TRANSITDATA2019, 5th International Workshop and Symposium
8-9-10 July 2019. Paris, France

Mining regional passenger mobility activity using multiple data sources: A case study in Paris area

Yao Zhou[a], Xiaoyan Xie[a*] and Fabien Leurent[a]
[a]*LVMT, UMR-T 9403, Ecole des Ponts, IFSTTAR, UPEM, UPE, 77455, Champs-sur-Marne, France*

Abstract

In the era of big data, traditional passenger mobility analysis approaches were challenged. There is a need to not only update traditional approaches, but also apply new data analyses approaches, such as machine learning approaches and multiple data fusion approaches. To accomplish that, this paper aims to develop a methodology to recover passenger mobility information from smart ticketing data, combing with other mobility and geo-location datasets on the large scale Public Transport (PT) network of region Île-de-France (Paris area), France. By linking passenger trip type identification with post behavior performance assessment through explorative statistical analysis, this research provides applications of Unsupervised Learning and Supervised Learning approaches for mining outstanding relations of passenger trips to specific mobility activities, such as trip time profiles and purposes. To confront the results of UL and SL, major passenger trip types were identified. A posterior analysis was proposed. The association analysis revealed that the identify types had statistically significant differences on the usage of different PT transport modes. A tool for both big data management and multiple data sources application related to passenger regional mobility analyses for Île-de-France (Paris area) is developed. This study offers a big data driven instance for future regulation and better anticipating the PT network performance by targeting the main users of different PT modes.

Keywords: Passenger activity, large network, machine learning, multiple data sources, Paris area

Introduction

Understanding passenger mobility patterns on a Public Transport (PT) network is crucial not only for passenger behavior analyses, demand characterization and forecast, and transit evaluation and planning (Kurauchi and Schmöcker, 2016), but also for city evaluation and planning (Zhong et al., 2016, 2014). A core element of mobility patterns was passenger mobility activity (Alsger et al., 2018, 2016, 2015). However, passenger mobility activity analysis on a PT network were still difficult tasks, as there was lack of passenger's complete and big sample journey trajectory data.
In the era of big data, over the last decade, more diverse, accurate, abundant and large-scale new passive data together with related novel data processing and mining methodologies have become available owing to the deployments of Intelligent Transportation Systems (ITSs) and of Information and Communications Technology (ICT), and the increasing computational capacities. The new data, which were collected by automatic observation systems, are named modern data. A new branch of knowledge was emerged, the modern data measurement and data-driven methods for passenger mobility analysis in PT field [Aguiléra *et al.*, 2014; Leurent and Xie, 2018; Pelletier *et al.*, 2011; Yue *et al.*, 2014; Zhu *et al.*, 2017a, 2017b]. Indeed, traditional passenger mobility analysis approaches were challenged. There was a need to not only update traditional approaches, but also apply new data analyses approaches, such as machine learning approaches (Lopez et al., 2017) and multiple data fusion approaches (Alsger et al., 2018, 2016, 2015; Wu et al., 2018).

Research contributions

To address the above-mentioned problems, the objective of this paper is to develop a methodology to recover passenger mobility information from smart ticketing data - Automated Fare Collection (AFC) data, combing with other mobility and geo-location datasets on the large scale PT network of region Île-de-France (Paris area), France. By linking passenger trip type identification with post behavior performance assessment through explorative statistical analysis, this research provides applications of Unsupervised Learning (UL) and Supervised Learning (SL) approaches for mining outstanding relations of passenger trips to specific mobility activities, such as trip time profiles and purposes. Thus, this study offers a big data driven instance to recover mobility patterns, the knowledge of which can be useful to improving network regulation and

* Corresponding author. *E-mail address: xiaoyan.xie@enpc.fr*.

planning.

Methodology

The aim of this part is to design an approach for both big data management and multiple data sources application related to passenger regional mobility analyses for Île-de-France, France. We are interested in mining and analyzing passenger activity in Île-de-France. A trip is a complete journey between one pair of OD stations for an activity including short time intermediate activities in or near an intermediate station. Thus, a trip is composed of one or several trip-legs, where a trip-leg is a journey between one pair of OD stations along a single line (Figure 1).

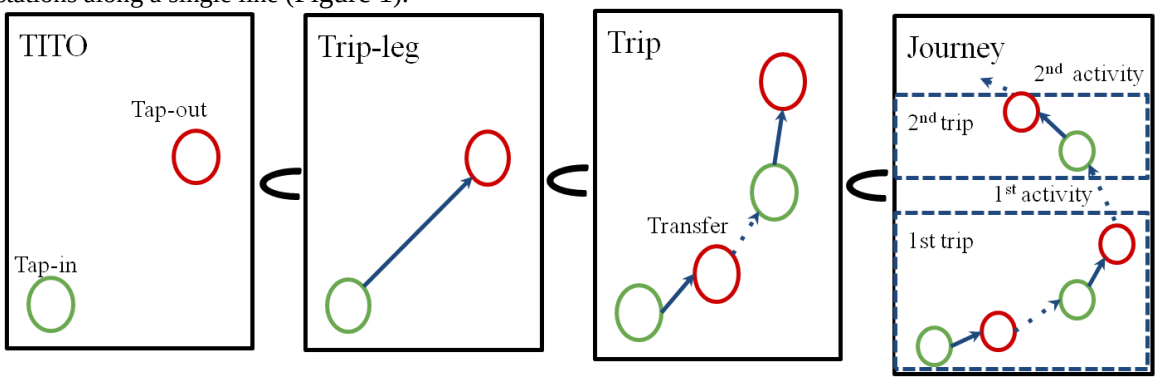

Figure 1. Components of a passenger journey: TITO, trip-leg, trip, transfer and activity.

