Université Gustave Eiffel

univ-gustave-eiffel.fr

PhD
Multi-source data fusion for mobility analysis

Starting date: September 2020
Closing date for application: April 15th, 2020
Main localisation: Champs-Sur-Marne (77, France)
Contract: PhD Contract (36 months)
Main advisor : Latifa Oukhellou, Co-direction : Nour-Eddin El Faouzi
Encadrements : Angelo Furno, Etienne Côme
Inscription : Université Gustave Eiffel (ED MSTIC)
Context : The PhD is part of the **ANR Mobitic project**, done in collaboration with the SENSE laboratory of Orange Labs. A remunerated complementary mission with Orange is possible.

Keywords : Mobility, data fusion, machine learning, data-mining

Context

Current approaches for measuring people's mobility are traditionally based on data such as surveys, traffic counts, etc. Digital data open up opportunities for the dynamic analysis of mobility at very high levels of geographical and temporal accuracy and may provide stakeholders with the potential to better monitor changes in people's mobility, which is a requirement to drive sustainable mobility policies. However, single-source data are partial and biased, and their ability to grasp complex urban phenomena is therefore reduced. Their combination with other conventional and digital data (mobile phone data, smart card data, traffic data) will make it possible to take advantage of the benefits of each type of data: relevance, representativeness, and reliability at very fine spatial and temporal levels. The ANR MobiTIC project (2020-2023), coordinated by GRETTIA and in which LICIT is a partner, aims to combine multi-source data to measure the mobility and presence of people in urban and peri-urban areas. This project involves Orange, Insee, and Géographie-Cités as partners. Kéolis Rennes is also involved in the project as a supplier of smart card data. The research work to be carried out is in line with the research work carried out at GRETTIA and LICIT on the use of digital data for the analysis of mobility and urban functions. The thesis envisaged brings together the skills of the two laboratories to propose a methodological corpus in the field of data mining and machine learning aimed at merging several data sources for the mobility analysis.

Objectives

This thesis aims to exploit multi-source data for mobility analysis by taking advantage of each data source: mobile phone data, survey data, ticketing data, and road traffic data. Mobile phone data cannot provide the information required to build mobility indicators. In addition to adjustments using socio-demographic information, mobility indicators can benefit from information extracted from other data sources to overcome their inner individual limitations and biases.

Two aspects will be investigated in the thesis: (i) the first is dedicated to the estimation of origin-destination (OD) matrices for each mode and travel purpose, (ii) the second, for exploratory purposes, consists in extracting a reduced set of transport demand profiles by exploiting multi-source data.
The first part of the thesis will focus on developing methods for estimating OD matrices by mode and travel purpose to provide estimates over short periods and at a fine spatial scale. Two strategies will be explored, namely:
- working at the individual signaling trace scale as proposed in [1] and integrate into the multi-modal map-matching procedure aggregate information on the different modes provided by other sources. Such an approach could find synergies with the work carried out within the framework of the ANR JCJC Promenade project (led by A. Furno - LICIT and associated with this thesis), in particular, the ongoing thesis work (the Ph.D. defense is planned for October 2021) carried out by L. BONNETAIN, funded by the Ministry of Ecological and Solidarity Transition (ENTPE). In particular, this component will benefit from methods under development for cleaning up mobile phone traces, detecting static and mobile phases of users and the first solutions (using both mobile phone and ticketing data) for inference of typical and popular travel times in a multimodal context. The objective will be to extend these methods for the characterization of travel patterns and the quantification of modal shares by investigating, for example, Spatio-temporal clustering techniques and statistical imputation approaches to complement and enrich individual mobility traces. In this respect, the higher sampling frequency guaranteed by Orange signaling data, and more complete coverage of the uses of different modes of transport, made possible by the heterogeneity of the available data, appear both fundamental and promising for the quality of the methods to be developed.
- combining data sources once aggregation operations (spatial and temporal) have been completed [2]. Although potentially less rich, this approach simplifies processing and does not raise difficulties related to the processing of individual data. Also, the data sources are complementary: mobile phone data provide information on the structure of the OD matrix for all modes, while ticketing and loop data by their mono-modal nature provide volume information on certain nodes or arcs of the multimodal transport network. Such approaches have already been successfully implemented [3,4] using a simulator (microscopic traffic model) for matrix assignment. However, these works did not deal in detail with the question of modes and travel purposes, contrary to what is envisaged in the thesis
For the second part of the thesis, pioneering work on the application of latent variable models such as LDA (Latent Dirichlet Allocation) model, stochastic block models [5,6], tensor factorization [7,8] or exploratory factor analysis [9] was carried out, some directly by the researchers involved in the supervision of the thesis. These models allow extracting a reduced number of standard profiles of the transport demand. The spatial analysis of the profiles thus obtained provides information on both the mobility of people and the underlying urban dynamics. These models will be extended to jointly exploit multi-source data to build joint spatial and temporal analyses. This work will be based on advances in generative models for multiview data [10,11].

Schedule
The schedule for this thesis is as follows:
1) Bibliographic study on latent variable models and generative models for multiview data ;

2) Pre-processing of the data collected for the estimation of dynamic origin-destination (OD) matrices by mode and by travel purpose: preparation of the data for subsequent statistical processing (imputation of modes, imputation of travel purposes, Propensity score weighting). This step is closely linked to the ANR Promenade project piloted at Licit;

univ-gustave-eiffel.fr

3) Exploratory analysis for the extraction of transport demand profiles; Development of generative models for multiview data;

Candidate profile

The candidate for this thesis pursue a Master 2 in Data Science or Statistics with an interest in mobility and transport applications. Good knowledge of R and/or Python languages is also required.

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

Bibliography :

[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.