Université Gustave Eiffel

univ-gustave-eiffel.fr

Thèse
Fusion de données multi-sources pour l'analyse de la mobilité

Date de démarrage : Septembre 2020
Date limite de candidature : 15 Avril 2020
Localisation principale : Champs-Sur-Marne (77)
Durée du contrat : Contrat doctoral de 36 mois
Direction de thèse : Latifa Oukhellou, Co-direction : Nour-Eddin El Faouzi
Encadrements : Angelo Furno, Etienne Côme
Inscription : Université Gustave Eiffel (ED MSTIC)
Contexte : Thèse dans le cadre du projet **ANR Mobitic**, en collaboration avec le laboratoire SENSE d'Orange Labs. En plus de l'allocation doctorale, le doctorant effectuera une mission expertise chez Orange durant sa thèse et percevra un complément de salaire.

Mots clés : Mobilités, données multi-sources, fusion de données, apprentissage automatique, fouille de données

Contexte

Les approches actuelles pour mesurer les mobilités des personnes reposent sur des données classiques (enquêtes, comptages de trafic, etc.). Les données numériques, ouvrent des perspectives d'analyse dynamique à des niveaux de précision géographique et temporelle très fins, et offrent aux acteurs le potentiel d'un suivi rapproché des évolutions de la mobilité des personnes et des politiques de mobilité durable visant à les infléchir. Cependant, les données numériques prises isolément sont partielles et biaisées et leur capacité à saisir des phénomènes urbains complexes et inter-reliés est encore réduite. Leur combinaison à d'autres données classiques et numériques (téléphonie, billettiques, données de trafic) permettra de tirer parti des atouts de chaque type de données : pertinence, représentativité et fiabilité à des niveaux spatio-temporels très fins. Dans cette optique, le projet ANR MobiTIC (2020-2023), coordonné par le GRETTIA et dans lequel le LICIT est partenaire, ambitionne de combiner des données multi-sources, pour mesurer les mobilités et présences des personnes en milieu urbain et péri-urbain. Ce projet implique Orange, Insee et Géographie-cités comme partenaires. Kéolis Rennes est également associé au projet en tant que fournisseur de données billettiques. Les travaux envisagés s'inscrivent dans la continuité des travaux de recherche menés au GRETTIA et au LICIT sur l'exploitation des données numériques pour l'analyse de la mobilité et des fonctions urbaines. La thèse envisagée met en synergie les compétences des deux laboratoires pour proposer un corpus méthodologique dans le domaine de l'apprentissage visant à fusionner plusieurs sources de données.

Objectifs

Cette thèse vise à exploiter les données numériques multi-sources pour l'analyse de la mobilité en tirant le meilleur parti des différentes sources de données : les données de téléphonie mobile, les données d'enquêtes, les données billettiques et les données de trafic routier. Les données de téléphonie ne peuvent à elles seules fournir les informations nécessaires à la construction des

indicateurs recherchés. En plus des redressements mobilisant les informations socio-démographiques, les indicateurs sur la mobilité peuvent bénéficier des informations issues des sources tierces.

Deux volets seront investigués dans la thèse : (i) le premier est dédié à l'estimation de matrices origines-destinations (OD) par mode et par motif de déplacement, (ii) le second, à but exploratoire, consiste à extraire des profils types de la demande de transport en exploitant des données multi-sources.

Le premier volet de la thèse s'attachera à développer des méthodes d'estimation de matrices OD par mode et par motif avec l'objectif de fournir des estimations sur des périodes temporelles courtes et à une échelle spatiale fine. Deux stratégies seront explorées :

- travailler à l'échelle de la trace de signalisation individuelle tel que proposé dans [1] et intégrer au sein de la procédure de map-matching multi-modal des informations agrégées sur les différents modes en provenance des autres sources. Une telle approche pourra trouver des synergies avec les travaux effectués dans le cadre du projet ANR JCJC Promenade (porté par A. Furno – LICIT et associé à cette thèse), notamment les travaux de thèse (en cours - fin prévue pour Octobre 2021) menés par L. BONNETAIN, financée par le Ministère de la Transition écologique et solidaire (ENTPE). En particulier, ce volet pourra bénéficier des méthodes en cours de développement pour le nettoyage de traces de téléphonie mobile, la détection des phases statiques et mobiles des utilisateurs et des premières solutions multi-sources (exploitant à la fois les données de téléphonie et les données billettiques) pour l'inférence des temps de parcours typiques et des parcours populaires dans un contexte multimodal. L'objectif sera d'étendre ces méthodes pour la caractérisation des motifs de déplacements et la quantification des parts modales en mobilisant, par exemple, des techniques de **clustering spatio-temporelles** et des approches d'imputation statistique pour compléter et enrichir les traces de mobilité individuelles. A cet égard, la fréquence d'échantillonnage plus élevée garantie par les données de signalisation Orange, et une couverture plus complète des usages de différents modes de transport, permise par l'hétérogénéité des données disponibles, apparaissent à la fois fondamentaux et prometteuses pour la qualité des méthodes qui seront développées.

