Modélisations de séquences spatialisées dans les réseaux d'ordre supérieur

Lysa Corcuff, François Queyroi

Modélisations de séquences spatialisées dans les réseaux d'ordre supérieur

Lysa Corcuff*
François Queyroi*

*LS2N, UMR CNRS 6004, Université de Nantes
nom.prenom@univ-nantes.fr,
https://www.ls2n.fr/equipe/duke/

Résumé. L'analyse des mobilités requiert souvent une représentation des flux en modèle markovien d'ordre 1. De nombreux travaux envisagent l'utilisation d'ordres supérieurs afin de construire des réseaux fournissant des meilleures représentations des séquences de déplacement observées. Nous proposons ici l'analyse et la comparaison des qualités prédictives et de la taille de tels modèles sur différents jeux de données géographiques. Nous allons également nous intéresser à la prise en compte de variables exogènes telles que la position ou des catégories de lieux visités qui offre des pistes de recherche intéressantes. Nos expériences indiquent notamment que le modèle HON (Xu et al. (2016)) permet d'obtenir des modèles parcimonieux qui conservent une bonne qualité prédictive même si certains résultats n'ont pu être reproduits. En particulier, aucune stratégie analysée ici ne permet d'obtenir des meilleures prédictions que le modèle d'ordre fixe (Rosvall et al. (2014)).

1 Introduction et Contexte

L'analyse des mobilités se base sur l'étude de séquences de déplacements entre lieux effectués par des personnes ou des véhicules. Dans ce cadre, la représentation par un réseau origine-destination est une approche courante qui peut être fouillée pour obtenir de l'information sur le système sous-jacent (Ducruet et Berli (2018)). Si les données en entrée correspondent à un ensemble de suites de lieux $s_1 s_2 s_3 \ldots$ alors ce réseau correspondra généralement à un graphe dont le poids d'une arête (s_1, s_2) est le nombre de transitions observées entre les lieux s_1 et s_2. Les paires $s_1 s_2$ et $s_2 s_3$ sont ainsi considérées comme des mouvements indépendants. Cette représentation correspond aussi à un modèle prédictif. En effet, des marches aléatoires sur ce graphe peuvent être utilisées pour répondre à la question : quel sera le lieu visité après s_2 ? En utilisant pour probabilités de transition les fréquences relatives, on aboutit à un modèle de Markov d'ordre 1 ou «sans-mémoire» dans lequel seule la dernière position du marcheur importe. Une question légitime est de savoir si ce modèle représente bien les séquences observées en réalité. En effet, de nombreux algorithmes de fouille de graphes se basent sur la notion de marche aléatoire (*e.g. PageRank*). Des travaux récents (Rosvall et al. (2014)) ont remis en cause la pertinence de ces modèles sans mémoire et proposent de construire des réseaux permettant de tenir compte des états antérieurs d'un marcheur et ainsi de capturer les dépendances

indirectes entre les lieux. Ces réseaux correspondraient alors à des réseaux d'«ordre supérieur» dans lesquels les dépendances indirectes sont représentées par des «noeuds-mémoires» encodant des séquences de déplacements (Xu et al. (2016)).

Pour des séquences de déplacements données, la représentation de réseau d'ordre 1 n'est pas unique car on peut définir de plusieurs façons les probabilités de transition (Queyroi (2019)). C'est également le cas pour les modèles d'ordre supérieur. Un premier enjeu de ce travail est ainsi de comparer différentes définitions de ces probabilités de transitions. En plus des modèles existants de la littérature basés sur les fréquences relatives, on va utiliser des mécanismes issus de la compression de textes et aussi définir un modèle basé sur les positions géographiques. Un autre enjeu est ici la taille des modèles et *a fortiori* des réseaux créés (du fait de l'introduction des «noeuds-mémoires»). En effet, la prise en compte des lieux précédents mène à une explosion combinatoire du nombre de cas à considérer. On peut dans ce contexte se demander si l'utilisation de variables exogènes ou d'hypothèses géographiques ne peut pas permettre d'obtenir des modèles «plus simples» qui capturent pourtant la complexité des trajectoires observées. Ces deux aspects de la modélisation auront un impact sur la qualité de prédiction. Nous allons comparer les différents modèles sur trois jeux de données géographiques précédemment utilisés dans la littérature.

