Mining passenger's regional intermodal mobility from smartcard data

Xiaoyan Xie, Fabien Leurent, Yazi Zhu

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2020) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus

Mining passenger's regional intermodal mobility from smartcard data

Xiaoyan Xie[a,*], Fabien Leurent[a], and Yazi Zhu[b]

[a] *LVMT, UMR-T 9403, École des Ponts, UGE (ex IFSTTAR, ex UPEM), 77455, Champs-sur-Marne, France*
[b] *INP ENSIACET-Toulouse, 4 Allée Emile Monso, 31030, Toulouse, France*

Abstract

Park & Ride (P&R) enables railway users to access the transit network by means of their own cars. Its usage at the regional level can be analyzed on the basis of Household Travel Surveys (HTS). To overcome the HTS limitations in sample size and refreshment, this paper is aimed to combine such HTS for learning the car2rail intermodality phenomenon of individual mobility, with an Automated Fare Collection (AFC) database for inferring it over a very large set of individual trips. The approach involves three steps: (i) the HTS-based featuring of Origin-Destination (O-D) trips; (ii) the treatment of the AFC dataset using the dynamic path search and ad-hoc rules based on General Transit Feed Specification (GTFS) data, to yield AFC rail O-D trips; and (iii) the supervised machine learning of P&R usage based on the HTS and AFC data, considering three methods (Support Vector Machine, Decision Tree and Artificial Neural Network). Application to the Paris – Ile-de-France region with 2010 HTS and 2019 AFC data revealed three types of intermodal trips by an unsupervised machine learning algorithm, two of them at morning peak hours with either short or long rail distances, and the last one after the evening peak.

Peer-review under responsibility of the scientific committee of the 23rd EURO Working Group on Transportation Meeting.
Keywords: Intermodality; Activity based trip-chain model; Machine learning; AFC data; Park and Ride

1. Introduction

In big cities endowed with railway lines for mass passenger transit, Park & Ride (P&R) is a well-known travel strategy to make use of the railway network by accessing to it using one's car (Kimpton et al., 2020). P&R facilities are purported to build up such travel practices and make full use of the rail network and its environmental advantages over the car mode. In the Paris – Ile-de-France region, 46 P&R facilities were labeled by the Region using Public Transport (PT) smartcard Navigo in 2018, and it is scheduled to label 34 other car parks by 2021 (IdFM, 2020).

* Corresponding author. Tel.:+33 1 81 66 88 77; fax: +33 1 64 15 51 40.
E-mail address: xiaoyan.xie@enpc.fr or xyan.xie@gmail.com

2352-1465
Peer-review under responsibility of the scientific committee of the 23rd EURO Working Group on Transportation Meeting.

Up to now, it has been costly to monitor P&R usage. Parking surveys would be required on all sites, notably to make a distinction between plain car parking and intermodal trips. Household Travel Surveys (HTS) at the regional level are adequate tools to gain a statistical understanding of the intermodal phenomenon, e.g., the Paris – Ile-de-France region (Leurent and Polacchini, 1995); yet these tools are limited in sample size and refreshment, thereby missing spatial details. As Automated Fare Collection (AFC) data of smartcards have become widely available over the past decade, it is thus tempting to use them to depict an almost comprehensive picture of rail usage in the territory under scrutiny. Among previous research on that topic, the works by (Alsger et al., 2018, 2016; Zhou et al., 2019) have succeeded to generate Origin-Destination (O-D) trips between stations on the public transport network and to infer the trip purposes by machine learning based on an HTS, but only monomodal O-D trips on the regional RER network in Paris – Ile-de-France (Zhou et al., 2019).

This paper develops this line of analysis by studying intermodal rail usage as a phenomenon of individual travel, on the two datasets, HTS and AFC dataset. Special care is taken to rebuild individual rail O-D trips, by involving not only the AFC-monitored tap-ins and tap-outs, but also a GTFS model of railway services, dynamic path search and ad-hoc rules to delimit and feature out O-D trips in terms of utilized stations, passage times and travel time. Our methodology is devised to study the P&R travel practice in the Paris - Ile-de-France region. Special attention is given to the officially labeled P&R facilities – the 46 such facilities as of 2018 are taken here as a spatial filter to focus on a sub-population of trips. The resulting sub-populations of trips coming from the HTS and the AFC dataset, respectively, are subjected to machine learning in order to extract meaningful information from the HTS dataset and make P&R inference based on the AFC dataset.

