanges entre elles. Chacun de ces modules produisent en sortie un vecteur de représentation z. Ces représentations sont ensuite agrégées, et traitées par un module entièrement connecté.

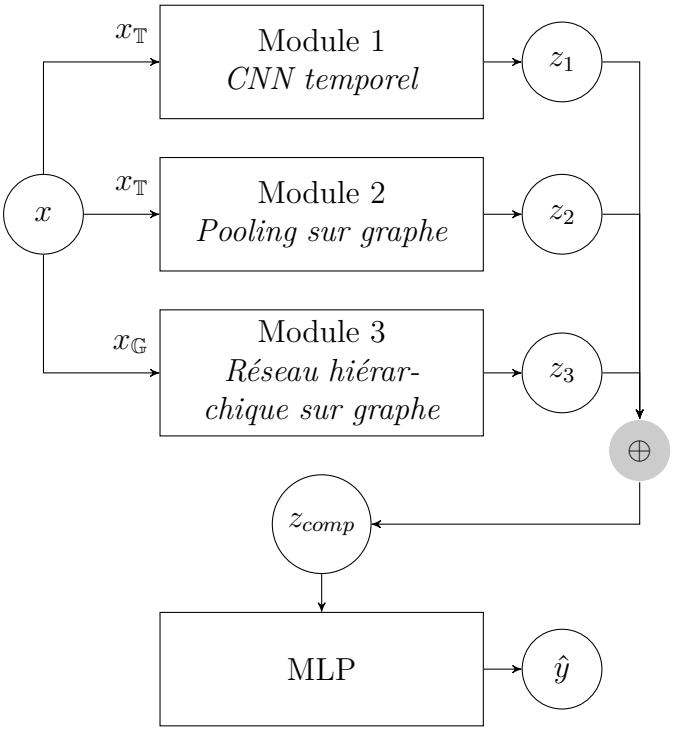


FIGURE 4.8 – Exemple de structure de réseau composite à 3 modules. Le modèle se veut générique, mais nous spécifions ici le type de module utilisé dans notre expérimentation en italique dans chaque bloc.

4.5 Évaluation

4.5.1 Expérimentation

Pour évaluer la pertinence de cette approche, nous définissons un protocole d'expérimentation : Pour chacun des niveaux de test, on apprend les paramètres d'un ensemble de modèles sur une tâche de classification multi-label. La donnée en entrée du modèle est la trajectoire, la prédiction en sortie est les labels associés au profil démographique du joueur.

Protocole On évalue la performance d'un modèle pour chaque groupe de label indépendamment des autres en en mesurant le Kappa de Cohen κ, métrique prenant en compte l'équilibre des classes. Étant donné que les caractéristiques démographiques sont la variable prédite ici, on ne peut pas utiliser la mesure de l'importance par permutation. On définit donc le score κ normalisé comme proxy de l'importance : une caractéristique mieux prédite est une caractéristique plus dépendante du signal en entrée.

Pour chaque niveau, quatre modèles différents sont testés :

ConvNN un réseau convolutif comme décrit dans la sous-section 4.2.1,

GraphMLP un modèle TreeNN avec comme signal le signal comportemental comme décrit dans la sous-section 4.3.4

GraphPool un réseau convolutif avec pooling graphe comme décrit dans la sous-section 4.3.3

CompSNN un réseau composite combinant les trois sous-réseaux définis ci-dessus. Les vecteurs de représentation des différents modules sont définis dans $\mathbb{R}^{64}$.

Pour chaque niveau, on évalue chaque modèle en calculant le κ moyen avec une validation croisée 5-plis toutes les cinq epochs d'apprentissage.

4.5.2 Résultats

La figure 4.9 présente les résultats de cette expérimentation. On note que la structure GraphMLP favorise très largement le genre, et n'arrive pas à prédire les caractéristiques autres que le genre, l'âge, et l'environnement. La structure ConvNN elle performe un peu moins bien sur le genre, mais un peu plus sur l'âge. Par contre, la troisième caractéristique qu'elle réussi le mieux à prédire n'est pas l'environnement, mais le niveau d'études. Les structures GraphPool et CompSNN, qui sont des structures hybrides, ont un peu le même

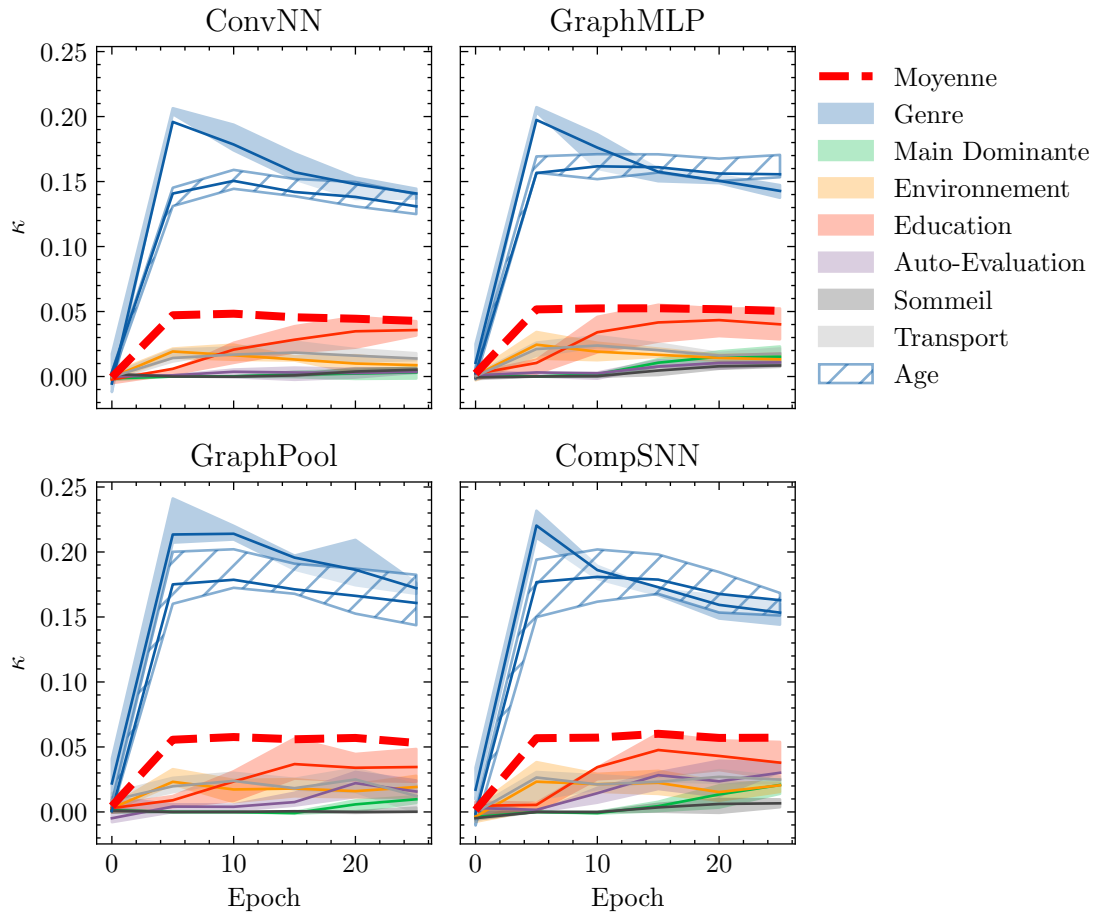


FIGURE 4.9 – Performances des 4 modèles sur tous les niveaux. L'intervalle rempli représente les premiers et troisièmes quartiles. Le modèle GraphMLP permet une meilleure prédiction de l'âge, on peut penser que c'est son ajout au CompSNN qui permet d'y retrouver ce trait.

profil. On remarque là encore une large prédominance de l'âge et du genre, mais ces modèles arrivent plus facilement à prédire les autres caractéristiques. Un point commun à ces quatre modèles est la perte de performance sur le genre et l'âge (à l'exception de GraphMLP pour l'âge) au fil des epochs, qui se fait au profit des autres caractéristiques. On remarque avec un délai la même chose avec le niveau d'éducation chez CompSNN.

Importance des caractéristiques démographiques

Pour mesurer l'importance nous gardons pour chaque combinaison modèle/niveau la meilleure epoch du point de vue du κ moyen.

Niveau	Modèle	Meilleur κ moyen
8	CompSNN	**0.048069**
	ConvNN	0.036807
	GraphMLP	0.039274
	GraphPool	0.043827
32	CompSNN	**0.075830**
	ConvNN	0.055880
	GraphMLP	0.061122
	GraphPool	0.070811
36	CompSNN	**0.058154**
	ConvNN	0.049522
	GraphMLP	0.051089
	GraphPool	0.054681
56	CompSNN	**0.066114**
	ConvNN	0.052222
	GraphMLP	0.058557
	GraphPool	0.064108
67	CompSNN	**0.073394**
	ConvNN	0.050063
	GraphMLP	0.057064
	GraphPool	0.062869

TABLE 4.1 – Performances des différents modèles sur les données de chaque niveau utilisé pour l'expérimentation. Le modèle composite CompSNN est systématiquement meilleur que les trois autres, alors que le modèle temporel ConvNN est systématiquement le moins bon.

4.6 Conclusion

4.6.1 Apports

Dans ce chapitre, nous avons essayé de répondre aux problèmes posés par l'application de méthodes tirées de l'état de l'art sur le jeu de données Sea Hero Quest. Pour ce faire, nous avons proposé une méthode de pooling permettant de passer d'un signal vu sous l'angle temporel à un signal vu sous l'angle spatial, et une architecture de réseau de neurones spécifique pour le traitement d'un signal spatial défini dans un espace non euclidien hiérarchisé. Nous avons aussi proposé une architecture de modèle à réseau de neurones combinant différentes représentations spécifiques d'un signal pour l'analyser sous

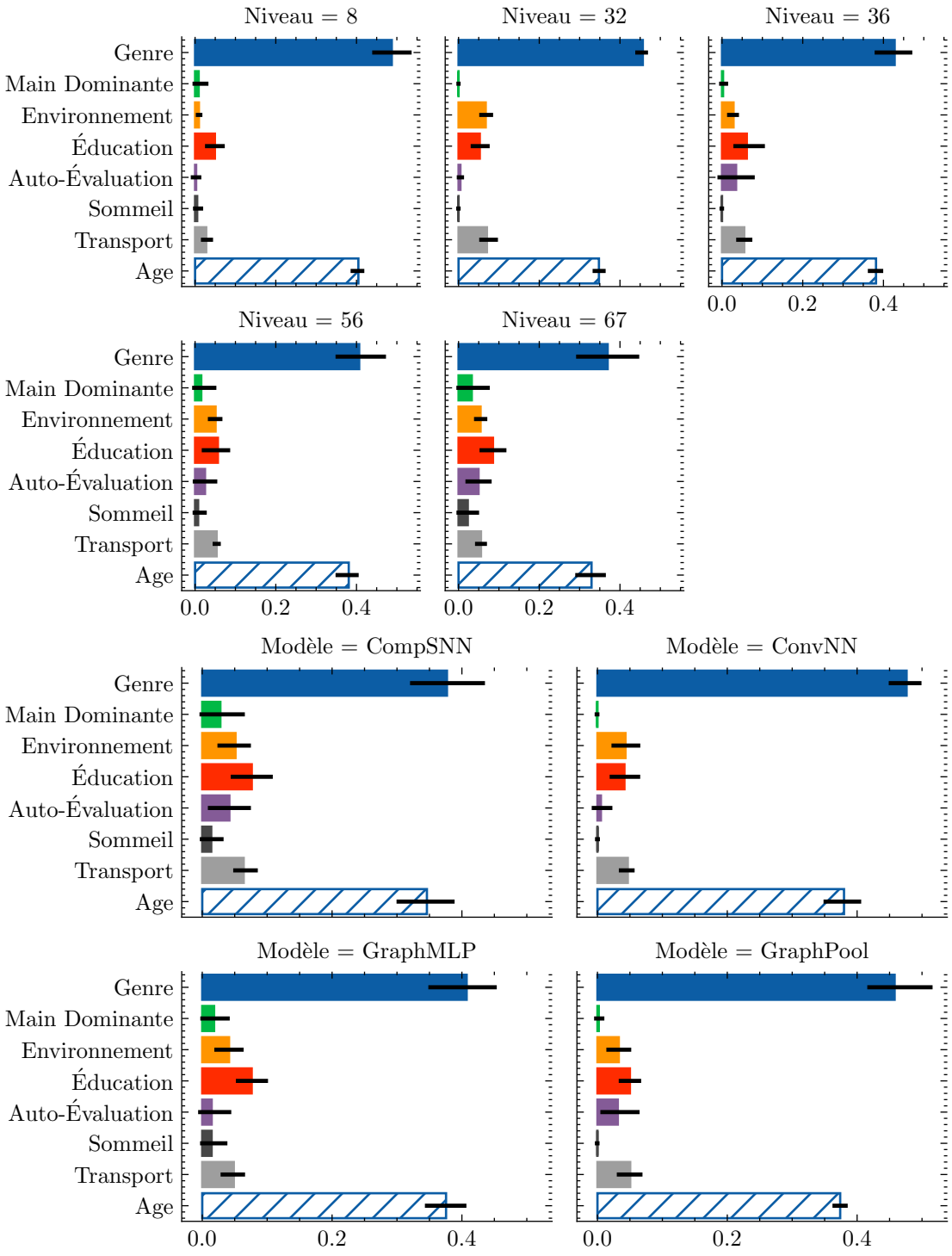


FIGURE 4.10 – Importances des différentes caractéristiques démographiques en fonction du niveau et du modèle utilisé.

plusieurs angles, afin de saisir pleinement la complexité des signaux de trajectoires.

Ces différents modèles et approches ont été évalués pour prédire les caractéristiques démographiques du joueur à partir de sa trajectoire dans un niveau donné. Les résultats de l'expérimentation nous permettent de dire que le réseau composite permet une meilleure prédiction de la démographie, et que les deux modèles spatiaux GraphPool et GraphMLP sont meilleurs que le modèle temporel ConvNN. Ils confirment également les résultats sur l'importance de chaque dimension démographique vis-à-vis du comportement spatial obtenus précédemment avec la métrique OpCorr (voir figure 4.11).