La prédiction de trajectoires spatiales est un domaine de recherche très développé. Beaucoup d'études portent par exemple sur les traces GPS des véhicules (Froehlich et Krumm (2008)). L'objectif pour les réseaux d'ordre supérieur n'est pas de fournir le meilleur modèle prédictif des mobilités mais plutôt une représentation agrégée fidèle et analysable de celles-ci. On a ainsi pour contrainte de représenter les trajectoires par une suite d'événements discrets sans dimension temporelle. D'autres plateformes permettent la fouille de trajectoires incluant cette dimension (Giannotti et al. (2007)). Dans le domaine de l'analyse de réseaux, Xu et al. (2016) ont certes discuté le lien entre leur représentation (modèle HON détaillé ci-dessous) et un modèle notamment utilisé en bioinformatique proche de la compression textuelle. Toutefois seul le modèle d'ordre fixe (modèle FO détaillé plus bas) est empiriquement utilisé pour les comparaisons. Les auteurs indiquent que leur modèle permet d'éviter un sur-apprentissage inhérent au modèle à ordre fixe ; cependant les tests n'ont été faits que pour un unique jeu de données réel à notre connaissance.

2 Modèles de prédiction utilisés

Pour étudier les différents modèles liés aux réseaux d'ordres supérieurs, il n'est pas nécessaire de construire ces réseaux mais seulement de définir les probabilités de transitions. Ce sont des estimations de la probabilité $P(\sigma|s)$ où $\sigma \in A$ est un symbole (ou *lieu*) et s est appelé le *contexte* (séquences d'éléments de A). L'ensemble A peut être considéré comme un *alphabet* de tous les symboles (lieux) possibles. Dans le cadre des réseaux origine-destination, A sera l'ensemble des sommets du graphe. Les modèles décrits ici se basent sur la quantité $N(s\sigma)$ qui représente le nombre de fois où la séquence s suivie de σ a été observée durant l'apprentissage du modèle. Cette quantité peut être facilement évaluée en utilisant un arbre des suffixes (Begleiter et al. (2004)).

Ordre fixe (FO). Le modèle à ordre fixe (Rosvall et al. (2014)) est le plus direct et une base pour les autres étudiés ici. L'idée est d'énumérer l'ensemble des sous-séquences d'une longueur donnée k pour déterminer la probabilité du lieu suivant. Formellement, étant donné un contexte s, on utilise la fréquence relative de transition vers σ à la suite de la séquence s :

$$P_{fo}(\sigma|s) = \frac{N(s\sigma)}{\sum_{\sigma' \in A} N(s\sigma')} \tag{1}$$

Pour ce modèle et les suivants, l'estimation pour un contexte s tel que $|s| > k$ ou qui n'est pas observé durant l'apprentissage (dénominateur de l'Eq. 1 nul) se fera avec $P_{fo}(\sigma|s')$ où s' est le suffixe de longueur au plus k de s rencontré dans la base. La taille de ce modèle est de $\mathcal{O}(|A|^{k+1})$. Comme pour les réseaux origine-destination «classiques» (d'ordre 1) on observe une densité bien plus faible en pratique ; de nombreuses transitions ayant une probabilité nulle.