The body of the paper is in 5 sections. After introducing the study case and the multiple data sources (Section 2), a three-step methodology is developed: (i) O-D trip map-matching from the HTS, with selection in relation to the P&R labeled sites (Section 3), (ii) O-D trip generation from the AFC dataset, with feature reconstitution involving GTFS dataset, dynamic shortest path search and ad-hoc rules (Section 4), (iii) supervised machine learning (Section 5). Its results will be taken to reveal three types of intermodal rail trips using the k-means algorithm (Section 5). The last Section brings about conclusions and suggests some directions for further research (Section 6).

2. Study case of Paris – Ile-de-France region and data sources

In France, the "Île-de-France" region encompasses the greater Paris conurbation: about 10,6 million inhabitants as of 2010, spread over about 12,500 km², with 80% of people living in 30% of the regional space (OMNIL, 2012). The region is endowed with three main systems of rail transit (IdFM, 2019): firstly, 14 metro lines (M1-14) constitute a semi-closed system equipped with tap-in gates only. Secondly, the regional railways include 8 "Transilien" suburban lines (Train H, J, K, L, N, P, R, U) and 5 RER (Regional Express Network) lines (RER A, B, C, D, E), along which most stations are equipped with both tap-in and tap-out gates, save for some suburban stations that only have tap-in machines on their platforms Thirdly, there are 11 tramway lines (T1-11) that constitute a semi-closed system equipped with on-board tap-in machines. A transfer inside any one of the 3 systems requires neither tap-out nor tap-in, with the exception of in-tapping on boarding a tram. On transferring from one to another of the three sub-modes, the tap-out is only required after alighting from a train-RER, but all tap-ins are required.

We availed ourselves of five datasets, namely: (1) the AFC data as of 2019, (2) the Household Travel Survey (HTS) data as of 2010, (3) the transit network geo-locations as of 2019; (4) the vehicle timetable data under General Transit Feed Specification (GTFS) as of 2019; (5) the list of officially labeled P&R facilities as of 2018. In the first stage of the study, we filtered the data, especially for AFC and HTS datasets.

The regional AFC system stems from fare integration at the regional level: there is a unique smartcard called Navigo. Our AFC dataset, provided by the regional mobility authority IdFM, contains the anonymous data of all Navigo smartcards on Monday 11th February 2019, about 13.6 million records. There are 12 attributes in the raw

data: card ID (anonymous number maintained over 3 months), tap instant, tap date, station ID, station name, tap-gate ID, tap type (access, egress, transfer), vehicle ID, run ID, transit mode, route ID, and operator. As those data include not only the rail transit data (the blue parts in Fig. 1a) but also the bus service data (the grey part in Fig. 1a), the AFC records of bus are filtered. The number of taps per card is depicted in the Fig. 1b. As there are some abnormal data, e.g. more than 100 taps per card per day, we keep only the cards which have 2-20 taps each per day. We got about 97.10% of the records. The distributions of tap-in records (blue bars) and tap-out records (orange bars) per rail station are shown in Fig. 1c. There are large numbers of taps in suburban stations, but more stations in the central area – Paris city.