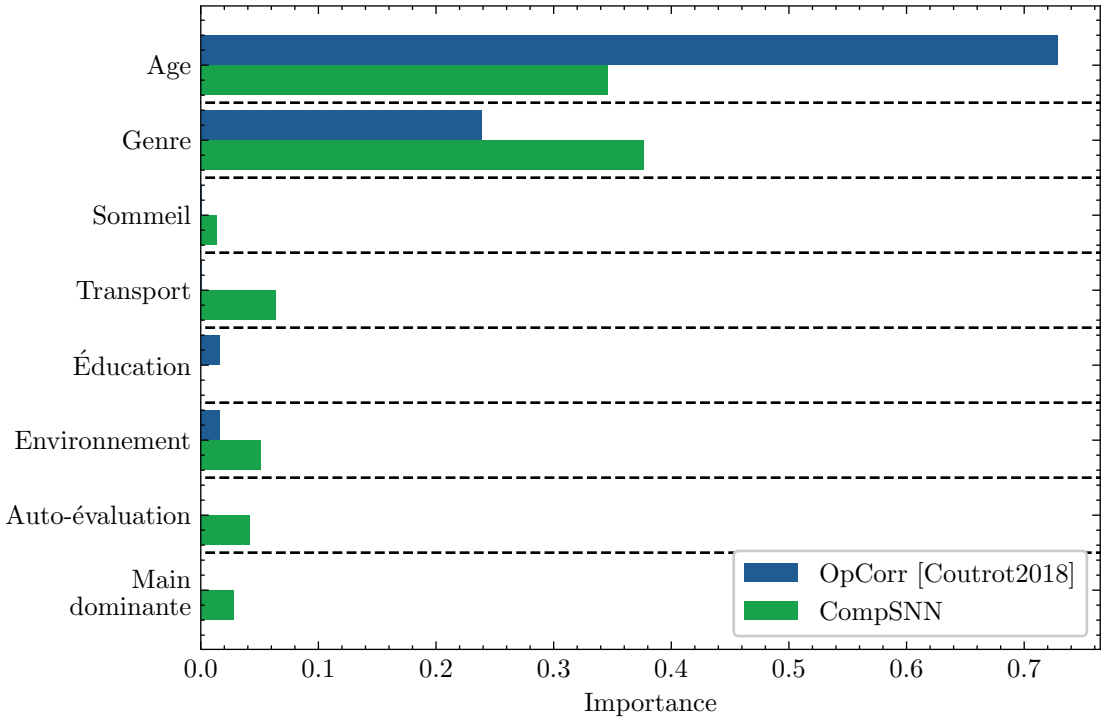


FIGURE 4.11 – Comparaison entre l'importance donnée aux différentes caractéristiques démographiques par le modèle CompSNN à celle donnée par une régression linéaire entrainée à prédire le score OpCorr.

4.6.2 Problème d'incertitude

Un problème identifié avec cette méthode est la question de *l'incertitude dans la cible*. La figure 4.9 illustre bien ce problème : assez vite durant l'apprentissage, le score moyen ne progresse plus (voir baisse légèrement). Par contre, le score des caractéristiques facilement prédites chute, au profit d'autres plus dures à prédire. On formule l'hypothèse que ce phé-

nomène est dû à une plus grande incertitude dans l'influence de certaines caractéristiques sur le comportement spatial : on demande à un modèle de prédire des caractéristiques pour lesquelles on ne sait pas si elles ont un impact sur le processus cognitif d'orientation, et, si elles en ont un, de quelle magnitude il est.

Les méthodes développées précédemment sur les données Sea Hero Quest n'avaient pas ce problème : on ne cherchait pas à prédire le profil démographique entier à partir d'une représentation de la trajectoire, on se concentrait soit sur la prédiction d'un sous ensemble réduit de caractéristiques pour lesquelles on sait déjà qu'elles ont un rôle important, soit sur la prédiction de la représentation (une métrique simple univariée dans la plupart des cas) à partir du profil démographique, ce qui permet au modèle d'apprendre à ignorer plus ou moins certaines caractéristiques démographiques.

La prédiction de la trajectoire telle quelle à partir du profil est une tâche beaucoup trop complexe, il faut donc passer par une représentation.

En résumé :

✓ L'utilisation parallèle de différentes représentations de trajectoires et de modèles adaptés permet une meilleure analyse de nos données.

✓ L'absence de plus de vérité terrain sur l'importance des différentes caractéristiques démographiques sur le sens de l'orientation nécessite de changer d'approche.

Chapitre 5

L'entropie contextuelle

Chez lui elle marche sur la tête ; il suffit de la remettre sur les pieds

— Karl Marx

L'approche décrite dans le chapitre précédent, où l'on essaie de prédire la démographie à partir du comportement, pose un problème important dans la formalisation du problème et sa solvabilité : il n'y a aucune garantie que, par exemple, être droitier ou gaucher ait un impact sur le comportement spatial, et donc qu'il soit possible de le prédire à partir d'une trajectoire. C'est d'ailleurs ce que tendent à montrer les résultats de l'expérimentation menée sur nos données, même s'il est impossible d'en tirer une conclusion définitive (notre modèle pourrait tout simplement ne pas être adapté pour l'extraction de caractéristiques liées à la préférence manuelle).

Pour répondre à ce problème, nous allons retourner la tâche. Au lieu de chercher à prédire le profil démographique du joueur à partir de son comportement, on va chercher à prédire son comportement à partir de son profil démographique.

5.1 Mesurer un comportement anormal

Chercher à prédire toute une trajectoire à partir d'un profil démographique est une tâche impossible : nos données présentent beaucoup trop de variance pour cela, les processus cognitifs impliqués dans la planification et la navigation spatiale étant complexes et influencés par beaucoup plus de facteurs que les quelques informations que nous avons. Un joueur jouant deux fois au même niveau aura potentiellement deux comportements très différents, et deux joueurs aux profils très différents peuvent parfaitement avoir un comportement très similaire.

Pour répondre à ce problème, nous allons chercher à produire une métrique qui nous permettrait de quantifier la *normalité* d'une trajectoire.

5.1.1 Définition de la normalité

Nous allons dans ce chapitre utiliser les notions de comportement normal et de comportement limite, dans un sens purement statistique de ces termes. Il ne s'agit en aucun cas d'un jugement qualitatif sur les stratégies adoptées, ni d'une intention normative de notre part. On définit comme comportement normal un comportement moyen, très présent dans les données, et comme comportement limite un comportement rare, peu représenté.

On définit un échantillon normal comme un échantillon avec une forte vraisemblance. Dans le cas de trajectoires définies dans un environnement, avec ses règles, on pourrait essayer de modéliser cet environnement et utiliser cette modélisation pour mesurer la vraisemblance des différentes trajectoires. Le problème d'une telle approche est qu'elle demanderait une modélisation complexe et une connaissance exhaustive de l'environnement, ce qui n'est pas forcément possible. Pour résoudre ce problème, nous proposons ici de partir des données de trajectoires pour apprendre une représentation de l'environnement, du *contexte*.

On définit un échantillon moyen par analogie avec la moyenne d'une distribution gaussienne : est *moyen* le point qui maximise la probabilité d'émission par la distribution. Cependant, nos données ne suivant pas une distribution gaussienne, on ne peut pas estimer la normalité d'une trajectoire en apprenant les paramètres d'une distribution gaussienne. Si on considère les coordonnées des points des trajectoires comme nos données, leur distribution est multimodale, il faut donc définir la normalité pour le cas multimodal.

On définit comme normaux les points qui forment des groupes denses dans l'espace des données, et réciproquement un point de l'espace des données est considéré comme normal s'il maximise localement la probabilité d'émission par la distribution apprise à partir des données.

En découle donc qu'un élément important de la tâche devient la définition d'une méthode pour apprendre cette distribution.

Ici, nos données étant des séries temporelles multivariées définies dans un espace bidimensionnel, on propose de définir un espace x, y, t dans lequel apprendre la distribution des points. Cela nous permet de saisir la dimension spatio-temporelle de nos données, là où par exemple la méthode utilisée pour définir les graphes dans le chapitre précédent détruisait (volontairement) la dimension temporelle en s'appliquant uniquement dans l'espace x, y.

Tous les points de toutes les trajectoires sont concaténés dans un seul vecteur de dimension $N \times 3$, avec N le nombre total de points. En pratique, et parce qu'on a un

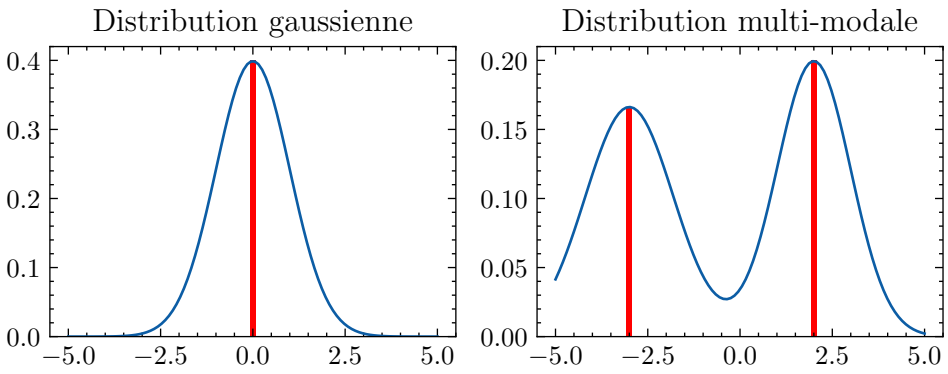


FIGURE 5.1 – Comparaison des cas normaux pour les distributions gaussiennes et multi-modales. Les barres rouges indiquent les points les plus normaux de la distribution.

nombre suffisant de points, on sous échantillonne le vecteur (voir figure 5.2).

Ensuite, on applique un algorithme d'Estimation de Densité par Noyau (KDE) sur ce vecteur. On utilise ici un noyau Epanechnikov, parce qu'il permet de réduire la complexité de l'algorithme en utilisant une optimisation via arbre k-d. On en obtient une fonction de densité $d : \mathbb{R}^3 \to [0, 1[$.

On se sert ensuite de cette fonction de densité d apprise sur un ensemble de points pour associer à chaque échantillon d'une trajectoire un score de densité p. On définit donc une série temporelle S_d à partir de la trajectoire s de L_s échantillons.

$$p(s(t)) = d(s(t)_x, s(t)_y, t); \tag{5.1}$$

$$S_d = \{p(s(t = i/L_s))\}_{i=0}^{L_s} \tag{5.2}$$

5.1.2 L'entropie contextuelle

A partir du vecteur s_d défini préalablement, on peut proposer une définition d'une mesure d'entropie d'un échantillon. La notion d'entropie est ici utilisée dans un sens large, puisqu'elle ne mesure pas la diversité d'un système, et parce que $\sum s_d \neq 1$. Elle nous permet cependant de proposer une quantification de l'information contenue dans une trajectoire, une trajectoire peu informative étant une trajectoire très probable, et inversement.

On utilise la formule de l'entropie de collision, qui est l'entropie de Rényi pour $\alpha = 2$. Elle est choisie parce que contrairement à l'entropie de Shannon, $-log(p^2)$ est strictement

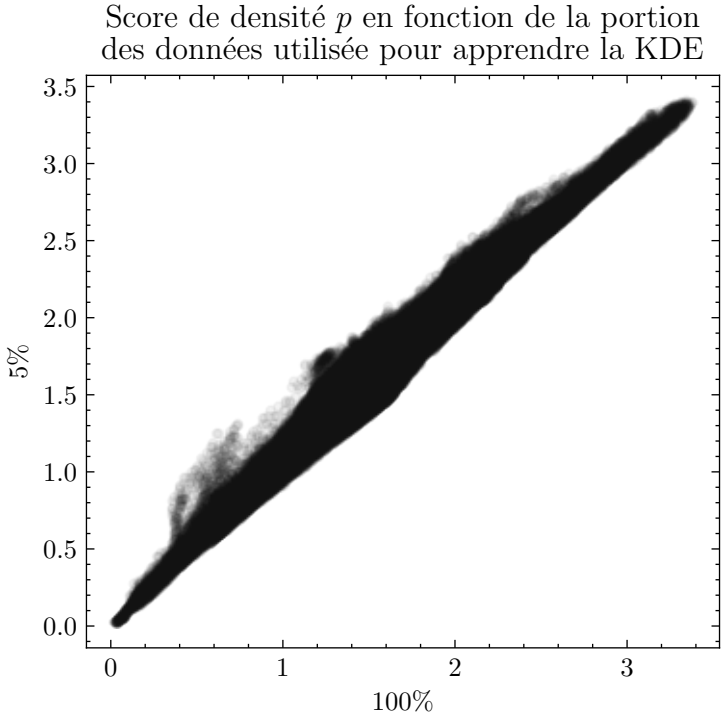


FIGURE 5.2 – Scores p calculés à partir de 1000 trajectoires du niveau 67, avec $N = 334508$. La réduction du volume de données utilisé pour apprendre la KDE à 5% des points permets de diviser le temps de calcul par 16 (de 2264 secondes à 139 secondes sur cette expérimentation) sans perdre trop d'information.

décroissant sur l'intervalle $]0, 1]$, ce qui nous permet de discriminer entre les trajectoires limites et les trajectoires normales. Si on utilisait l'entropie de Shannon, on aurait une discrimination entre d'un coté les trajectoires moyennement probables, et de l'autre regroupées les trajectoires peu et fortement probables.

$$H_2(s) = -log \int_0^1 p(s(t))^2 dt$$

Étant donné que l'on a un échantillonnage discret de la trajectoire, on a :

$$H_2(s) = -log \frac{1}{l} \sum_{i=0}^{l} p(s(i/l))^2$$

De bout en bout, on a donc une fonction qui nous permet de passer d'une trajectoire de longueur variable à un score, et qui nous permet donc de comparer entre elles ces trajectoires sans avoir à faire des mesures point à point entre chaque paire.

5.2 Détection de comportements limites à partir de la démographie

5.2.1 Discrétisation du comportement

Une fois ces scores associés à chaque trajectoire, on va chercher à les prédire à partir des informations démographiques du joueur, c'est-à-dire que l'on cherche un groupe démographique qui aurait plus tendance à présenter un comportement limite.

Le problème que l'on rencontre d'abord en définissant cette tâche, c'est la difficulté induite par la distribution des scores. En effet, puisque l'on mesure la normalité, on a une distribution asymétrique des scores. Entrainer un régresseur sur ces données a de fortes chances de nous donner une solution dégénérée qui prédirait tout vers la moyenne.

Pour éviter ce problème, on propose de discrétiser nos données, de transformer nos scores en label, en utilisant un seuil basé sur les quantiles pour assurer une distribution uniforme entre les différentes classes (voir figure 5.3). Le nombre de labels devient un hyper-paramètre du modèle.