Modèle *Higher-Order Network* (HON) introduit dans Xu et al. (2016) et Saebi et al. (2020). Il a pour objectif de réduire la taille du modèle FO sans pour autant perdre en qualité de prédiction. L'intérêt pour les auteurs est de créer un graphe de taille raisonnable où des sommets sont dupliqués pour encoder les sous-séquences menant à des ensembles de sommets différents. Le modèle consiste à retirer du modèle FO les sous-séquences abc si la distribution $P(.|abc)$ est proche de $P(.|bc)$ pour obtenir un modèle plus parcimonieux et d'ordre variable. Pour un contexte $s_1^k = s_1 s_2 \dots s_k$, le modèle probabiliste correspondant aux constructions des auteurs peut s'écrire :

$$P_{ho}(\sigma|s_1^k) = \begin{cases} P_{ho}(\sigma|s_2^k) & \text{, si } KL(P_{fo}(.|s_1^k), P_{fo}(.|s_2^k)) \leq \frac{k}{\log_2 N(s_1^k)} \\ P_{fo}(\sigma|s_1^k) & \text{, sinon} \end{cases} \tag{2}$$

où $KL(P_a, P_b)$ est la divergence de Kullback-Leibler entre les distributions P_a et P_b. Ce filtrage des contextes peu pertinents peut avoir deux effets : a) éviter un sur-apprentissage du modèle b) rendre le modèle moins précis. Il nous faut vérifier expérimentalement l'impact de ces deux effets. C'est également vrai pour la taille du modèle ; théoriquement l'ordre de grandeur est toujours $\mathcal{O}(|A|^{k+1})$ et la réduction effective de la taille va dépendre du jeu de données utilisé pour la construction.

Prédiction par reconnaissance partielle (PPM). Le problème de prédiction des lieux peut être vu comme un problème de compression de symbole. Les modèles utilisent dans ce cadre des mécanismes permettant de gérer les fréquences nulles (Begleiter et al. (2004)). Cela peut permettre d'éviter un éventuel sur-apprentissage dans notre cas. On va utiliser le modèle PPM : pour un contexte $s_1^k = s_1 s_2 \dots s_k$, on pose

$$P_{ppm}(\sigma|s_1^k) = \begin{cases} \dfrac{N(s_1^k\sigma)}{|A_{s_1^k}| + \sum_{\sigma' \in A} N(s_1^k\sigma')} & \text{, si } N(s_1^k\sigma) > 0 \\ \dfrac{|A_{s_1^k}|}{|A_{s_1^k}| + \sum_{\sigma' \in A} N(s_1^k\sigma')} P_{ppm}(\sigma|s_2^k) & \text{, sinon} \end{cases} \tag{3}$$

où $A_s = \{x : N(sx) > 0\}$ *i.e.* les différents symboles observés après s. Lorsque $N(s_1^k\sigma) = 0$, une probabilité non nulle est attribuée à cette séquence. Elle dépend du nombre de symboles

différents rencontrés et de la probabilité du suffixe $s_2^k\sigma$. La séquence $s_1^k\sigma$ aura ainsi une probabilité plus importante si s_1^k est rarement apparu et que le nombre de possibilités pour le symbole suivant est important. Ainsi, moins un contexte donné est apparu, plus importante sera la probabilité donnée à n'importe quel autre symbole. Toutefois, cela va directement attribuer une probabilité plus faible aux séquences effectivement présentes dans la base d'apprentissage.

Modèle géographique (Geo). Ce modèle est, à l'instar de PPM, une redéfinition des probabilités de transition qui va intégrer la localisation des lieux. Une hypothèse est que les lieux ont plus de chances d'être visités si les lieux d'arrivée correspondant aux mêmes contextes sont proches. Cette stratégie peut sembler raisonnable lorsque les lieux sont issus de discrétisation de traces GPS (voir exemple des taxis ci-dessous). La probabilité dans ce modèle va correspondre à un mélange gaussien dont les densités sont centrées sur les points $\{x \in A\}$ avec des poids $\{N(sx)\}_{x\in A}$ et une dispersion notée $\gamma > 0$ (paramètre fixe). Les probabilités sont normalisées pour seulement tenir compte des positions possibles *i.e.* qui correspondent à un lieu de A. On a ainsi, pour un contexte s de longueur k