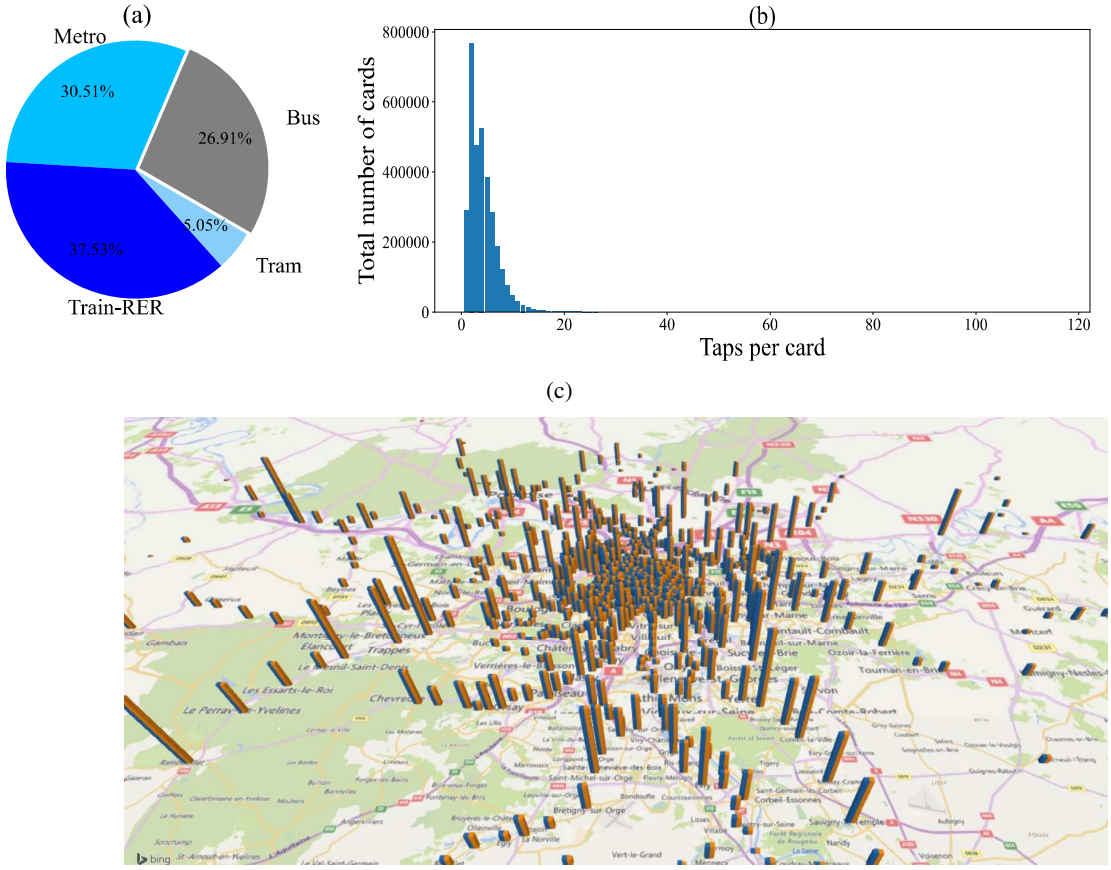


Fig. 1. AFC records: (a) the distribution among the 4 transit sub-modes; (b) the distribution of validation times per card on the surveyed day, and (c) the distribution of tap-in records (blue bars) and tap-out records (orange bars) at each rail station.

As for HTS, IdFM supplied us with the EGT2010 (Enquête Globale Transport in 2010), which is the only survey on the individual mobility of all region residents for all modes of transport and purposes. The survey was obtained by face to face interviews for 18 thousand households totaling 43 thousand people and 140 thousand trips per day as of 2009 (OMNIL, 2012). Three household samples were independently drawn from the built property file: 15 thousand on a weekday, 1,500 for a Saturday and 1,500 for a Sunday. The survey files include 4 distinct sheets:

- Households. It contains the characteristics of the household: the type of housing occupied, the equipment of the household for means of transport (cars and two-wheeled motor vehicles) and their level of income.
- Persons. Each person in the household is described in terms of age, gender, level of education, main occupation, whether or not they had a driver's license, a public transport subscription or shared bikes.

- Trips. For all persons aged 5+ who traveled the day before the interview, all trips made are recorded in terms of the times and places of origin and destination, as well as purposes (work, studies, errands, leisure, etc.).
- Means. The travel modes used in each trip are described sequentially; modal legs ends are timestamped and geo-located using a 100 m x 100 m grid.