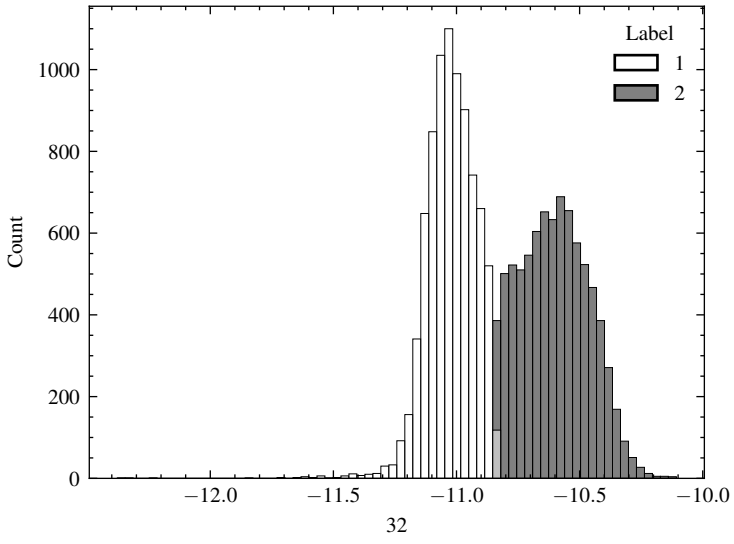


FIGURE 5.3 – Distribution des scores d'entropie contextuelle des trajectoires du niveau 32. Ici, on définit $K = 2$.

5.2.2 Expérimentation

Données Pour évaluer le modèle, on utilise le même sous-ensemble de joueurs et de trajectoires utilisé dans le chapitre précédent.

Protocole On évalue chaque configuration d'hyper-paramètres avec une validation croisée 5-plis, en utilisant le Kappa de Cohen κ comme métrique. La KDE est estimée via l'implémentation KDEpy de TreeKDE, la taille de fenêtre du noyau est fixée arbitrairement à $0,02$. Les classifieurs utilisés sont AdaBoost, Forêt Aléatoire, et Régression Logistique. On optimise les méta-paramètres (profondeur maximale pour les deux premiers, α pour la régression logistique). On mesure l'importance de chaque caractéristique démographique par permutation. On utilise ici aussi l'implémentation scikit-learn.

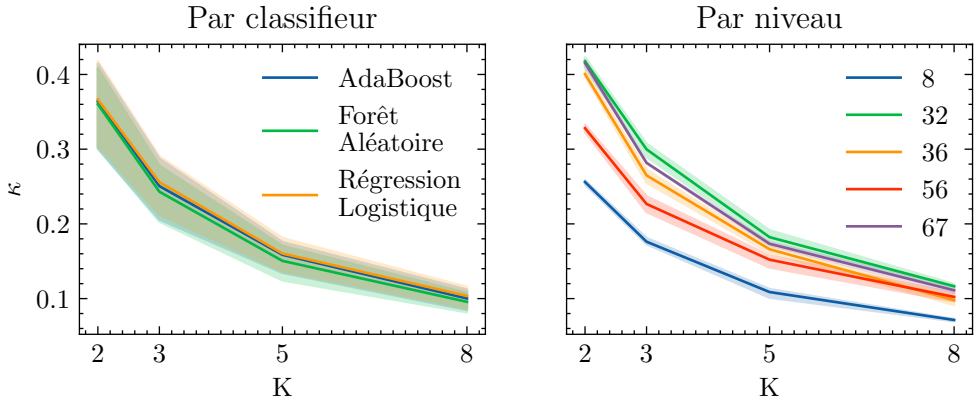


FIGURE 5.4 – Kappa de Cohen en fonction du nombre de groupes.

Résultats Comme on le voit sur la figure 5.4, les performances des modèles sont inversement corrélées au nombre de groupes. La tableau 5.1 et la figure 5.4 montrent que le choix de l'algorithme de classification n'importe que très peu. Les modèles ont par contre plus de problèmes à traiter les données des niveaux 8 et 56, ce qui peut s'expliquer par leur plus grande simplicité topologique, qui implique moins de choix par le joueur.

Les figures 5.5 et 5.6 montre que le choix du modèle et du niveau impactent peu l'importance associée à chaque caractéristique démographique. On peut donc extrapoler à partir de l'analyse du niveau 32 (choisi car ayant le meilleur κ).

On peut également visualiser les trajectoires des différents groupes, pour essayer de comprendre ce que mesure concrètement l'entropie contextuelle. La figure 5.7 montre deux exemples de trajectoires d'entropies différentes.

Niveau	Classifieur	κ
8	AdaBoost	0.2528
	Forêt Aléatoire	**0.2581**
	Régression Logistique	0.2572
32	AdaBoost	0.4211
	Forêt Aléatoire	0.4090
	Régression Logistique	**0.4228**
36	AdaBoost	0.3987
	Forêt Aléatoire	0.3980
	Régression Logistique	**0.4050**
56	AdaBoost	**0.3324**
	Forêt Aléatoire	0.3232
	Régression Logistique	0.3287
67	AdaBoost	0.4149
	Forêt Aléatoire	0.4122
	Régression Logistique	**0.4188**

TABLE 5.1 – Meilleur modèle pour chaque niveau.

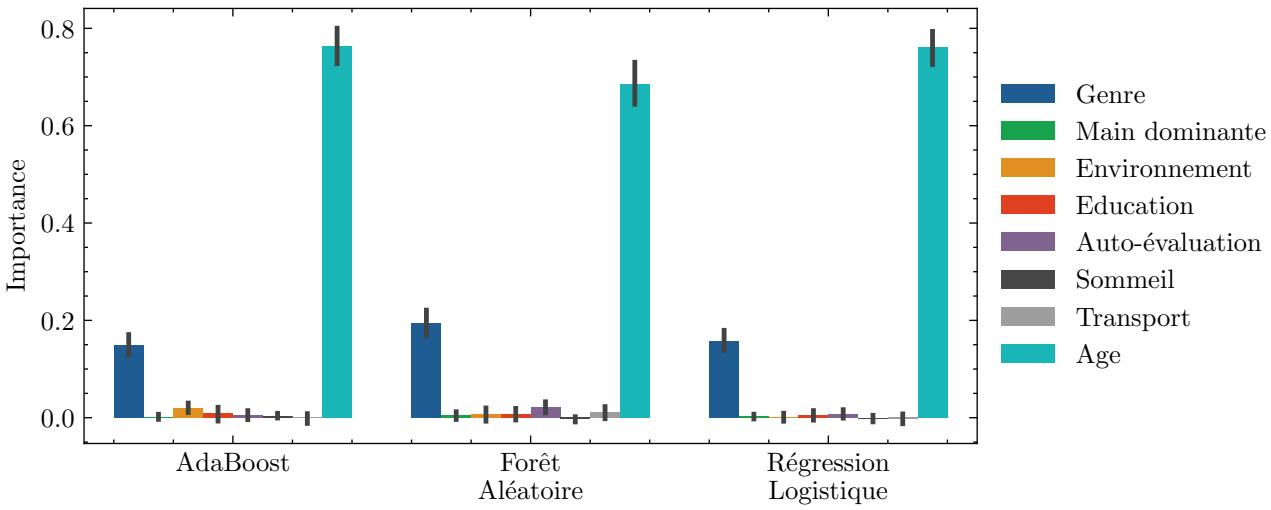


FIGURE 5.5 – Importance de chaque caractéristique démographique en fonction du modèle utilisé. Le choix du modèle ne change pas l'ordre d'importance pour les caractéristiques principales. Les barres d'erreur représentent l'intervalle de confiance à 95%.

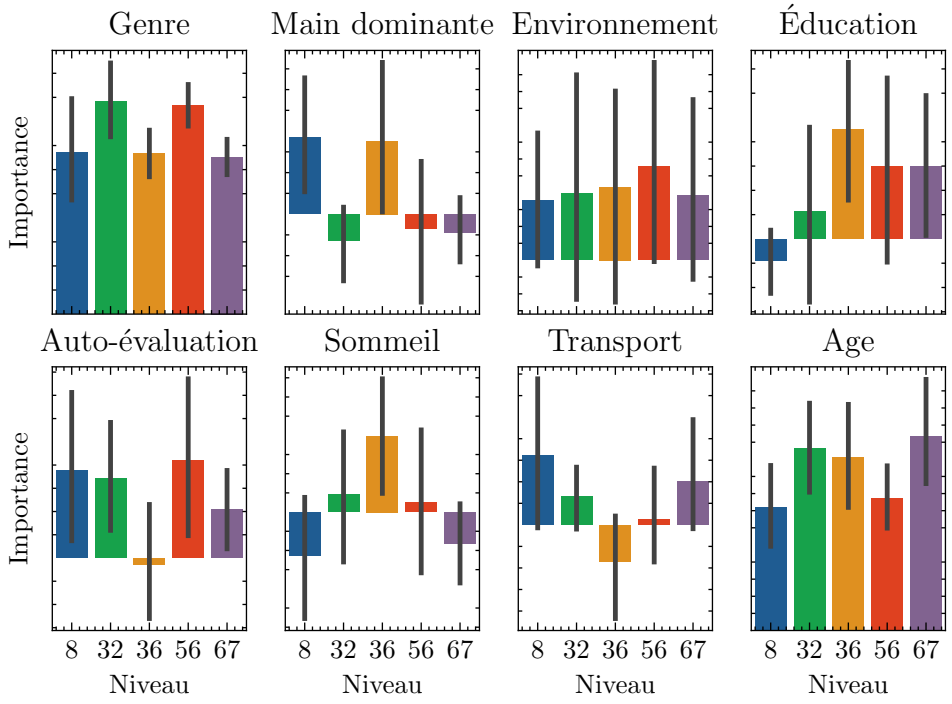


FIGURE 5.6 – Importance des caractéristique en fonction du niveau. Le choix du niveau semble avoir un impact, il serait intéressant de voir si cette différence tient à des caractéristiques spécifiques des niveaux. Les barres d'erreur représentent l'intervalle de confiance à 95%.

5.3 Conclusion

5.3.1 Apports

Dans ce chapitre, nous avons pu définir la notion d'entropie contextuelle, que nous avons utilisé pour essayer de prédire la normalité du comportement spatial à partir de la démographie.

Nous avons obtenu des résultats très satisfaisants, en obtenant un Kappa de Cohen jusqu'à 0.423 sur les données du niveau 32. En retournant le problème, nous avons réussi à grandement améliorer les performances de notre modèle par rapport à celui proposé dans le chapitre précédent, ce qui nous conforte dans l'idée que c'est la bonne direction.

Du point de vue de l'importance donnée aux caractéristiques démographiques (voir la figure 5.8), cette approche amplifie les effets observés en utilisant la métrique OpCorr : l'âge représente 80% de l'importance sur la prédiction.

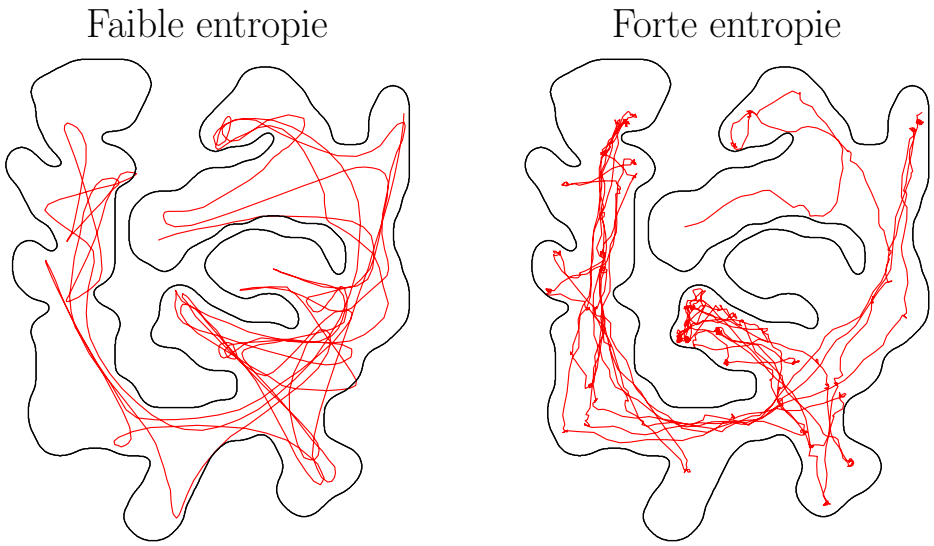


FIGURE 5.7 – Exemples de trajectoires tirées du niveau 32. On voit bien que, si la trajectoire de faible entropie n'est pas particulièrement courte et optimale, elle est beaucoup plus régulière que celle de forte entropie. On peut donc supposer que l'entropie mesure un doute "normal" chez le joueur.

5.3.2 Limitations

Cependant, si les performances du modèle sont grandement améliorées par rapport au chapitre précédent, on retrouve le même problème qu'avec les approches précédant cette thèse : La représentation de la trajectoire est fixe et définie. Si, contrairement au score OpCorr, cette métrique tire profit des données disponibles, elle reste cependant assez simple par rapport aux caractéristiques qui peuvent être extraites par les réseaux de neurones du chapitre précédent.

De plus, l'utilisation de l'estimation de la KDE impacte négativement la complexité algorithmique du modèle : comme dans plusieurs techniques de la littérature, les distances sont mesurées point à point. Même si la KDE permet des optimisations qui réduisent ce nombre de calculs de distance, le modèle n'est pas *paramétrique*.

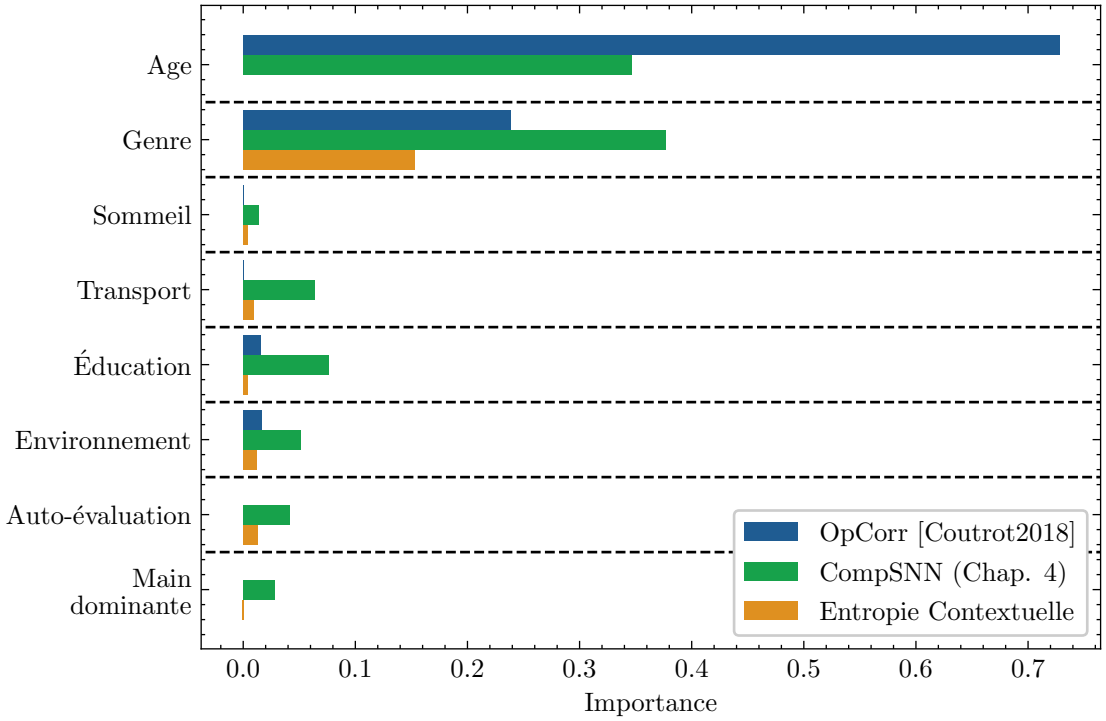


FIGURE 5.8 – Comparaison des importances attribuées à chaque caractéristique démographique. OpCorr fait référence aux résultats obtenus préalablement à partir des longueurs de trajectoires (voir section 1.4.1), CompSNN aux résultats obtenus sur le niveau 32 dans le chapitre 4

Chapitre 5en résumé :

✓ L'entropie contextuelle associe des trajectoires de longueurs variables à un score

✓ Ce score prends en compte le contexte, l'environnement dans lequel ces trajectoires s'inscrivent

✓ On peut prédire ce score à partir des caractéristiques démographiques de la personne ayant effectué la trajectoire, avec une très forte importance de l'âge.