$$P_{geo}(\sigma|s) = \frac{\sum_{x\in A} N(sx)K_\gamma(\sigma, x)}{\sum_{x\in A} N(sx)\sum_{a\in A} K_\gamma(a, x)} \tag{4}$$

avec $d(a, b)$ la distance et $K_\gamma(a, b) = \exp(-\gamma^{-1}d(s, x)^2) \in [0, 1]$ le noyau RBF donnant la proximité entre a et b. Notons que si $\gamma \to \infty$ alors tous les lieux sont équiprobables $P_{geo}(\sigma|s) = \frac{1}{|A|}$. En revanche, si $\gamma \to 0$ alors $P_{geo}(\sigma|s) = P_{fo}(\sigma|s)$ si $d(a, b) > 0$ pour $a \neq b \in A$.

Modèle Aller-Retour (TB). Ce modèle n'utilise pas d'informations exogènes mais se base sur l'hypothèse suivante : les séquences de déplacements peuvent être représentées par des allers-retours entre deux lieux. Un exemple peut être les trajets en avion qui consistent souvent en des allers-retours. Xu et al. (2016) note en effet la prévalence des allers-retours dans les réseaux d'ordre 2 construits expérimentalement. L'idée est donc de combiner un modèle sans mémoire (*i.e.* P_{fo} pour $k = 1$) avec le calcul d'une probabilité, pour chaque lieu, de revenir au lieu visité précédemment. Formellement, on dispose d'un contexte s_1s_2. Si la position suivante $\sigma = s_1$ alors ce mouvement correspond soit à un retour dont la probabilité est donnée par la fréquence de retours observés en s_2 soit à un transit pour lequel la probabilité de visiter s_1 est la même que pour FO avec $k = 1$. En combinant ces deux cas, on obtient la probabilité

$$P_{tb}(\sigma|s_1s_2) = (1 - r(s_2))P_{fo}(\sigma|s_2) + \frac{\sum_{x\in A} N(xs_2x)}{\sum_{x\in A}\sum_{y\in A} N(xs_2y)}\mathbb{1}(s_1 = \sigma) \tag{5}$$

avec $\mathbb{1}(a = b) = 1$ si $a = b$ et 0 sinon. Ce modèle a une taille équivalente à P_{fo} pour $k = 1$ auquel il faut ajouter les $\mathcal{O}(|A|)$ valeurs des probabilités d'aller-retour (opérande droite de l'Eq. 5). Si ce modèle obtient des résultats comparables aux modèles avec mémoire ($k > 1$) alors les trajectoires peuvent être résumées par un mécanisme plus parcimonieux d'allers-retours.

Modèles avec catégories. L'objectif est ici de construire un modèle parcimonieux des séquences de déplacements en réduisant, à l'instar de HON, le nombre de sous-séquences considérées. Pour cela, on peut faire l'hypothèse que certains lieux peuvent être associés à des

mêmes motifs. Par exemple, on peut supposer que les navires provenant de Chine et passant par le port de Singapour auront une distribution proche en terme de prochain port visité. Si c'est le cas, on peut définir une seule distribution au lieu d'en faire une pour chaque port chinois. Formellement, on considère que chaque lieu s appartient à une catégorie $C(s) \in \mathcal{C}$. Une séquence $s = s_1 \dots s_k$ sur l'alphabet A correspond alors à la séquence $C(s_1) \dots C(s_k)$ sur l'alphabet $\mathcal{C}$. En combinant les deux alphabets, on définit $N(C(s)\sigma) : \mathcal{C}^k \times A \to \mathbb{N}$ comme le nombre de fois où les catégories $C(s) = C(s_1) \dots C(s_k)$ suivies du symbole σ sont apparues durant l'apprentissage. Le modèle peut alors s'écrire

$$P_{\mathcal{C},0}(\sigma|s) = \frac{N(C(s)\sigma)}{\sum_{\sigma' \in A} N(C(s)\sigma')} \tag{6}$$