The weekday sample of EGT2010 is considered. It was drawn so that at least 4,000 people were interviewed in each of the 8 administrative districts composing the region ("departments"). There were about 120 thousand O-D trips. Their destinations are mapped in Fig. 2 (violet dots) in relation to the RER stations (red dots). They are mainly concentrated in the central part (the Paris city).

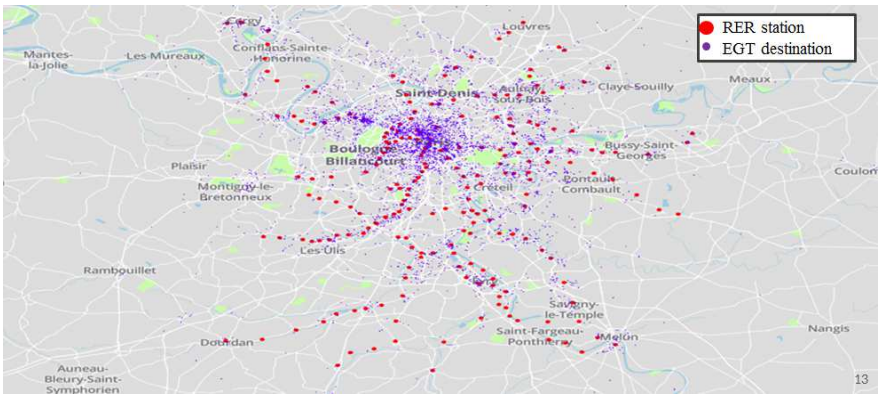


Fig. 2. Geo-locations of the HTS O-D trips: destinations (violet dots) versus the RER stations (red dots).

The GTFS data of Ile-de-France in 2019 are supplied as open source data by IdFM. The GTFS files contain the service routes and timetables, the transfer times and station geo-locations.

Lastly, a file of the 46 officially labeled P&R geo-locations as of 2018 was available open data from IdFM.

3. Generation of P&R related rail O-D trips from survey data

In this study, the intermodal trip connects a Private Vehicle (PV) O-D trip to a PT O-D trip either before the origin station as a go-trip (PV-PT) or after the destination station as a back-trip (PT-PV). A PV O-D trip is a complete journey between one pair of O-D places along a single route of a PV. A PT O-D trip is a complete journey between one pair of O-D stations for an activity including short time intermediate activities in or near an intermediate station. An O-D trip is composed of one or more than one PT trip-legs along the passenger's trajectory (Alsger et al., 2016, 2015; Xie and Leurent, 2020), where a trip-leg is a journey between one pair of O-D stations along a single transit line.

The P&R related rail HTS O-D trips are inferred by matching the HTS trips to the nearest rail transit stations of P&R facilities. This is achieved by a three-step algorithm:

Step 1: Extract the rail HTS O-D trips;

Step 2: Extract the nearest rail stations of P&R facilities by min(l_{p-s}), where l_{p-s} is the Haversine distance between the geo-location of a P&R parking and the nearest rail station;

Step 3: Match the rail HTS O-D trips to the nearest rail stations of P&R facilities by using the 100 m x 100 m cells, min(l_{z-ps}), where l_{z-ps} is the Haversine distance between the center of a cell and the nearest rail station of the P&R parking.

Among the P&R related rail HTS O-D trips, 22,231 are rail only and 1,254 are intermodal trips combining a car mode and a rail mode. Thus the full HTS weekday sample contains 1.01% intermodal trips, 17.9% rail only trips and 81.1% other trips (Fig. 3a). The distribution of intermodal trips per station is shown in Fig. 3b: orange bars for go-

trips, and blue bars for back trips. The numbers of go-trips and back-trips are almost the same, since the cars parked on go-trips need be recovered on back trips.

(a) (b)

Fig. 3. HTS data: (a) Distribution of HTS O-D trip types; (b) Distribution of intermodal trips per station: go-trips (orange) and back-trips (blue).