Chapitre 6

Modèle à deux têtes

6.1 Représentation de trajectoire

Dans le chapitre précédent, nous avons proposé une méthode utilisant l'apprentissage d'une distribution dans l'espace x, y, t pour identifier des groupes de trajectoires. Cette méthode, en se basant sur un apprentissage de distribution au moyen d'un algorithme d'Estimation de Densité par Noyau, permet une modélisation fine de la distribution des données, mais a un coût de calcul qui augmente avec la taille des données. De plus, la formulation entropie associe à chaque trajectoire une valeur univariée, ce qui peut être très pratique, mais conduit à une perte d'information conséquente.

Pour répondre à ces deux problèmes, nous proposons de remplacer la KDE par un modèle paramétrique, pour remplacer l'entropie par une représentation multivariée

6.1.1 Modélisation de la distribution par mixture de gaussiennes

Pour apprendre la distribution des points de trajectoire, nous choisissons d'utiliser un modèle à mixture de gaussiennes (GMM). Nous faisons ce choix pour plusieurs raisons :

- C'est un modèle paramétrique, dont le nombre de paramètres est fixe et indépendant du nombre de points utilisés pour l'apprendre
- Une mixture de gaussiennes peut théoriquement approximer toutes les distributions avec le nombre suffisant de composantes
- On peut récupérer pour un point le score de chaque gaussienne individuellement

On définit une mixture de gaussienne comme un ensemble $\mathcal{D}$ de k gaussiennes multivariées, chacune pondérée par un poids $\omega \in [0, 1]$, et paramétrée par un centre $\mu \in \mathbb{R}^3$ et un matrice de covariance $\Sigma \in \mathbb{R}^{3\times3}$. On a donc $\mathcal{D} = \{\mu_i, \Sigma_i, \omega_i\}_{i=0}^{k}$, avec $\sum_{i=0}^{k} \omega_i = 1$. On a $(3 + 3 \times 3 + 1) \times k$ paramètres pour un modèle.

On rappelle la fonction de densité d'une distribution gaussienne multivariée :

$$p_{\mu,\Sigma}(x) = [(2\pi)^k|\Sigma|]^{-\frac{1}{2}} \times exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

On a donc la fonction de densité de la mixture de gaussienne :

$$p(x) = \sum_{i=0}^{k} \omega_i p_{\mu_i,\Sigma_i}(x)$$

Les paramètres du modèle sont appris via l'algorithme EM. Nous utilisons l'implémentation scikit-learn de l'algorithme. Le paramètre k est fixé arbitrairement pour le moment, en utilisant une valeur haute qui permette quand même à l'algorithme EM de converger vers une solution.

6.1.2 Du point à la trajectoire

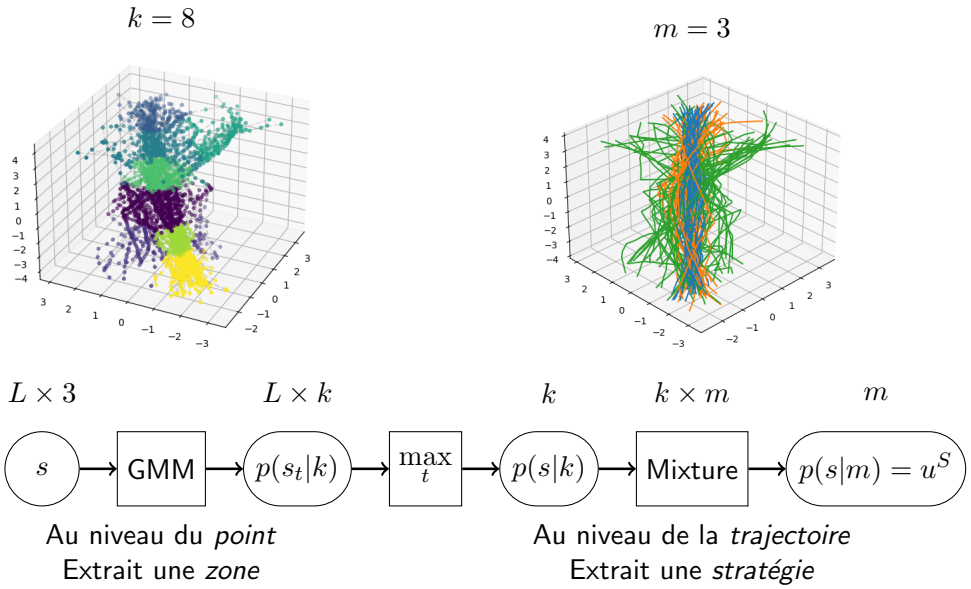


FIGURE 6.1 – Exemple de groupement obtenu.

Chaque point d'une trajectoire est représenté par un vecteur de dimension k, on a donc, pour une trajectoire de l points, une représentatione de dimension $l \times k$.

Parce que les composantes sont apprises dans l'espace spatio-temporel, on peut écraser ce vecteur le long de l'axe temporel sans en perdre l'information. On prend le maxima de chaque composante le long de la trajectoire, et obtient donc ainsi un vecteur de dimension

k pour représenter la trajectoire, ce qui nous permet de représenter des trajectoires de longueurs variées dans un même espace.

6.1.3 Groupement de trajectoire

Une fois les trajectoires transformées, on cherche à extraire des groupes de cette représentation. Pour ce faire, on va définir M modèles à mixture de gaussienne, dont les composantes sont fixes, ce sont les gaussiennes apprises préalablement à partir de tous les points, seuls les paramètres de pondération ω sont appris. Un groupement est donc un modèle de mixture de mixture de gaussiennes à $M \times k$ paramètres.

Critère d'évaluation On considère qu'une solution est bonne si elle maximise la vraisemblance des paramètres Ω. On la défini comme :

$$\max_{\Omega} \prod_{i} P(x_i|\Omega)$$

où On définit la probabilité d'un échantillon selon Ω comme la probabilité de l'échantillon selon le groupe de paramètres Ω_m qui la maximise :

$$P(x|\Omega) = \max_{m} P(x|\Omega_m)$$

Pour des raisons de stabilité numérique, on va plutôt chercher à maximiser le logarithme de la vraisemblance :

$$\max_{\Omega} \sum_{i} \log P(x_i|\Omega)$$

6.2 Groupement joint des trajectoires et des profils démographiques

L'utilisation d'une représentation multivariée pour les trajectoires nous contraint à changer de paradigme pour notre modèle : nous ne pouvons pas entrainer un modèle à prédire directement cette représentation à partir de la démographie, car il est probable que ce soit l'association de la visite de plusieurs zones dans l'espace $x; y; t$ qui soit associé aux profils démographiques.

Nous voulons donc une méthode de groupement simultané des profils démographiques et des représentation de trajectoires. Pour ce faire, il faut aussi proposer une méthode de groupement et une représentation adaptées pour les données démographiques.

6.2.1 Groupement de profils démographiques

Les informations démographiques étant sous forme catégorique, on les modélise comme un ensemble de distributions multinoulli, une par caractéristique démographique. Un groupe de profils démographiques est donc défini comme un *modèle à mixture de distributions multinoulli* de dimensions variées. La distribution associée à l'âge a la particularité d'être souple.

La probabilité d'un échantillon x selon le groupe paramétrisé par θ est définie comme :

$$P(x|\theta) = \sum_d w_d p(x_d|\theta_d)$$

Où d est une caractéristique démographique à laquelle on associe une distribution multinoulli de paramètres θ_d et de poids w_d.

Caractéristique démographique	Dimensionnalité
Genre	2
Main dominante	2
Environnement	4
Éducation	5
Auto-évaluation	4
Sommeil	3
Transport	3
Age	5

TABLE 6.1 – Dimensionnalité des distributions multinoulli associées à chaque caractéristique démographique

La tableau 6.1 donne les dimensionnalités de chacunes des composantes de la mixture. On a donc par mixture 28 paramètres.

Critère d'évaluation Une solution de groupement des profils démographiques est un modèle à $M \times 28$ paramètres qui maximise la vraisemblance des données.

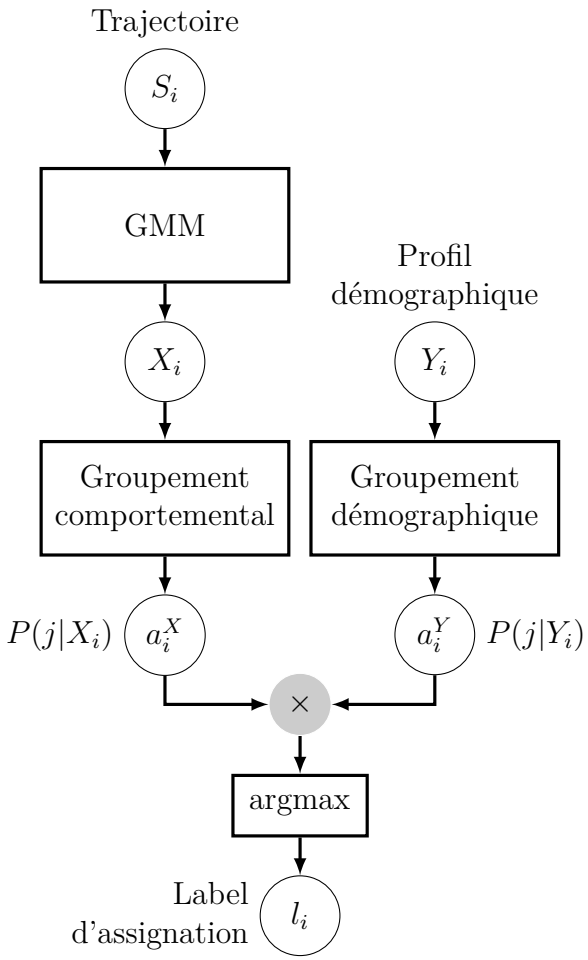


FIGURE 6.2 – Représentation du Modèle à Mélange de Distributions Mixtes (MDMM).

6.2.2 Le Modèle à Mélange de Distributions Mixtes

Les modèles proposés jusqu'ici pour les problèmes de groupement le sont de façon disjointe : on ne cherchait pas pour l'instant à faire le lien entre comportement et démographie.

Pour ce faire, on propose un nouveau modèle, chargé d'effectuer ce groupement de manière jointe, le *Modèle à Mélange de Distribution Mixtes* (MDMM) (voir la figure 6.2).

6.2.3 Optimisation des paramètres

Pour en optimiser les paramètres, on propose un algorithme de type Espérance-Maximisation (EM), inspiré des modèles à mixture (voir algorithm 1).

C'est un algorithme "faible", le critère d'évaluation n'est pas utilisé pendant l'appren-

tissage, on ne peut donc pas assurer que la solution vers laquelle il converge soit la meilleure absolument de ce point de vue. Cependant, le calcul d'assignation à un groupe qui se fait ligne 7 en multipliant la probabilité d'appartenir au j-ième groupe selon la représentation de la stratégie d'exploration avec celle obtenue selon la représentation démographique, permet d'avoir un proxy pour l'accord entre les groupes utilisé pour mettre à jour leurs paramètres. Pour biaiser l'étape de Maximisation vers notre objectif de correspondance entre les groupes comportementaux et démographiques, on pondère préférablement les points pour lesquels les deux labels sont identiques.

Dans un EM d'apprentissage de GMM classique, un échantillon est assigné à une gaussienne. Ici, un échantillon des données de comportement est un ensemble de points, il est donc assigné à un mélange de gaussiennes. Pour gérer ce cas particulier, et faciliter l'apprentissage, on introduit un paramètre τ, qui discrétise l'assignation à une gaussienne, en augmentant linéairement le seuil sur l'intervalle $[0.2, 0.8]$.

Un problème récurrent identifié durant le développement de cet algorithme était la forte chance de converger vers une solution dégénérée, où tous les échantillons finissaient assignés au même groupe. Pour palier à ce problème, la probabilité d'émission d'un échantillon X_i en fonction d'un groupe j est normalisée par le prior de ce groupe vis à vis de l'ensemble X. On fait la même chose pour les échantillons de l'ensemble Y.

6.2.4 Résultats

Avec 2 groupes

On fixe ici le paramètre du nombre de groupe $M = 2$, pour réduire le nombre de paramètres à apprendre et donc la complexité du problème tout en réduisant le risque de convergence vers une solution dégénérée. On fixe le poids des dimensions démographiques w_d à 1, pour ne pas biaiser l'apprentissage dans son importance associée à chacune et donc pouvoir analyser cette importance ensuite.

Pour évaluer le modèle proposé, on l'exécute 10 fois avec une initialisation aléatoire différente à chaque fois. La figure 6.3 donne les résultats de cette évaluation. On peut ensuite regarder les paramètres appris par les meilleurs modèles sur chaque niveau.

Estimation de l'importance des caractéristiques démographiques La figure 6.4 présente l'importance des caractéristiques démographiques dans le groupement joint. On remarque pour le genre notamment qu'une importance fortement négative lui est parfois

Algorithm 1: Apprentissage d'un Modèle à Mélange de Distribution Mixtes

Data: X l'ensemble des données de comportement, Y l'ensemble des données démographiques

Result: l les labels d'assignation aux groupes

$\omega \sim U^{M \times k}$;
$p \sim U^{M \times 28}$;
$\tau \leftarrow 0.2$;
while *pas de convergence* **do**
    `// Espérance`
    $a_i^X \leftarrow P_\omega(X_i|j)P(j|X)^{-1}$;
    $a_i^Y \leftarrow P_p(Y_i|j)P(j|Y)^{-1}$;
    $l_i \leftarrow \text{argmax}_j \, a_i^X \times a_i^Y$;
    `// Maximisation`
    if $\text{argmax}_j \, a_i^X = \text{argmax}_j \, a_i^Y$ **then**
        $w_i \leftarrow 1$;
    else
        $w_i \leftarrow \frac{1}{2}$;
    end
    $n_j = \sum w_{\{l_i=j\}}$;
    $\omega_j \leftarrow \frac{1}{n_j} \sum (w \times (X > \tau))_{\{l_i=j\}}$;
    $p_j \leftarrow \frac{1}{n_j} \sum (w \times Y)_{\{l_i=j\}}$;
    $\tau \leftarrow \tau + \dfrac{0.6}{maxiter}$;
end

donnée. La figure 6.5 montre que cette importance négative est corrélée à un moins bon score. On en déduit que le genre est donc bien une caractéristique importante.