pour une séquence $C(s)$ de longueur maximum k présente dans la base. Une variation de ce modèle correspondant à l'exemple donné plus haut va également être étudiée. Il faut pour cela représenter le fait que les lieux appartenant à la même catégorie peuvent correspondre à des comportements différents. On va ainsi utiliser les catégories des seuls $k-1$ premiers lieux. On obtient le modèle $P_{\mathcal{C},1}(\sigma|s_1^k)$ qui se traduit par

$$P_{\mathcal{C},1}(\sigma|s_1^k) = \frac{N(C(s_1^{k-1})s_k\sigma)}{\sum_{\sigma' \in A} N(C(s_1^{k-1})s_k\sigma')} \tag{7}$$

En supposant que le nombre de catégories est très inférieur au nombre de lieux, la taille de ces modèles sera significativement inférieure au modèle à ordre fixe avec des tailles de $\mathcal{O}(|A||\mathcal{C}|^k)$ et $\mathcal{O}(|A|^2|\mathcal{C}|^{k-1})$ respectivement.

3 Protocole expérimental

Le tableau 1 contient un récapitulatif des jeux de données utilisés. Nous n'utilisons qu'une seule variable catégorielle pour tester les modèles $P_{\mathcal{C},0}$ et $P_{\mathcal{C},1}$. Une implémentation des modèles et des tests réalisés est disponible à l'adresse `https://github.com/fqueyroi/GeoSeqPredModels`.

— **Maritime.** Ces séquences sont extraites de la base *Lloyd's List Intelligence*. Elles correspondent aux déplacements des navires de transports de marchandises entre le 1er Avril et le 31 Juillet 2009 (Ducruet et Berli (2018)). Une séquence correspond ainsi à un unique navire et chaque élément aux ports où ce navire a fait escale sur la période. Une autre partie de cette base a été utilisée dans Xu et al. (2016) pour démontrer la pertinence du modèle HON. Les catégories utilisées sont les pays des ports visités par les navires.

— **Taxis.** Ce jeu de données fait partie du challenge ECML/PKDD 2015.Les séquences correspondent aux relevés à intervalle régulier des positions GPS de 442 taxis de la ville de Porto sur les trois premières journées de Juillet 2013. Pour discrétiser la base, les coordonnées Latitude/Longitude sont associées au point d'intérêt (POI) le plus proche (commerces, arrêt de bus, *etc.*). Ces points d'intérêts sont obtenus en utilisant une extraction de base *Open Street Map*. Les catégories utilisées sont les 41 postes de police de la ville de Porto. Ces points découpent la ville en zones reflétant la densité de population. À l'inverse, Saebi et al. (2020)) ont utilisé les postes de police comme alphabet

de base. Nous voulons en revanche comparer les modèles sur des alphabets de tailles conséquentes.

— **Aéroports.** Ces séquences sont tirées de la base Origine-Destination (*RITA TransStat* 2014) qui regroupe les itinéraires de vols de passagers aux États-Unis durant le 1er trimestre de 2011. Chaque séquence correspond à la suite d'aéroports dans lesquels les voyageurs ont fait escale, commencé ou terminé leur voyage. Ces données furent notamment utilisées dans Rosvall et al. (2014). Les catégories correspondent aux états américains dans lesquels les aéroports sont situés.

Nom	Lieu / Catégories	Nb Seq	$\|A\|$	$\|\mathcal{C}\|$	Taille Moy.
Maritime	Port / Pays	4 298	910	227	28.83
Taxis	POI / Quartiers	8 816	2 453	41	49.11
Aéroports	Aéroport / États	2 751 486	446	52	3.98

TAB. 1 – *Description des jeux de données utilisés. Nous ne conservons que les séquences de longueur supérieure à trois après avoir retiré les répétitions. «POI» : point d'intérêt.*