Studied site and multiple data sources. The research revolved around analyzing passenger mobility on PT network in Île-de-France. Among the data provide by this region, in this study, we considered not only passenger related data, AFC data (Navigo data) which are about 13 million records per day (Siméon, 2017) and Transport Survey in Île-de-France (Enquête Globale Transport, EGT) (OMNIL, 2012), but also transit network and vehicle related data, GTFS (General Transit Feed Specification) data.

Trip generation. Four main steps are considered for trip generation: (1) data clearance, (2) trip-leg generation, (3) trip generation, and (4) passenger activity mining. This is accomplished by multiple data sources fusion approaches.

Passenger activity mining. At first, trip characteristic of each activity is studied. Since the generated trips do not have more information about passenger activity, except the generated time profiles of activities and trip information, for a direct analysis, we can only use Unsupervised Learning (UL) approach. The analysis is about trip time profiles, including travel time, departure time and arrival time of each trip.

Then, for deeper analyses, the generated trips are confronted with EGT data which include more passenger activity attributes, such as trip purpose attributes including work, professional affaire, school, leisure, shopping, personal affaire, and home. Firstly, we extract the data of mobility on PT network from EGT to adapt this study, and calculate the statistics of purposes in each zone. We convert secondly generated trips to EGT data format by calculating the purpose statistics of each station. Combing the trip purpose attributes with the former UL trip time profiles, a Supervised Learning (SL) is proposed for this study.

Implementation. These methods are implemented in Python with big data management and machine learning packages. A tool for both big data management and multiple data sources application related to passenger regional mobility analyses for Île-de-France is developed.

Preliminary results and discussion

As for the preliminary results, a total of 13 million records of AFC data (Navigo data) on the March 14th 2017, EGT data on 2008, GTFS data and geo-location data in OpenStreetMap were applied. To confront the results of UL and SL, major passenger trip types were identified based on their mobility features, which were defined as, (Type I) morning activity trips mixed professional affaire, work, leisure, shopping, personal affaire, and school trips; and (Type II) afternoon and evening activity trips mixed professional affaire, leisure, shopping, personal affaire, and home trips. A posterior analysis was proposed. The association analysis revealed that the identify types had statistically significant differences on the usage of different PT transport modes with a descend order metro, bus, RER and tram. Such results can be used for future regulation and better anticipating the PT network performance by targeting the main users of different PT modes.

However, due to the availability of data, future work can be done to scale the analysis by extending the scope spatially and temporally. Besides, by integrating with other data sources, more features can be recovered to describe the motility patterns, such as geo-matching with land use. More sophisticated machine learning method can also be explored to improve the process of clustering.

References

Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C., 2014. Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transp. Res. Part C Emerg. Technol. 43, 198–211.

Alsger, A., Assemi, B., Mesbah, M., Ferreira, L., 2016. Validating and improving public transport origin-destination estimation algorithm using smart card fare data. Transp. Res. Part C Emerg. Technol. 68, 490–506. https://doi.org/10.1016/j.trc.2016.05.004

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. Transp. Res. Part C Emerg. Technol. 87, 123–137. https://doi.org/10.1016/j.trc.2017.12.016

Alsger, A.A., Mesbah, M., Ferreira, L., Safi, H., 2015. Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. Transp. Res. Rec. J. Transp. Res. Board 2535, 88–96. https://doi.org/10.3141/2535-10

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27.

IdFM, 2019. Le réseau aujourd'hui. https//www.iledefrance-mobilites.fr/le-reseau/ (available 6 Feb. 2019) (in French).

Kurauchi, F., Schmöcker, J.-D., 2016. Public Transport Planning with Smart Card Data. CRC Press B. 1–261.

Leurent, F., Xie, X., 2018. On individual repositioning distance along platform during train waiting. J. Adv. Transp. 1–18. https://doi.org/https://doi.org/10.1155/2018/4264528

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., Van Lint, H., 2017. Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-14237-8

OMNIL, 2012. Enquête globale transport La mobilité en Île-de-France. EGT 2010-STIF-OMNIL-DRIEA N1, 1–20 (in French).

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568. https://doi.org/10.1016/j.trc.2010.12.003

Siméon, T., 2017. La gazette des validations. Numéro 1, 1er trimetre 2017 1–4 (in French).

Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Transp. Res. Part C Emerg. Technol. 96, 321–346. https://doi.org/10.1016/j.trc.2018.09.021

Yue, Y., Lan, T., Yeh, A.G.O., Li, Q., 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behav. Soc. 1, 69–78.

Zhong, C., Arisona, S.M., Huang, X., Batty, M., Schmitt, G., 2014. Detecting the dynamics of urban structure through spatial network analysis. Int. J. Geogr. Inf. Sci. 28, 2178–2199. https://doi.org/10.1080/13658816.2014.914521

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G., 2016. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. PLoS One 11. https://doi.org/10.1371/journal.pone.0149222

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M., 2017a. A probabilistic Passenger-to-Train Assignment Model based on automated data. Transp. Res. Part B Methodol. 104, 522–542. https://doi.org/https://doi.org/10.1016/j.trb.2017.04.012

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M., 2017b. Inferring Left behind Passengers in Congested Metro Systems from Automated Data. Transp. Res. Procedia, Transp. Res. Part C Emerg. Technol. (in Press. 23, 362–379. https://doi.org/10.1016/j.trpro.2017.05.021