- combiner les sources de données une fois les opérations d'agrégation (spatiale et temporelle) effectuées [2]. Bien que potentiellement moins riche, cette approche simplifie les traitements et ne soulève pas de difficultés liées au traitement de données individuelles. De plus, les sources de données sont complémentaires : les données de téléphonie fournissant des informations sur la structure de la matrice OD pour l'ensemble des modes alors que les données billettiques et de boucles de par leur nature mono-modale fournissent des informations de volumes sur certains nœuds ou arcs du réseau de transport multimodal. De telles approches ont déjà été mises en œuvre avec succès [3,4] en utilisant un simulateur (modèle microscopique de trafic) pour l'affectation de la matrice. Cependant ces travaux ne traitaient pas dans le détail la question des modes et des motifs contrairement à ce qui est envisagé dans la thèse.

Pour le deuxième volet exploratoire, des travaux précurseurs sur l'application des modèles à variables latentes tels que le modèle LDA (Latent Dirichlet Allocation), les modèles stochastiques par blocs [5,6], la factorisation de tenseur [7,8] ou l'analyse factorielle exploratoire [9] ont été menés dont certains directement par les chercheurs impliqués dans l'encadrement de la thèse. Ces modèles permettent d'extraire un nombre réduit de profils types de la demande de transport. L'analyse spatiale des profils ainsi obtenus renseigne à la fois sur la mobilité des personnes et les dynamiques urbaines sous-jacentes. Ces modèles seront étendus pour exploiter conjointement les données multi-sources

afin de construire des analyses spatio-temporelles conjointes. Ce travail s'inspirera des avancées réalisées dans le domaine des **modèles génératifs pour données multi-vue** [10,11].

Déroulement
Le planning prévu pour cette thèse est le suivant :

1) Etude bibliographique sur les modèles à variables latentes ; les modèles génératifs pour données multi-vue, les algorithmes de détection d'anomalies ;

2) Prétraitement des données collectées pour l'estimation de matrices origines-destinations (OD) dynamiques par mode et par motif : préparation des données en vue des traitements statistiques ultérieurs (imputation des modes, imputation des motifs, redressement). Cette étape est en étroite liaison avec le projet ANR Promenade piloté au Licit ;

3) Analyse exploratoire pour l'extraction de profils types de la demande en transport ; mise en œuvre des modèles génératifs pour données multi-vue ;

Profil du candidat
Le candidat recherché pour cette thèse suit un diplôme équivalent Master 2 orienté Science de données ou Statistique, avec un intérêt pour le domaine de la mobilité et les transports. Une bonne connaissance des langages R et/ou Python est également demandée.

Contacts :

Latifa Oukhellou Directrice de Recherche Latifa.oukhellou@univ-eiffel.fr 14-20 Bd Newton, 77 420 Champs-sur-marne Tél : +33 (0)1 81 66 87 19	Etienne Côme Chargé de Recherche Etienne.come@univ-eiffel.fr 14-20 Bd Newton, 77 420 Champs-sur-marne Tél : +33 (0)1 81 66 87 18	Angelo Furno Chargé de Recherche angelo.furno@univ-eiffel.fr 25 Avenue François Mitterrand, 69500 Bron Tel. +33 (0)4 78 65 68 70

Références bibliographiques :

[1] F. Asgari, A. Sultan, H. Xiong, V. Gauthier and M. A. El-Yacoubi, "CT-Mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network," Computer Communications, vol. 95, pp. 69-81, 2016.

[2] Friedrich, M. & Immisch, K. & Jehlicka, P. & Otterstätter, T. & Schlaich, J. (2010). Generating Origin-Destination Matrices from Mobile Phone Trajectories. Transportation Research Record: Journal of the Transportation Research Board. 2196. 93-101. 10.3141/2196-10.

[3] Zilske, M.l & Nagel, K. (2015). A Simulation-based Approach for Constructing All-day Travel Chains from Mobile Phone Data. Procedia Computer Science. 52. 468-475. 10.1016/j.procs.2015.05.017.

[4] I, Md Shahadat & C., Charisma & W., Pu & Gonzalez, Marta C.. (2014). Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies. 40. 63–74. 10.1016/j.trc.2014.01.002.

[5] P-A. Laharotte, R. Billot, E. Côme, L. Oukhellou, A. Nantes, N-E El Faouzi (2015) Spatiotemporal Analysis of Bluetooth Data: Application to a Large Urban Network. IEEE Transactions on Intelligent Transportation Systems 16(3): 1439-1448.

[6] A. E. Papacharalampous et al., Multi-Modal Data Fusion for Big Events. IEEE Intelligent Transportation Systems Magazine, vol. 7, no. 4, pp. 5-10, 2015.

[7] Côme, E., Randriamanamihaga, A., Oukhellou L. and Aknin, P. Spatio-temporal analysis of Dynamic Origin-Destination data using Latent Dirichlet Allocation. Application to the Vélib' Bike Sharing of Paris. In Proc of 93rd Annual Meeting of the Transportation Research Board, 2014.

[8] Sun, L., Axhausen, K.W. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transportation Research Part B: Methodological 91, 511–524.

[9] Furno A, Fiore M, Stanica R, Ziemlicki C, Smoreda Z (2017). A Tale of Ten Cities : Characterizing Signatures of Mobile Traffic in urban Areas, IEEE TMC 16(10).

[10] Airoldi, E., Wang, X., & Lin, X. Multi-way blockmodels for analysing coordinated high-dimensional responses. The Annals of Applied Statistics,7(4), 2431-2457. 2013.

[11] Shimamawari, T., Eguchi, K., Takasu, A. Bayesian Non-parametric Inference of Multimodal Topic Hierarchies, Journal of Information Processing, 2016, Volume 24, Issue 2, Pages 407-415.