4. Generation of P&R related rail O-D trips from smartcard data

An AFC record of a smartcard is either a tap-in or a tap-out. To reconstitute O-D trips from the AFC dataset, we need to concatenate all records of tap-in and tap-out for each card ID over the selected day. Two issues of missing information in the Île-de-France AFC data deserve special investigation: (i) the lack of tap-in or tap-out record; and (ii) the lack of transfer information between the branches of the same line or between different lines of a given sub-mode. To cope for both issues, an Activity based Trip-chain Model (ATM) is developed for the semi-closed three-mode network, extending our ATM for the closed mono-modal RER-only network in (Zhou et al., 2019). Two main stages are involved: (1) the inference of O-D pairs, and (2) the inferences of transfers and O-D trips.

4.1. Inference of O-D pairs

Based on the properties of the AFC data in Île-de-France, the data are divided into three cases: Case a, Case b and Case c. The algorithm of the inference of O-D pairs for the three cases is as following:

Case a: A tap-in n is followed by a tap-out $n+1$.
We get directly one O-D pair $(n, n+1)$.

Case b: A tap-in n is followed by a tap-in $n+1$.
The tap-in station $n+1$ is taken as the missing tap-out station of tap-in station n, while the tap-in stations n and $n+1$ are different. If the tap-in staions n and $n+1$ are the same, we pass to next tap-in until the tap-in staions are different:

$$\text{Tap-in } (n, n\epsilon[1, N-1]) \neq \text{Tap-in } (n+p, \min(p\epsilon[1, N-n]))$$
$$\Rightarrow \text{Tap-out } (n, n\epsilon[1, N-1]) = \text{Tap-in } (n+p, \min(p\epsilon[1, N-n])).$$

where N is the total number of taps of a card.

Case c: There lacks the final tap-out of the day.
We suppose that the last trip of the day always returns to the starting station of the day.

Per card ID, the records are processed by a waiting queue algorithm under First-In, First-Out (FIFO) discipline to generate O-D pairs for the three cases. We got about 7.5 million O-D pairs (Fig. 4): most of the O-D pairs fall in Case b. As the travel times of the inferred O-D pairs are not checked in this stage, the inferred O-D pairs will be further adjusted in the next stage according to the PT supply.

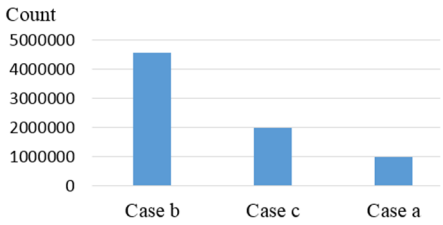
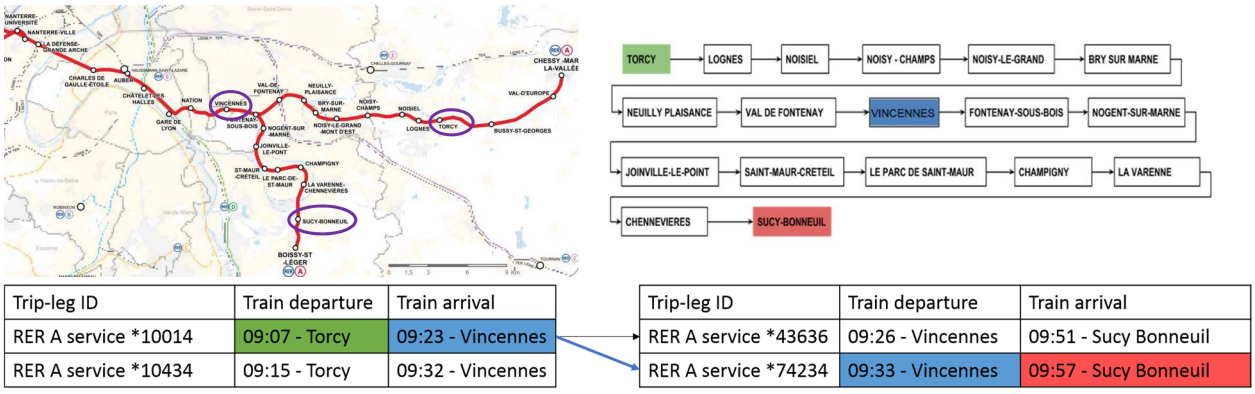


Fig. 4. Inferred O-D pairs.