Pour mesurer l'importance finale donnée à chaque caractéristique par cette approche, nous pondérons la contribution de chaque modèle par son score, en retirant les modèles ayant un score négatif pour ne pas avoir de déviation standard négative, ce qui nous empêcherait de calculer l'intervalle de confiance. La figure 6.6 présente les résultats.

Avec 5 groupes

On suppose que la dominance du genre est causée par sa binarité dans nos données : puisqu'on cherche extraire deux groupes, la caractéristique démographique qui présente déjà deux groupes est une solution plus simple pour le modèle. On va donc répéter l'expérimentation avec $M = 5$, en supposant que ce changement entrainera une plus forte importance donnée à une caractéristique de dimension 5, comme montré sur la figure 6.7.

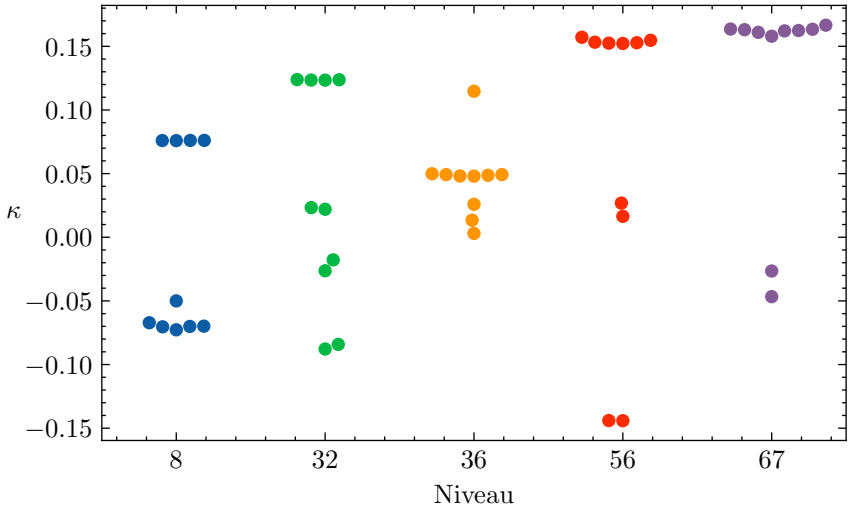


FIGURE 6.3 – Évaluation du modèle sur 10 itérations par niveau. Chaque point représente le score d'une itération. On remarque plusieurs points de convergences.

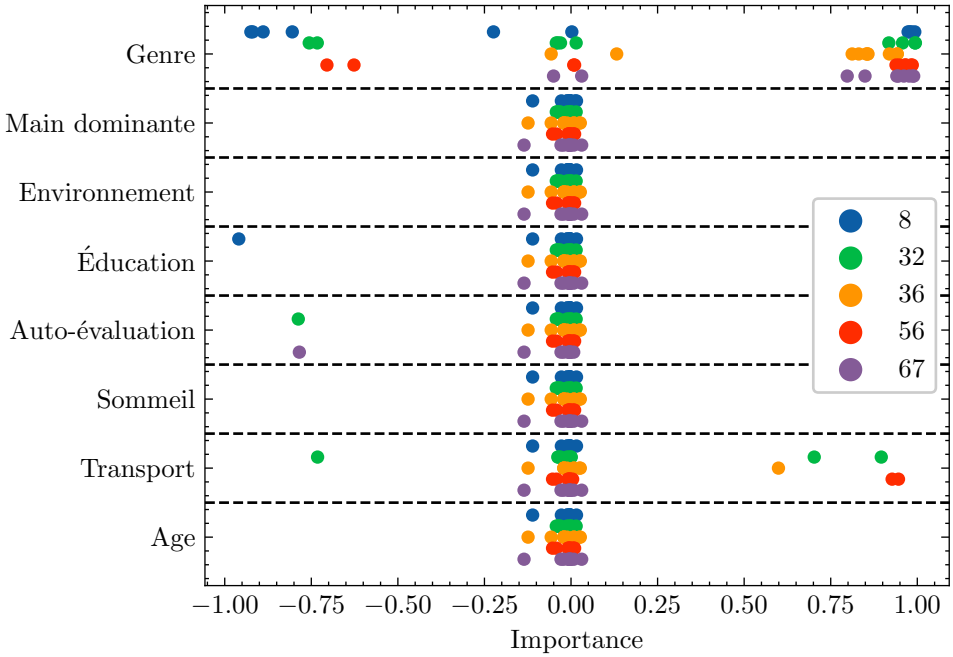


FIGURE 6.4 – Importances des caractéristiques démographiques en fonction du niveau. On remarque une très forte dominance du genre, qui présente une distribution non-normale. On remarque aussi que le choix du niveau a un impact sur l'importance des caractéristiques, le niveau 8 par exemple donne parfois une forte importance négative au genre.

Si on répète maintenant la mesure de l'importance moyenne pondérée, on remarque que le niveau d'éducation, qui est une caractéristique de dimensionnalité 5, augmente en

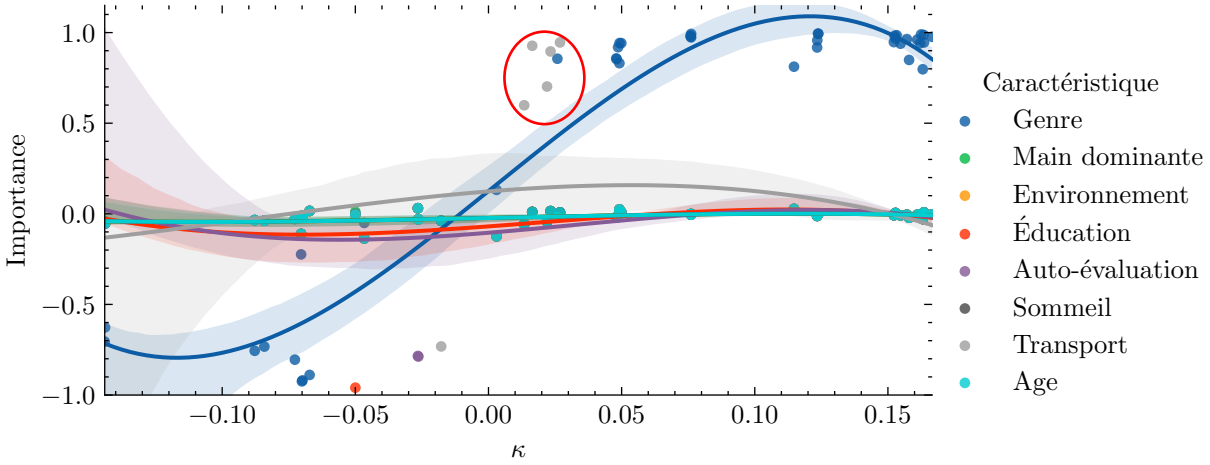


FIGURE 6.5 – Influence de l'importance donnée aux caractéristiques sur la performance des modèles. Le cercle rouge montre un ensemble de modèles qui font figure d'outliers étant donné l'importance qu'ils donnent à la caractéristique *Transport*. Cette forte importance n'est pas corrélée à un meilleur score.

importance (voir figure 6.8)

Avec 7 groupes

Pour réduire l'influence de la dimensionnalité des caractéristiques sur leurs importances, on répète maintenant le même processus avec 7 groupes (premier nombre premier

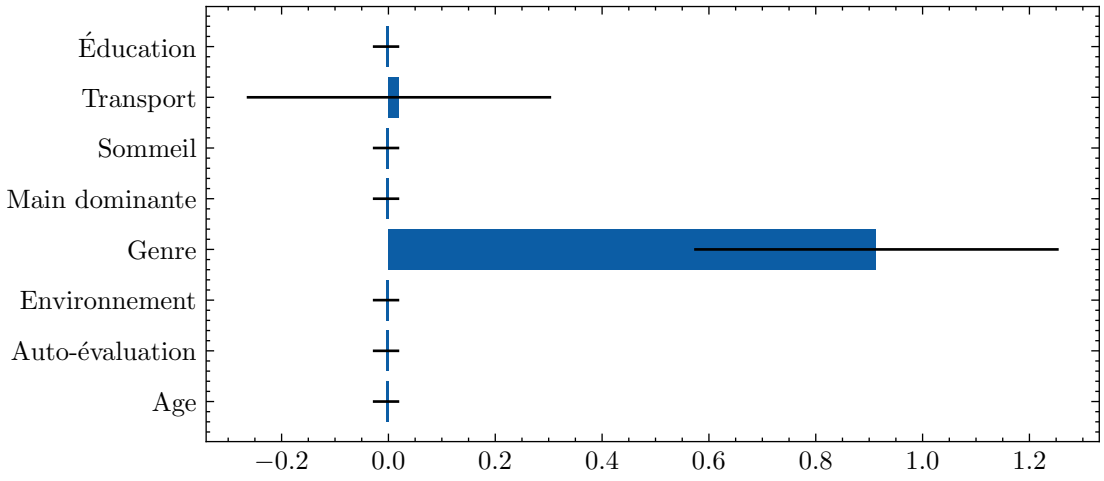


FIGURE 6.6 – Importance moyenne des caractéristiques démographiques. La barre d'erreur représente l'intervalle de confiance à 95%.

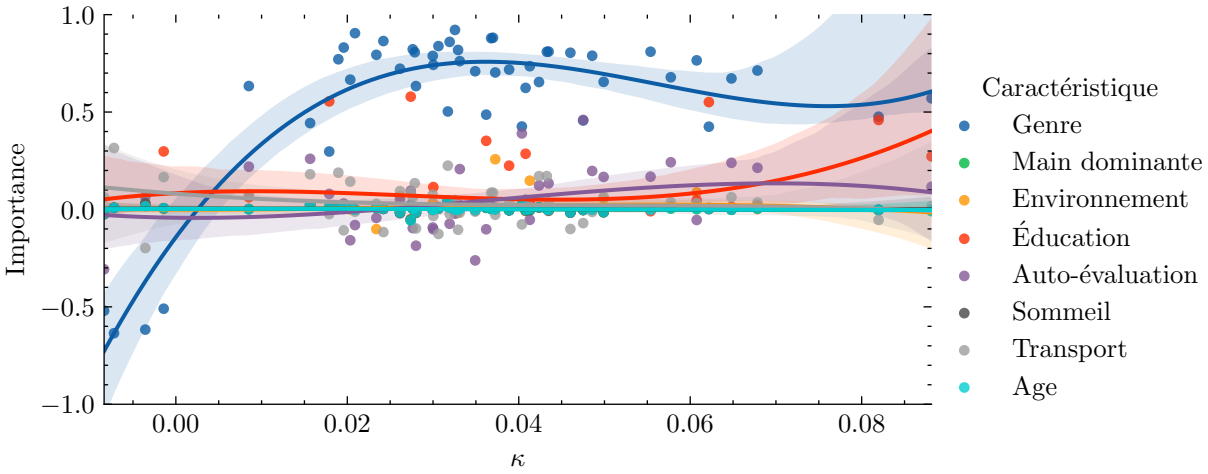


FIGURE 6.7 – Influence de l'importance donnée aux caractéristiques sur la performance des modèles avec $M = 5$. On remarque qu'une augmentation en performances s'accompagne d'une baisse de l'importance du genre et d'une augmentation de l'importance du niveau d'éducation.

strictement supérieur à la dimensionnalité maximale).

Là encore, l'importance du genre baisse au profit de celles de l'éducation, du transport, et du type d'environnement, comme montré par la figure 6.9.

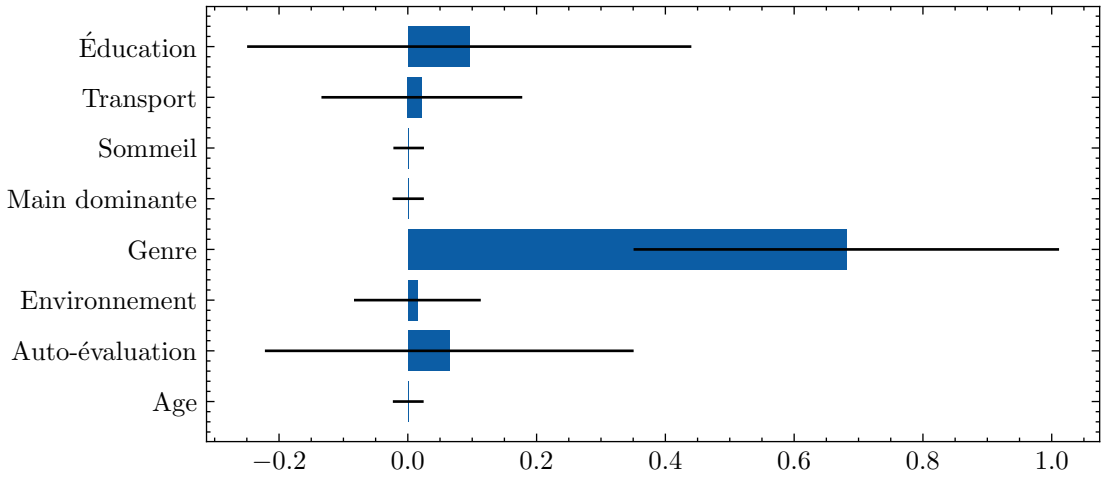


FIGURE 6.8 – Importance moyenne des caractéristiques démographiques avec $M = 5$. La barre d'erreur représente l'intervalle de confiance à 95%.