Une mesure classique dans la prédiction de séquence est la *log-loss* moyenne d'un modèle $\hat{P}$. Elle correspond au logarithme de la vraisemblance d'une séquence de test s_i^k. Cette mesure ne pourra donc pas être directement utilisée car nous utilisons des modèles qui peuvent donner une probabilité nulle à certaines séquences. Le score que nous utilisons correspond à la probabilité moyenne de détecter le bon symbole *i.e.*

$$score(\hat{P}, s_1 \dots s_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(s_i | s_1 \dots s_{i-1}) \tag{8}$$

Pour chacun des jeux de données, nous utilisons 10% des séquences comme jeu de tests et nous calculons la moyenne de l'Eq. 8 sur ces séquences comme score du modèle. On exprimera ce dernier en pourcentage moyen dans la section suivante. On s'intéresse également ici à la parcimonie des modèles obtenus. Nous allons simplement définir la taille comme le nombre total d'entrées dans les structures de données nécessaires aux estimations pour chaque modèle. Ainsi, pour le modèle fixe d'ordre 1 ($P_{fo}(1)$), cette taille correspond aux nombres d'arêtes dans le graphe pondéré des flux entre lieux. Notons que cette approche est surtout pertinente pour évaluer les modèles FO, HON et par catégories. Dans le cas de Geo et PPM, le nombre de transitions non-nulles seront très probablement supérieures aux valeurs rapportées par la suite.

4 Résultats et Prespectives

Les résultats de nos expérimentations sont disponibles dans le tableau 2. Notons que le modèle $P_{fo}(1)$ correspond à la précision obtenue en utilisant un réseau origine-destination construit à partir des sous-séquences de longueur 2. C'est donc une base importante de comparaison pour les autres modèles. Ainsi, la différence de précision entre $P_{fo}(1)$ et $P_{fo}(3)$ est révélatrice de l'intérêt de modéliser ces systèmes comme des réseaux d'ordre supérieur.

Modèles	Maritime		Taxis		Aéroports	
	score	taille	score	taille	score	taille
$P_{fo}(1)$	13.08	9K	27.69	16K	4.65	12K
$P_{fo}(2)$	31.08	29K	41.31	46K	11.08	242K
$P_{fo}(3)$	44.80	58K	43.60	95K	12.90	962K
$P_{ppm}(2)$	26.23	29K	37.40	46K	10.77	242K
$P_{ppm}(3)$	36.41	58K	37.75	95K	11.73	962K
$P_{hon}(2)$	30.50	27K	40.11	34K	11.07	240K
$P_{hon}(3)$	40.70	36K	40.81	38K	11.71	697K
P_{tb}	14.05	9K	27.09	17K	7.66	13K
$P_{\mathcal{C},0}(1)$	7.53	5K	1.69	6K	3.06	6K
$P_{\mathcal{C},0}(2)$	15.87	17K	3.41	12K	7.08	89K
$P_{\mathcal{C},0}(3)$	26.01	37K	4.87	21K	8.46	419K
$P_{\mathcal{C},1}(2)$	24.62	24K	30.47	31K	8.69	131K
$P_{\mathcal{C},1}(3)$	35.72	48K	33.02	48K	10.04	558K
$P_{geo}(1, 10^{-6})$	10.64	10K	19.74	18K	3.77	13K
$P_{geo}(2, 10^{-6})$	25.42	30K	30.84	48K	8.69	243K
$P_{geo}(3, 10^{-6})$	36.83	59K	32.71	97K	10.19	963K
$P_{geo}(3, 10^{-8})$	44.46		43.25		12.89	
$P_{geo}(3, 10^{-10})$	44.79		43.60		12.90	

TAB. 2 – *Résultats obtenus sur les jeux de données pour les différents modèles. La ligne $P_X(k)$ correspond au modèle X avec un contexte de longueur k (si nécessaire). Le second paramètre pour P_{geo} correspond à la valeur de γ utilisée. Le score est la moyenne (en %) de la formule 8 pour les séquences de tests. La taille correspond au nombre d'entrées total (arrondi au milliers) dans les différentes structures de données nécessaires.*

Le modèle Aller-Retour obtient des résultats globalement très faibles comparé aux autres modèles ayant un ordre maximal de 2 (légèrement meilleurs pour le réseau aérien). Cela suggère qu'un tel mécanisme n'est probablement pas suffisant pour représenter les relations indirectes entre lieux.