4.2. Inferences of transfers and O-D trips

A transfer between two successive O-D trip-legs is inferred by combining the generated O-D pairs, the transit network model, and some ad-hoc rules. The transfer inference algorithm involves Dijkstra's algorithm to find out the shortest path on the dynamic bigraph of the rail network. Firstly, the dynamic bigraph of the rail PT network is built up from GTFS data: it contains more links between stations than in the static network model, since the dynamic graph integrates train arrival and departure times from the GTFS timetables. Those times give the temporal attributes of routes and O-D pairs. Secondly, the generated O-D pairs are matched to the dynamic bigraph of the rail network by removing bus O-D pairs of transfer inference. In general, three ad-hoc rules are postulated for adapting to the local mobility characteristics and local transit infrastructure structure:

- Rule 1: The travel time of an O-D trip $\Delta t_{\text{O-D trip}}$ must be less than 2 hours; otherwise, the candidate O-D trip is split into two O-D trips.
- Rule 2: The transfer time $\Delta t_{\text{transfer}}$ lies in [2, 10] min, and the transfer walking distance l_{transfer} lies in [160, 800] m. Otherwise, the candidate O-D trip is split into two O-D trips. When the transfer time $\Delta t_{\text{transfer}}$ is greater than 10 min, it is considered as an activity time. Thus, $\Delta t_{\text{activity}} > 10$ min.
- Rule 3: The least egress time Δt_{egress} is not smaller than 1 min.



Trip-leg ID	Train departure	Train arrival
RER A service *10014	09:07 - Torcy	09:23 - Vincennes
RER A service *10434	09:15 - Torcy	09:32 - Vincennes

Trip-leg ID	Train departure	Train arrival
RER A service *43636	09:26 - Vincennes	09:51 - Sucy Bonneuil
RER A service *74234	09:33 - Vincennes	09:57 - Sucy Bonneuil

Fig. 5. Inferred transfer (Vincennes) between Origin (Torcy) and Destination (Sucy-Bonneuil), and generated O-D trips.

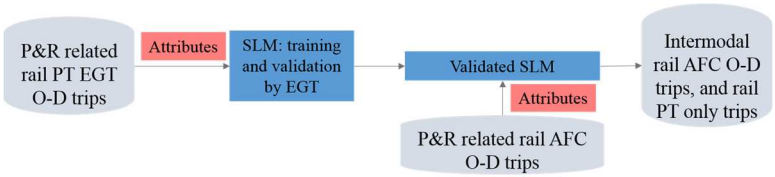
By applying Rule 1, when $\Delta t_{\text{O-D pair}} \leq 2$h, an O-D pair is used to find out an itinerary between the Tap-In and Tap-Out (TITO) times. Otherwise, $\Delta t_{\text{O-D pair}} > 2$h, two O-D trips are generated: one is based on the tap-in time, the tap-in station, the tap-out station, and $\Delta t_{\text{O-D pair}} \leq 2$h; and the other one is based on the tap-out time, the tap-in station, the tap-out station, and $\Delta t_{\text{O-D pair}} \leq 2$h. Furthermore, train run choice and route choice are both considered to find out a unique transfer between two successive trip-legs: this amounts to a dynamic passenger assignment involving Rule 2 for checking the transfer time and the transfer walking distance. All trip-legs of an O-D trip are concatenated as a chain. For example, Fig. 5 introduces the transfer inference between the O-D pair of Torcy (tap-in at 9h05) and Sucy-Bonneuil (tap-out at 9h58). The unique transfer occurs at the Vincennes station where two train runs arrive

from Torcy at 9h23 and 9h32, and two train runs depart at 9h26 and 9h57 towards Sucy-Bonneuil. When more than one train runs are feasible, we choose the last one and recalculate the train circulation time of each trip leg along the O-D trips by applying Rule 3 (see the trip-legs linked by the blue arrow in Fig. 5). We finally got 35,020 P&R related rail AFC O-D trips: 17,734 go-trips and 17,286 back-trips involving the 46 P&R related rail stations.