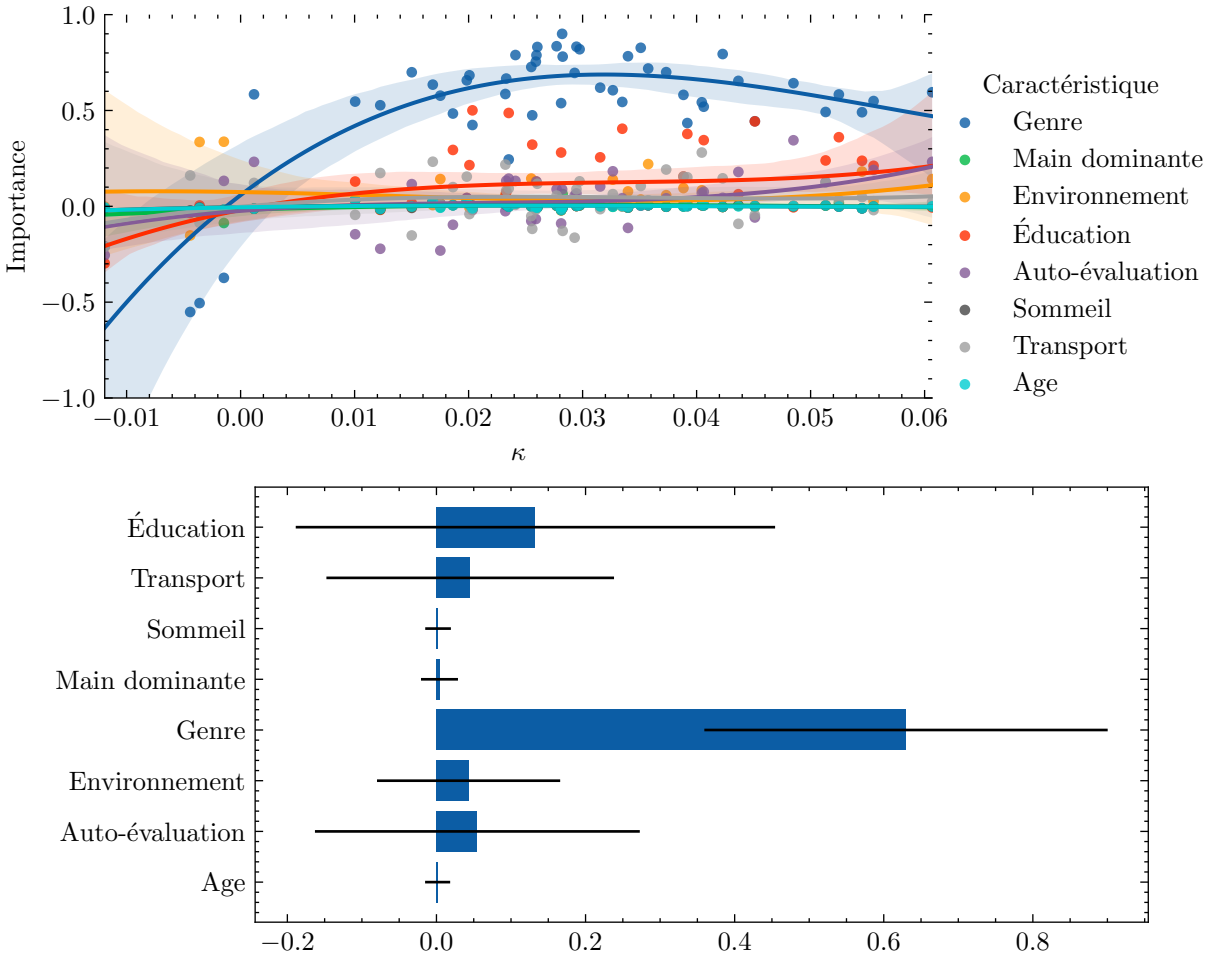


FIGURE 6.9 – Influence de l'importance donnée aux caractéristiques sur la performance des modèles et importance moyenne pondérée avec $M = 7$.

6.3 Conclusion

6.3.1 Apports

Dans ce chapitre, nous avons proposé une méthode de groupement joint ainsi que les représentations de trajectoires et de profils démographiques adaptées.

Le modèle ainsi proposé montre une importance principale du genre, et une importance secondaire du niveau d'éducation corrélée au nombre de groupes. Une particularité des résultats de ce chapitre est l'importance nulle associée à l'âge, qui est la caractéristique la plus importante selon les résultats obtenus avec OpCorr ou l'Entropie Contextuelle.

On formule l'hypothèse que l'importance plus basse de l'âge avec CompSNN, qui est

nulle avec MDMM, tient de la spécificité de ces approches d'intégrer une représentation spatiale des trajectoires, ce qui nous fait penser que le genre influe sur la stratégie adoptée (par où je passe), alors que l'âge influe sur la capacité à la réaliser (combien de temps je mets à le faire).

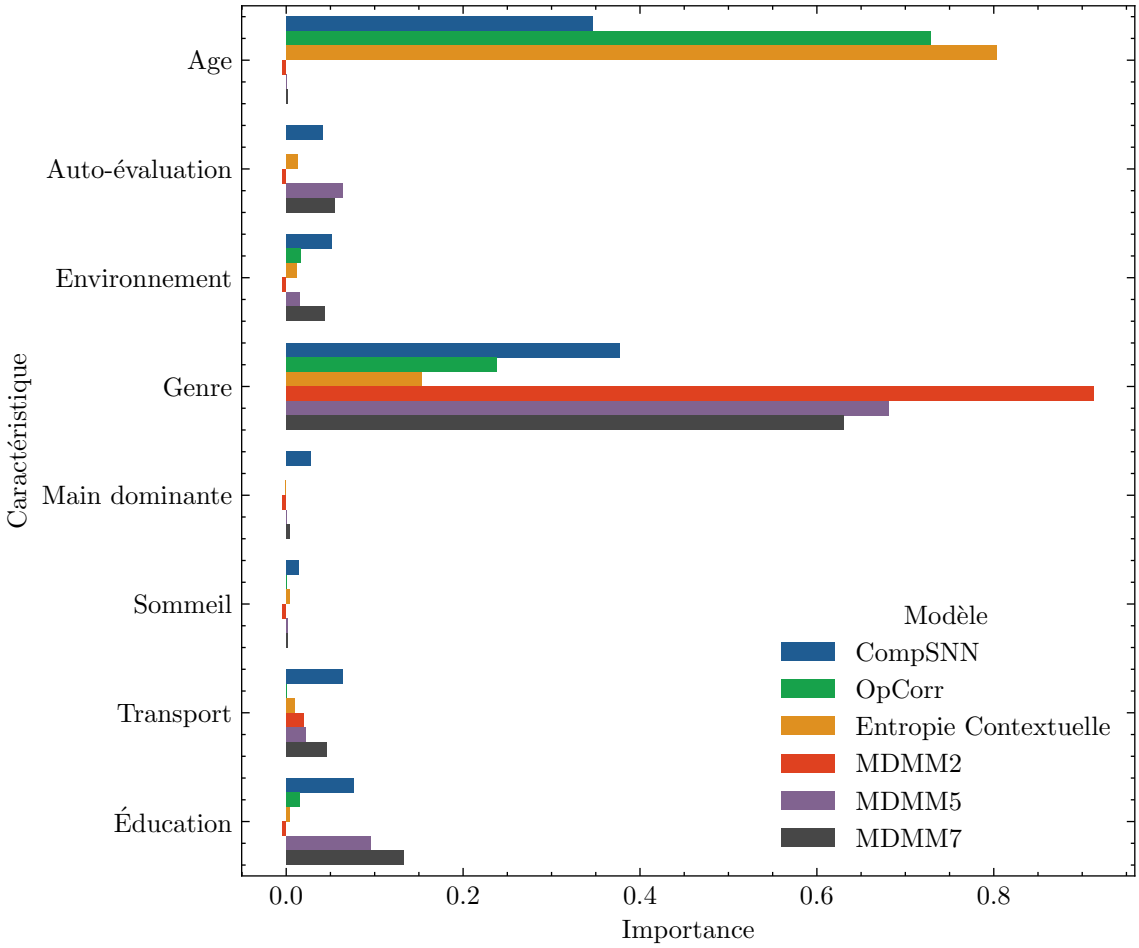


FIGURE 6.10 – Comparaison des importances attribuées à chaque caractéristique démographique. OpCorr fait référence aux résultats obtenus préalablement à partir des longueurs de trajectoires (voir section 1.4.1), CompSNN aux résultats obtenus sur le niveau 32 dans le chapitre 4, Entropie Contextuelle aux résultats obtenus dans le chapitre 5, et le nombre à la suite de MDMM au nombre de groupes cherchés.

6.3.2 Limitations

L'algorithme proposé dans ce chapitre est une ébauche qui nous sert de preuve de concept. Les résultats de l'expérimentation, particulièrement en fixant le nombre de

groupes à deux, indique un problème de convergence vers des solutions dégénérées, avec un score négatif (voir figure 6.3).

Un développement direct de l'algorithme MDMM proposé ici consisterait à y intégrer la phase de groupement des points de trajectoire. En effet, le résultat dépend directement d'une représentation fixe, qui n'est pas intégralement apprise a-priori de la démographie. L'intégration "bout en bout" d'un traitement de la trajectoire pourrait permettre de répondre à ce problème.

Une piste pour répondre à ces deux problèmes pourrait être l'utilisation d'un méthode à descente de gradient pour l'optimisation des paramètres, qui permettrait l'intégration de modules à réseaux de neurones pour traiter la trajectoire. Cette piste a été approchée, mais n'a pour l'instant pas porté ses fruits, l'optimisation convergeant quasi-systématiquement vers une solution dégénérée qui groupait tous les échantillons dans le même groupe. Plus de recherche est nécessaire pour approfondir cette piste.

Un autre aspect qui n'a pas été exploré est la pondération des différentes caractéristiques démographiques dans l'algorithme de groupement. Puisque nous ne voulions pas biaiser le modèle dans l'identification de l'importance des caractéristiques, nous en avons fixé les poids à 1, ce qui peut expliquer le score plutôt faible que nous obtenons : en effet, les dimensions démographiques identifiées comme peu importantes a posteriori induisent un bruit que l'algorithme ne peux pas filtrer.

En résumé :

✓ On peut effectuer le groupement de deux ensembles de données de manière jointe, il n'est plus nécessaire de choisir un sens pour nos modèles.

✓ La représentation des trajectoires d'un point de vue spatial favorise l'importance du genre par rapport à l'age, ce qui pourrait indiquer que l'âge modifie la capacité à exécuter une stratégie plutôt que sa conception.

✓ Il reste à développer cette approche pour la rendre plus robuste et intégrer l'apprentissage automatique du poids des caractéristiques démographiques et des composantes de la mixture de gaussienne utilisée pour représenter les trajectoires.

Chapitre 7

Conclusion et perspectives

À la fin du chapitre 2, nous dressions un cahier des charges pour les méthodes à définir :

- Tenir compte de la nature spatio-temporelle des données de trajectoire
- Permettre d'évaluer l'impact des différentes caractéristiques démographiques sur le comportement spatial
- Permettre d'identifier des groupes homogènes d'un point de vue comportemental et démographique.

Nous allons à présent évaluer si nous avons pu répondre à ce cahier des charges, comment nous l'avons fait, et quelles perspectives sont ouvertes à partir de là sur ces différents axes.

7.1 Modélisation spatio-temporelle du comportement

Ce point représentait un des enjeux techniques principaux de ce travail de recherche. Notre revue de l'état de l'art nous a amenée à juger que les techniques développées jusqu'à maintenant pour l'analyse de trajectoires choisissait un point de vue, soit spatial, soit temporel, et espérait que les modèles traitant ces données soient assez bons pour reconstruire d'une certaine façon l'autre point de vue quand celui ci était nécessaire à la réalisation de la tâche.

Dans le cas de l'analyse des données de trajectoires issues de Sea Hero Quest, nous ne savions pas par quel angle prendre le problème, et la faible variance dans les données rendait la tâche déjà suffisamment compliquée pour que nous ne puissions espérer que les modèles reconstruisent d'eux mêmes une partie de l'information que nous leur retirions.

Il a donc fallu proposer des modèles capables d'analyser les données de trajectoires aussi bien sous l'angle spatial que temporel.

7.1.1 Traitement parallèle spatial/temporel

La première méthode explorée, présentée dans le chapitre 4, propose d'utiliser une architecture de modèle à réseaux de neurones composée de plusieurs sous-modules mis en parallèle. Chacun de ces modules est chargé d'analyser la trajectoires sous un angle donné et d'en produire une représentation. Ces représentations sont ensuite concaténées et fournies à un module de sortie chargé de répondre à la tâche à partir d'elles, ici la prédiction du profil démographique du joueur ayant produit la trajectoire.

Cette approche nous a permis de confirmer l'intérêt de la concaténation de ces différentes représentations, puisque leur utilisation conjointe permet de mieux prédire toutes les caractéristiques démographiques, là où un modèle entrainé sur une seule représentation avait plus tendance à se spécialiser sur une caractéristique.

7.1.2 Extraction de points d'intérêts

Cette approche est présente dans toutes les méthodes proposées ici. Que ça soit dans le chapitre 4, où on essaie de segmenter la carte du niveau pour projeter les trajectoires sur un graphe, dans le chapitre 5, où on essaie de mesurer la probabilité qu'un joueur se soit trouvé à un endroit donné du niveau à un instant donné de sa navigation, ou dans le chapitre 6, où on utilise l'information de visite d'une zone de la carte à un instant donné pour grouper les trajectoires, nous avons essayé d'utiliser la densité de visite de l'espace des niveaux pour en analyser les trajectoires.

Un apport des chapitres 5 et 6 est d'avoir proposé d'utiliser l'information temporelle et l'information spatiale simultanément, comme dimensions des données de trajectoire, maintenant définies dans un cube et non plus sur un plan.

La différence des résultats du chapitre 6 par rapport aux analyses précédent la thèse et aux résultats des deux autres approches confirme la pertinence de cette modélisation, qui mériterait d'être approfondie.

7.2 Influence de la démographie sur le comportement spatial et l'orientation

L'utilisation de la mesure de l'importance par permutation (comme définie dans le section 3.2.1) nous a permis de comparer une variété de modèles aux mécanismes différents. Grâce à ça, nous avons pu comparer l'importance relative donnée à chacune des

caractéristiques démographiques présentes dans nos données : la figure 7.1 présente ces résultats.

Par rapport à l'analyse utilisant la métrique OpCorr (voir section 1.4.1), les approches que nous avons proposées permettent

- d'en confirmer les résultats, comme c'est le cas avec l'utilisation de l'Entropie Contextuelle définie dans le chapitre 5, en donnant une forte importance à l'âge
- de les compléter, avec le modèle CompSNN défini dans le chapitre 4, en augmentant l'importance donnée au genre,
- ou de montrer une dynamique toute autre, comme avec le Modèle à Mixtures de Distributions Mixtes présenté dans le chapitre 6, en donnant une importance nulle à l'âge.