Les modèles PPM et Geo engendrent une importante baisse de la précision si on le compare à P_{fo}. Les tests pour Geo ont été effectués pour des valeurs de $\gamma = \{10^{-i}, i = 6 \dots 10\}$ et la précision augmente systématiquement avec γ *i.e.* lorsque le modèle tend vers P_{fo}. C'est également vrai dans le cas des données taxis pour lesquelles on pouvait s'attendre à une amélioration vu l'imprécision de la discrétisation spatiale. L'utilisation de fréquences relatives pour définir les probabilités de transitions entre lieux semble ainsi plus pertinente que les stratégies PPM et Geo. Toutefois notre prise en compte des coordonnées n'est pas l'unique façon de procéder et d'autres solutions pourraient être envisagées.

Le modèle HON produit généralement des plus mauvaises prédictions qu'avec un ordre fixe. Ceci est en contradiction avec les résultats de Xu et al. (2016) : le filtrage de ce modèle ne semble pas éviter un sur-apprentissage. Néanmoins, la perte de précision est très inférieure au gain en terme de taille. C'est particulièrement le cas pour les taxis, où la précision n'est inférieure que de trois points pour une taille deux fois moins importante. Ce constat ne se

retrouve pas avec les modèles utilisant des catégories. Bien que ces derniers soient de taille relativement faible, les pertes en terme de précision sont importantes. HON produit parfois des modèles de plus petite taille et avec une meilleure précision. Un problème intéressant serait de trouver une classification en un nombre donné de groupes aboutissant au meilleur score pour le modèle $P_{\mathcal{C},1}(k)$ pour $k > 1$. Dans le cas de $P_{\mathcal{C},0}(1)$, cette classification correspondrait à une partition des sommets du graphe origine-destination en groupe de sommets ayant des ensembles proches de voisins.

Références

Begleiter, R., R. El-Yaniv, et G. Yona (2004). On prediction using variable order markov models. *Journal of Artificial Intelligence Research 22*, 385–421.

Ducruet, C. et J. Berli (2018). Mapping the globe. the patterns of mega-ships. *Port Technology International 77*, 94–96.

Froehlich, J. et J. Krumm (2008). Route prediction from trip observations. In *SAE Technical Paper*. SAE International.

Giannotti, F., M. Nanni, F. Pinelli, et D. Pedreschi (2007). Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, New York, NY, USA, pp. 330–339. ACM.

Queyroi, F. (2019). Comparing Static and Dynamic Graphs built from Mobility Traces. In *MARAMI 2019*, Actes Marami 2019, Dijon, France.

Rosvall, M., A. V. Esquivel, A. Lancichinetti, J. D. West, et R. Lambiotte (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications 5*(1), 1–13.

Saebi, M., J. Xu, L. M. Kaplan, B. Ribeiro, et N. V. Chawla (2020). Efficient modeling of higher-order dependencies in networks : from algorithm to application for anomaly detection. *EPJ Data Science 9*(1), 15.

Xu, J., T. L. Wickramarathne, et N. V. Chawla (2016). Representing higher-order dependencies in networks. *Science advances 2*(5), e1600028.

Summary

Transport Network analysis often requires to model transitions as order 1 markovian models. Previous works suggest the use of higher order models in order to build networks that can more accurately predict observed sequences. In this work, we compare these models' prediction capabilities and size using different real world trajectories datasets. Beside generic models, we introduce models that include exogenous variables such as the location or the categories of the visited places. They provide further research opportunities. Our experimental results suggest that the HON model (Xu et al. (2016)) offers a good compromise between predictive capabilities and parsimony. However, some claimed properties of this model could not be reproduced. Indeed, none of the strategies used here results in better predictions than the fix-order model (Rosvall et al. (2014)).