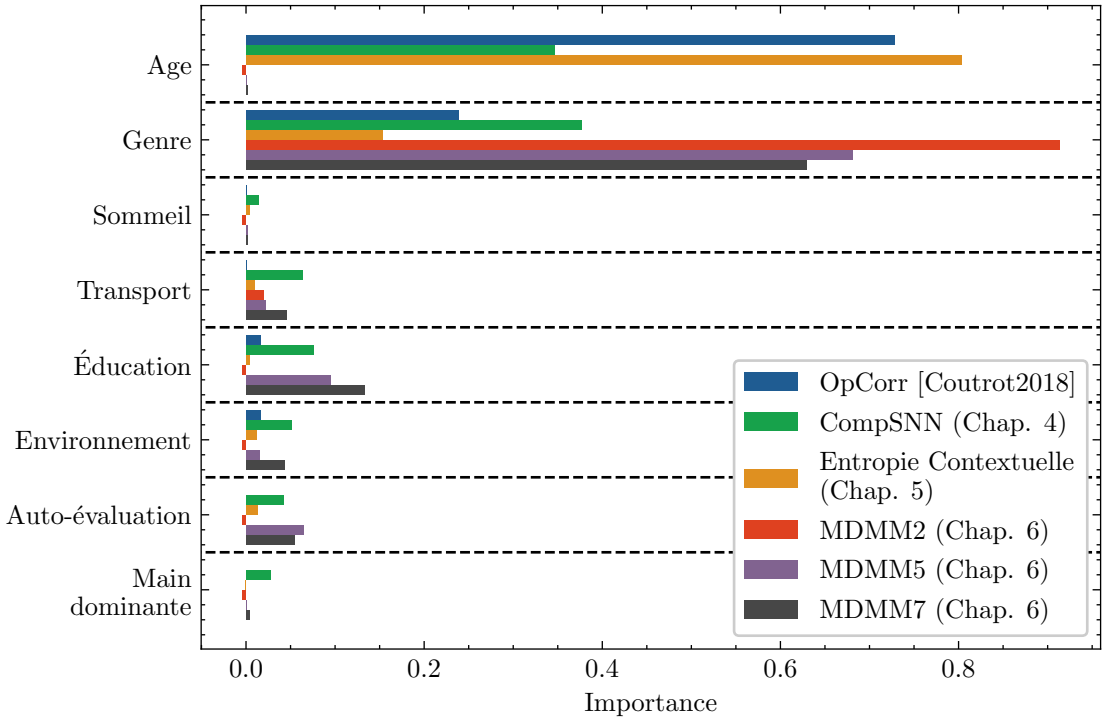


FIGURE 7.1 – Importances données aux différentes caractéristiques par tous les modèles comparés dans cette thèse.

De plus, le développement d'une méthode de groupement joint du comportement et de la démographie dans le chapitre 6 ouvre des perspectives sur l'analyse de ces données, en permettant de gérer le problème posé par la faible quantité d'information dans les données.

7.2.1 Interprétation

Nous formulons l'hypothèse que ces différences d'importance données aux différentes caractéristiques démographiques, en particulier avec l'utilisation du modèle MDMM, montrent une différence dans l'impact de ces caractéristiques sur le comportement :

- l'âge aurait plutôt une influence sur l'application d'une stratégie d'orientation, en modifiant la vitesse d'exécution, le contrôle fin de la direction, etc.
- alors que le genre impacterait plutôt la formulation de la stratégie, en modifiant le choix des endroits par lesquels passer.

7.3 Limitations et perspectives

Il y a un rapport généalogique entre les différentes approches proposées dans ce manuscrit : chaque proposition fait suite à la précédente, en essayant de répondre aux problèmes posés par elle. Les conclusions que nous tirons ici ne sont donc pas spécifiques à l'une d'entre elles, mais s'inscrivent dans ce même processus itératif.

7.3.1 Extraction des points d'intérêts et définition de graphe

Les méthodes utilisées pour extraire les points d'intérêts à partir des trajectoires présentent plusieurs problèmes :

- la méthode proposée dans le chapitre 4, qui utilise une segmentation de la carte à partir de la distribution des points de trajectoire dans le plan xy, nécessite l'ajustement manuel de nombreux paramètres pour chaque carte,
- les méthodes utilisées dans les chapitres 5 et 6 ne nécessitent pas l'ajustement de nombreux paramètres, mais sont particulièrement lourdes et couteuses en calcul. Ceci rend compliqué de les définir dynamiquement, et ne permet pas de construire un graphe à partir des points identifiés.

Nous pensons que les pistes suivantes pourraient permettre de dépasser ces limitations.

Extraction de graphe par squelettisation

La squelettisation[75-77] permet de produire une représentation compacte d'une image tout en en conservant la topologie, ce qui la rend particulièrement propice à l'analyse de données définies dans un espace non-euclidien. On peut l'appliquer sur la distribution

des points visités par les joueurs dans l'espace x, y, t. Cette approche perd l'information de distribution, mais permet d'extraire des zones sémantiques de la carte : les courbes du squelette en sont les couloirs, et les points où se rejoignent ces courbes en sont les croisements. On peut construire un graphe à partir du squelette, avec pour arêtes les courbes, et pour nœuds les points de croisement.

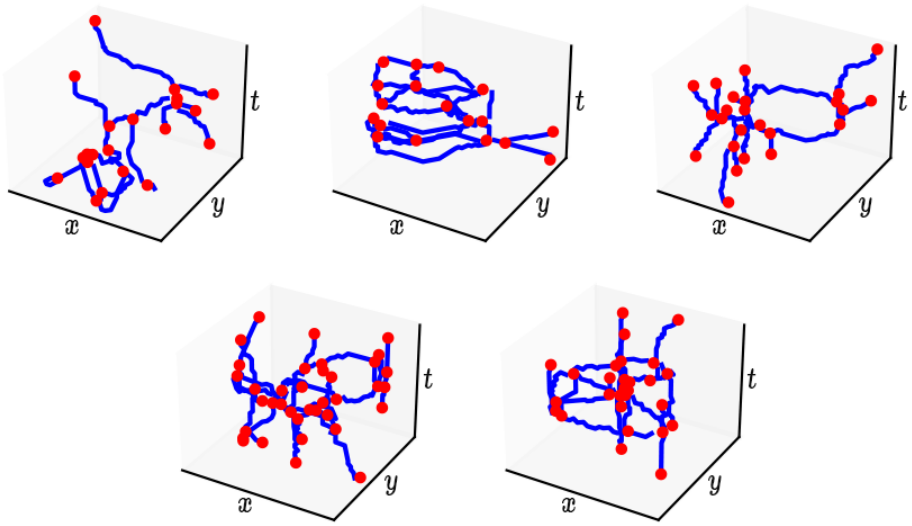


FIGURE 7.2 – Exemple de graphes construits à partir du squelette de la carte du niveau dans un espace xyt. Les points rouges représentent les croisements, les lignes bleues les couloirs.

Apprentissage automatique d'un mélange hiérarchique de gaussiennes

Nous proposions dans le chapitre 4 une méthodes de pooling sur graphe et de réseau de neurones hiérarchique, et dans le chapitre 6 l'utilisation de modèle à mixture de gaussiennes.

L'hybridation de ces deux approches pourraient permettre d'utiliser des techniques de traitement du signal sur graphe, facilitées par la structure hiérarchique du graphe, au prix d'une perte sémantique sur l'interprétation du graphe qui serait possible avec l'utilisation d'une squelettisation.

Nous l'envisageons de la manière suivante :

– On définit un arbre binaire, aux feuilles duquel on assigne une gaussienne définie dans l'espace xyt des trajectoires. On peut ensuite utiliser la méthode de pooling

sur graphe pour projeter la trajectoire sur ces feuilles, dont le signal est ensuite aggloméré de manière hiérarchique en remontant l'arbre, jusqu'à en obtenir une représentation unique à la racine.

– Les centres des gaussiennes associées aux feuilles sont des paramètres du modèle, ils évoluent donc au fur et au mesure de l'apprentissage, ce qui nous permet d'apprendre une mixture de gaussiennes pertinente par rapport à la tâche à résoudre.

7.3.2 Groupement joint des données démographiques et de trajectoire

Dans le chapitre 6, nous proposions l'algorithme MDMM pour grouper conjointement les profils démographiques et les trajectoires des joueurs. Pour ça, nous présentions un algorithme de type espérance-maximisation. Nous avons choisi cette classe d'algorithme car elle nous permet d'utiliser des opérations qui cassent le gradient, et car elle est relativement simple. Cependant, elle présente quelques problèmes :

– la mesure de vraisemblance utilisée ne mesure pas strictement l'accord entre les deux aspects du groupement,

– l'algorithme est très sensible à l'initialisation des poids,

– la convergence arrive très vite, trop vite.

Une solution pour répondre à ces problèmes pourrait être le développement d'un modèle à réseau de neurones spécifique et l'utilisation d'une méthode d'optimisation à descente de gradient. Si nous ne l'avons pas fait ici, c'est parce que l'utilisation de réseaux de neurones pour l'apprentissage non-supervisé pose plusieurs problèmes, notamment du fait qu'on ne peut pas utiliser de fonction discrétisante. Mais plusieurs méthodes d'apprentissage non-supervisé profond avec maximisation d'information[78, 79]ont été proposées, ce qui nous laisse penser qu'une telle approche est possible et nécessite juste d'être approfondie pour être appliquée à notre problème.

Points à développer :

✓ Développer de nouvelles méthodes d'extraction de graphes pour la représentation de données spatio-temporelles

✓ Améliorer l'algorithme MDMM et penser une optimisation par méthode de gradient

✓ Proposer un modèle sémantique capable d'exprimer ce qui caractérise le comportement d'un groupe démographique donné.

BIBLIOGRAPHIE

1. Bonnes, M. & Carrus, G., in *Encyclopedia of Applied Psychology* 801-814 (Elsevier, 2004), https://doi.org/10.1016/b0-12-657410-3/00252-x.
2. Evans, D. J. & Herbert, D. T., *Behavioural geography and criminal behaviour* 1989, https://www.ojp.gov/ncjrs/virtual-library/abstracts/behavioural-geography-and-criminal-behaviour-geography-crime-p-161.
3. Stanley, K., *Spatial Behaviour | Kevin Stanley* 2014, https://www.cs.usask.ca/faculty/kgs325/monitoring-human-behaviour/spatial-behaviour.html.
4. Tang, B., Yiu, M. L., Mouratidis, K. & Wang, K., *Efficient Motif Discovery in Spatial Trajectories Using Discrete Fréchet Distance* in (2017).
5. Besse, P., Guillouet, B., Loubes, J.-M. & Royer, F., Review & Perspective for Distance Based Clustering of Vehicle Trajectories, *IEEE Transactions on Intelligent Transportation Systems* (mai 2016).
6. Ardakani, I., Hashimoto, K. & Yoda, K., in *Distributed, Ambient and Pervasive Interactions : Technologies and Contexts* (éd. Streitz, N. & Konomi, S.) (Springer International Publishing, 2018).
7. Lu, W., Wei, X., Xing, W. & Liu, W., Trajectory-based motion pattern analysis of crowds, *Neurocomputing* **247,** 213-223, https://doi.org/10.1016/j.neucom.2017.03.074 (juill. 2017).
8. Coughlan, G., Laczó, J., Hort, J., Minihane, A.-M. & Hornberger, M., Spatial navigation deficits — overlooked cognitive marker for preclinical Alzheimer disease ?, *Nature Reviews Neurology* **14,** 496-506, https://doi.org/10.1038%2Fs41582-018-0031-x (juill. 2018).
9. Hu, S. *et al.,* Urban function classification at road segment level using taxi trajectory data : A graph convolutional neural network approach, *Computers, Environment and Urban Systems* **87,** 101619 (2021).

10. Huang, X. *et al.,* TrajGraph : A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data, *IEEE Transactions on Visualization and Computer Graphics* **22,** 160-169, https://doi.org/10.1109/tvcg.2015.2467771 (jan. 2016).

11. Chen, C., Liao, C., Xie, X., Wang, Y. & Zhao, J., Trip2Vec : a deep embedding approach for clustering and profiling taxi trip purposes, *Personal and Ubiquitous Computing* **23,** 53-66, https://doi.org/10.1007/s00779-018-1175-9 (juill. 2018).

12. Wen Chang, H., chin Tai, Y. & jen Hsu, J. Y., Context-aware taxi demand hotspots prediction, *International Journal of Business Intelligence and Data Mining* **5,** 3, https://doi.org/10.1504/ijbidm.2010.030296 (2010).

13. Pan, G., Qi, G., Wu, Z., Zhang, D. & Li, S., Land-Use Classification Using Taxi GPS Traces, *IEEE Transactions on Intelligent Transportation Systems* **14,** 113-123, https://doi.org/10.1109/tits.2012.2209201 (mars 2013).

14. Liu, Y., Wang, F., Xiao, Y. & Gao, S., Urban land uses and traffic 'source-sink areas' : Evidence from GPS-enabled taxi data in Shanghai, *Landscape and Urban Planning* **106,** 73-87, https://doi.org/10.1016/j.landurbplan.2012.02.012 (mai 2012).

15. You, L. *et al.,* A Spatio-Temporal Schedule-Based Neural Network for Urban Taxi Waiting Time Prediction, *ISPRS International Journal of Geo-Information* **10,** 703, https://doi.org/10.3390/ijgi10100703 (oct. 2021).

16. Liu, H., Jin, S., Yan, Y., Tao, Y. & Lin, H., Visual analytics of taxi trajectory data via topical sub-trajectories, *Visual Informatics* **3,** 140-149, https://doi.org/10.1016/j.visinf.2019.10.002 (sept. 2019).

17. Castro, P. S., Zhang, D., Chen, C., Li, S. & Pan, G., From taxi GPS traces to social and community dynamics, *ACM Computing Surveys* **46,** 1-34, https://doi.org/10.1145/2543581.2543584 (nov. 2013).

18. Jamshidi, S. & Pati, D., A Narrative Review of Theories of Wayfinding Within the Interior Environment, *HERD : Health Environments Research ; Design Journal* **14,** 290-303, https://doi.org/10.1177/1937586720932276 (juin 2020).

19. O'Keefe, J. & Nadel, L., *The Hippocampus as a Cognitive Map* en (Oxford University Press, London, England, nov. 1978).

20. Lynch, K., *The image of the city* (MIT Press, London, England, jan. 1960).

21. Symonds, P., Brown, D. H. & Iacono, V. L., Exploring an Absent Presence : Wayfinding as an Embodied Sociocultural Experience, *Sociological Research Online* **22,** 48-67, https://doi.org/10.5153/sro.4185 (fév. 2017).

22. Lidwell, W. & Holden, K., *Universal principles of design, revised and updated* 2e éd. (Rockport, jan. 2010).

23. Bettio, L. E., Rajendran, L. & Gil-Mohapel, J., The effects of aging in the hippocampus and cognitive decline, *Neuroscience ; Biobehavioral Reviews* **79,** 66-86, https://doi.org/10.1016/j.neubiorev.2017.04.030 (août 2017).

24. Ghisletta, P., Rabbitt, P., Lunn, M. & Lindenberger, U., Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared, *Intelligence* **40,** 260-268 (2012).

25. Anguera, J. A. *et al.,* Video game training enhances cognitive control in older adults, *Nature* **501,** 97-101 (2013).

26. Lindenberger, U., Human cognitive aging : corriger la fortune ?, *Science* **346,** 572-578 (2014).

27. Linn, M. C. & Petersen, A. C., Emergence and characterization of sex differences in spatial ability : A meta-analysis, *Child development,* 1479-1498 (1985).

28. Reilly, D. & Neumann, D. L., Gender-role differences in spatial ability : A meta-analytic review, *Sex roles* **68,** 521-535 (2013).

29. Schwab, K. *et al., The global gender gap report 2017* en, Report, ISBN : 9781944835125 (World Economic Forum, nov. 2017), https://apo.org.au/node/208501 (2022).

30. Coutrot, A. *et al.,* Global Determinants of Navigation Ability, *Current Biology* (sept. 2018).

31. Coutrot, A. *et al.,* Entropy of city street networks linked to future spatial navigation ability, *Nature* **604,** 104-110, https://doi.org/10.1038/s41586-022-04486-7 (mars 2022).

32. Spiers, H. J., Coutrot, A. & Hornberger, M., How the environment shapes our ability to navigate, *Clinical and Translational Medicine* **12,** https://doi.org/10.1002/ctm2.928 (juin 2022).

33. Lewis, D., Memory and Intelligence in Navigation - East is a Big Bird. Thomas Gladwin. Harvard University Press (distr. Oxford), 1970, vii × 232 pp., illus., maps, £4·75. *Journal of Navigation* **24,** 423-424, https://doi.org/10.1017/s0373463300048426 (juill. 1971).

34. Qiu, C., Kivipelto, M. & von Strauss, E., Epidemiology of Alzheimer's disease : occurrence, determinants, and strategies toward intervention, *Dialogues Clin Neurosci* **11,** 111-128, issn : 1294-8322, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181909/ (2022) (juin 2009).

35. Georges, J., Miller, O. & Bintener, C., Estimating the prevalence of dementia in Europe, en, http://rgdoi.net/10.13140/RG.2.2.16880.81923 (2020).

36. Agüero-Torres, H. *et al.,* Dementia is the major cause of functional dependence in the elderly : 3-year follow-up data from a population-based study. *American journal of public health* **88,** 1452-1456 (1998).

37. Wimo, A., Winblad, B. & Jönsson, L., An estimate of the total worldwide societal costs of dementia in 2005, *Alzheimer's & Dementia* **3,** 81-91 (2007).

38. Qiu, C., De Ronchi, D. & Fratiglioni, L., The epidemiology of the dementias : an update, *Current opinion in psychiatry* **20,** 380-385 (2007).

39. Fratiglioni, L., Paillard-Borg, S. & Winblad, B., An active and socially integrated lifestyle in late life might protect against dementia, *The Lancet Neurology* **3,** 343-353 (2004).

40. *How Is Alzheimer's Disease Diagnosed ?* en, https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed (2022).

41. Coutrot, A. *et al.,* Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance (éd. Zamarian, L.) (mars 2019).

42. Pishdadian, S., Coutrot, A., Hornberger, M., Spiers, H. & Rosenbaum, S., *Big data meet deep data : Characterizing spatial navigation in hippocampal amnesia* Cognitive Neuroscience Society, Poster, mars 2021, https://hal.archives-ouvertes.fr/hal-03268852.

43. Spiers, H. J., Coutrot, A. & Hornberger, M., Explaining World-Wide Variation in Navigation Ability from Millions of People : Citizen Science Project Sea Hero Quest, *Topics in Cognitive Science,* https://doi.org/10.1111/tops.12590 (déc. 2021).

44. Shamoun-Baranes, J. *et al.,* From Sensor Data to Animal Behaviour : An Oystercatcher Example, *PLoS ONE* **7** (éd. de Polavieja, G. G.) e37997, https://doi.org/10.1371/journal.pone.0037997 (mai 2012).

45. Liu, S., Liu, Y., Ni, L. M., Fan, J. & Li, M., *Towards mobility-based clustering* in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10* (ACM Press, 2010), https://doi.org/10.1145/1835804.1835920.

46. Noureddine, H., Ray, C. & Claramunt, C., A hierarchical indoor and outdoor model for semantic trajectories, *Transactions in GIS* **26,** 214-235, https://doi.org/10.1111/tgis.12841 (oct. 2021).

47. Jin, M. & Claramunt, C., A Semantic Model for Human Mobility in an Urban Region, *Journal on Data Semantics* **7,** 171-187, https://doi.org/10.1007/s13740-018-0092-4 (août 2018).

48. Hosseinpour, M., Malek, M. R. & Claramunt, C., Socio-spatial influence maximization in location-based social networks, *Future Generation Computer Systems* **101,** 304-314, https://doi.org/10.1016/j.future.2019.06.024 (déc. 2019).

49. Antoniadis, A., Brossat, X., Cugliari, J. & Poggi, J.-M., *Clustering functional data using wavelets* Research Report RR-7515, Rapport de recherche publié : hal-00942684 (INRIA Grenoble - Rhone-Alpes, jan. 2011), 30, https://hal.inria.fr/inria-00559115.

50. Hargreaves, J. K., Knight, M. I., Pitchford, J. W., Oakenfull, R. J. & Davis, S. J., Clustering Nonstationary Circadian Rhythms using Locally Stationary Wavelet Representations, *Multiscale Modeling ; Simulation* **16,** 184-214, https://doi.org/10.1137/16m1108078 (jan. 2018).

51. He, H., Tan, Y. & Xing, J., Unsupervised classification of 12-lead ECG signals using wavelet tensor decomposition and two-dimensional Gaussian spectral clustering, *Knowledge-Based Systems* **163,** 392-403, https://doi.org/10.1016/j.knosys.2018.09.001 (jan. 2019).

52. Jiang, Z., Lin, R., Yang, F. & Wu, B., A Fused Load Curve Clustering Algorithm Based on Wavelet Transform, *IEEE Transactions on Industrial Informatics* **14,** 1856-1865, https://doi.org/10.1109/tii.2017.2769450 (mai 2018).

53. Selim, H., Prieto, M. D., Trull, J., Romeral, L. & Cojocaru, C., Laser Ultrasound Inspection Based on Wavelet Transform and Data Clustering for Defect Estimation in Metallic Samples, *Sensors* **19,** 573, https://doi.org/10.3390/s19030573 (jan. 2019).

54. Kieu, T., Yang, B., Guo, C. & Jensen, C. S., *Outlier Detection for Time Series with Recurrent Autoencoder Ensembles.* in *IJCAI* (2019), 2725-2732.

55. Yin, C., Zhang, S., Wang, J. & Xiong, N. N., Anomaly detection based on convolutional recurrent autoencoder for IoT time series, *IEEE Transactions on Systems, Man, and Cybernetics : Systems* **52,** 112-122 (2020).

56. Guo, Y. *et al., Multidimensional time series anomaly detection : A gru-based gaussian mixture variational autoencoder approach* in *Asian Conference on Machine Learning* (2018), 97-112.

57. Lubba, C. H. *et al.,* catch22 : CAnonical Time-series CHaracteristics, *Data Mining and Knowledge Discovery* **33,** 1821-1852, https://doi.org/10.1007/s10618-019-00647-x (août 2019).

58. Hosseinpoor, A., Ali Abbaspour, R., Claramunt, C. & Chehreghan, A., Inferring geometric similarities of trajectories by an abstract trajectory descriptor, *Earth Observation and Geomatics Engineering* **4,** https://doi.org/10.22059/eoge.2020.299971.1080 (juin 2020).

59. Fulcher, B. D., Little, M. A. & Jones, N. S., Highly comparative time-series analysis : the empirical structure of time series and their methods, *Journal of the Royal Society Interface* **10,** 20130048 (2013).

60. Šimon, M., Vašát, P., Poláková, M., Gibas, P. & Daňková, H., Activity spaces of homeless men and women measured by GPS tracking data : A comparative analysis of Prague and Pilsen, *Cities* **86,** 145-153, https://doi.org/10.1016/j.cities.2018.09.011 (mars 2019).

61. Cáceres, M. D. *et al.,* Trajectory analysis in community ecology, *Ecological Monographs* **89,** https://doi.org/10.1002/ecm.1350 (jan. 2019).

62. Ardakani, I., Hashimoto, K. & Yoda, K., in *Distributed, Ambient and Pervasive Interactions : Technologies and Contexts* 3-22 (Springer International Publishing, 2018), https://doi.org/10.1007/978-3-319-91131-1_1.

63. Bian, J., Tian, D., Tang, Y. & Tao, D., A survey on trajectory clustering analysis, *arXiv :1802.06971 [cs]* (fév. 2018).

64. Dendorfer, P., Osep, A. & Leal-Taixe, L., *Goal-GAN : Multimodal Trajectory Prediction Based on Goal Position Estimation* in *Proceedings of the Asian Conference on Computer Vision (ACCV)* (nov. 2020).

65. Wang, S., Cao, J. & Yu, P. S., Deep Learning for Spatio-Temporal Data Mining : A Survey (juin 2019).

66. Guo, D., Liu, S. & Jin, H., A graph-based approach to vehicle trajectory analysis, *Journal of Location Based Services* **4,** 183-199, https://doi.org/10.1080/17489725.2010.537449 (sept. 2010).

67. Ricaud, B., Borgnat, P., Tremblay, N., Gonçalves, P. & Vandergheynst, P., Fourier could be a data scientist : From graph Fourier transform to signal processing on graphs, *Comptes Rendus Physique* **20,** 474-488, https://doi.org/10.1016/j.crhy.2019.08.003 (juill. 2019).

68. Lu, K.-S. & Ortega, A., *A graph laplacian matrix learning method for fast implementation of graph fourier transform* in *2017 IEEE International Conference on Image Processing (ICIP)* (IEEE, sept. 2017), https://doi.org/10.1109/icip.2017.8296567.

69. Lee, D., Gu, Y., Hoang, J. & Marchetti-Bowick, M., Joint interaction and trajectory prediction for autonomous driving using graph neural networks, *arXiv preprint arXiv :1912.07882* (2019).

70. Cao, D., Li, J., Ma, H. & Tomizuka, M., *Spectral Temporal Graph Neural Network for Trajectory Prediction* in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, mai 2021), https://doi.org/10.1109/icra48506.2021.9561461.

71. Cohen, J., A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* **20,** 37-46, https://doi.org/10.1177/001316446002000104 (avr. 1960).

72. Breiman, L., *Machine Learning* **45,** 5-32, https://doi.org/10.1023/a:1010933404324 (2001).

73. Dubois, H., Le Callet, P., Hornberger, M., Spiers, H. J. & Coutrot, A., *Capturing and Explaining Trajectory Singularities using Composite Signal Neural Networks* in *2020 28th European Signal Processing Conference (EUSIPCO)* ISSN : 2076-1465 (jan. 2021), 1422-1426.

74. Mahmoudi, R. & Akil, M., Analyses of the Watershed Transform (2011).

75. Zhang, J., Wu, F., Chang, W. & Kong, D., Techniques and Algorithms for Hepatic Vessel Skeletonization in Medical Images : A Survey, *Entropy* **24,** 465, `https://doi.org/10.3390/e24040465` (mars 2022).

76. Higuchi, R. & Fujimoto, Y., *Path Extraction for Autonomous Mobile Robot Using Skeletonization* in *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society* (IEEE, oct. 2021), `https://doi.org/10.1109/iecon48115.2021.9589792`.

77. Pan, J., Zhang, J., Luo, S., Zhang, J. & Liang, Y., Automatic annotation of liver computed tomography images based on a vessel-skeletonization method, *International Journal of Imaging Systems and Technology* **30,** 704-715, `https://doi.org/10.1002/ima.22411` (fév. 2020).

78. Gomes, R., Krause, A. & Perona, P., *Discriminative Clustering by Regularized Information Maximization* in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1* (Curran Associates Inc., Vancouver, British Columbia, Canada, 2010), 775-783.

79. Hjelm, R. D. *et al., Learning deep representations by mutual information estimation and maximization* in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (OpenReview.net, 2019), `https://openreview.net/forum?id=Bklr3j0cKX`.

COLLEGE DOCTORAL PAYS DE LA LOIRE
ED MATHEMATIQUES STIC

Nantes Université

Titre : Apprentissage automatique pour l'analyse de trajectoires spatiales : extraction conjointe de caractéristiques démographiques et comportementales

Mot clés : Apprentissage automatique, séries temporelles multivariées, traitement du signal

Résumé : La façon dont les humains se déplacent dans un environnement donné est liée à certaines de leurs caractéristiques démographiques et cliniques, comme leur âge ou leur statut cognitif. Dans cette thèse, nous avons cherché à quantifier l'interaction entre le profil des navigateurs et leur comportement spatial via trois approches complémentaires. Nous avons notamment utilisé les données issues d'un jeu vidéo de navigation spatiale - Sea Hero Quest - donnant accès aux trajectoires de millions de joueurs aux profils démographiques variés. La première approche propose une architecture de modèle à réseaux de neurones parallèles, afin de prendre en compte la nature spatio-temporelle des trajectoires. La seconde associe à chaque trajectoire une entropie calculée à partir de la distribution des trajectoires, pour prendre en compte le contexte et identifier la singularité du navigateur. La troisième permet de produire un groupement joint sur d'un coté les données comportementales et de l'autre démographiques. Les expérimentations que nous avons menées nous ont permis de valider les résultats obtenus antérieurement avec des métriques et des méthodes d'analyse simples, mais également de les compléter, en explicitant par exemple la nature des effets de l'âge et du genre sur le comportement spatial. Ces travaux permettront aux neuroscientifiques de mieux comprendre les facteurs sous-tendant les différences individuelles en terme de sens de l'orientation.

Title: Machine learning for spatial trajectory processing: joint analysis of demographic and behavioral characteristics

Keywords: Machine learning, multivariate time-series, signal processing

Abstract: How humans move in a given environment is influenced by some of their demographics and clinical characteristics, such as their age or cognitive state. In this thesis, we tried to quantify the interaction between the navigator's demographics and their spatial behavior using three complementary approaches. We used data from a wayfinding video game - Sea Hero Quest - which gives access to the trajectories of millions of players with various demographic profiles. The first approach proposes a parallel neural network architecture that takes into account the spatio-temporal nature of the trajectories. The second one computes an entropy metric from the distribution of all trajectories, in order to learn context and identify the singularity of the navigator. The third approach allows us to produce a joint clustering from both behavioral and demographic data. The experiments we conducted allowed us to validate the results previously obtained with simple metrics and analysis methods, but also to complete them, by clarifying for example the nature of the effects of age and gender on spatial behavior. This work will allow neuroscientists to better understand the factors underlying individual differences in terms of sense of orientation.