5. Mining P&R intermodal trips and user types from smartcard data

Since AFC O-D trips do not record passenger intermodal information, the AFC O-D trips are confronted with HTS O-D trips for mining the intermodal trips by a Intermodality Inference Model (IIM). The IIM is a Supervised Learning Model (SLM) based on the common temporal and spatial attributes of the two datasets (Fig. 6a). The common temporal attributes include train departure time at origin station, train arrival time at destination station, and travel time. The common spatial attributes include origin station, destination station, distance between origin and destination, and total number of transfers. The training and validation of the SLM is based on the P&R related rail PT HTS O-D trips.

(a)



(b)

SLM	Accuracy (weighted average)		
	Precision	Recall	F1-score
SVM	0.78	0.77	0.77
Decision Tree	0.92	0.91	0.91
ANN in TensorFlow	0.93	0.74	0.81

Fig. 6. IIM: (a) SLM for internal AFC O-D trip generation; (b) SLM model evaluation metrics.

Three algorithms were considered to build SLMs in the TensorFlow environment: Support Vector Machine (SVM), Decision Tree and Artificial Neural Network. They were cross-compared according to the following model evaluation metrics: precision, recall, and F1-score: the higher the accuracy the better the model. Fig. 6b shows that Decision Tree performed best in this study, so it was further used to predict the intermodal rail AFC O-D trips. Finally, we got 3 025 intermodal rail AFC O-D trips, about 8.64% of the P&R related rail AFC O-D trips.

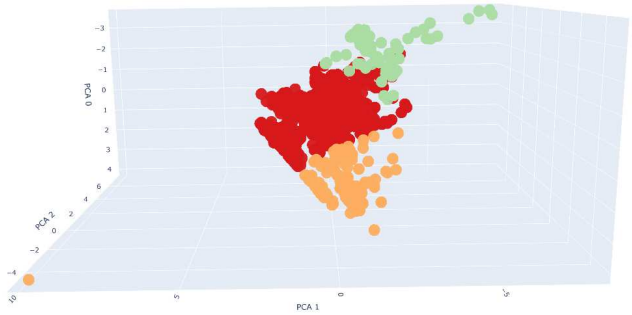


Fig. 7. Types of the intermodal rail AFC O-D trips.

In order to reveal typical usage situations, the intermodal rail AFC O-D trips were classified by an unsupervised learning algorithm of clustering analysis (K-means), based on five attributes per trip: train departure time at origin station, train arrival time at destination statin, travel time, distance between origin and destination, and total number of transfers. The number of clusters, K, can be determined according to the average silhouette value (Rousseeuw, 1987) or the elbow method (Thorndike, 1953). We thus obtained the following K = 3 types of intermodal trips:

- Class 1: Morning peak trips with long travel times, many transfers, and short travel distances. The origins and destinations are very close to Paris city center.
- Class 2: Evening off-peak trips with short travel times, only one or two transfers, and long travel distances.

The origins are very close to Paris city center, but the destinations are relatively far from the city center.

- Class 3: Morning peak trips with long travel times, only one or two transfers, and long travel distances. The origins are far away from Paris urban area, but the destinations are in Paris urban area.

The 3 Classes are illustrated in Fig. 7 by projecting the 5 attributes to a 3D graph (pca0, pca1, pca2) by the Principal Component Analysis (PCA): Class 1 (red), Class 2 (orange) and Class 3 (green).

6. Conclusions and perspectives

We developed a methodology to learn about P&R practices from an HTS and to infer it in a broader, more comprehensive way by using an AFC dataset in combination to a GTFS model of rail services, a dynamic path search algorithm and three ad-hoc identification rules. Data treatment at the trip level is a key methodological step to enrich the AFC dataset with information on the used stations, the passage times and the trip travel time. The resulting trip attributes enable us to "learn" about intermodal practice from the HTS and infer it from the enriched AFC dataset. Machine learning was implemented by testing three supervised learning algorithms, namely SVM, Decision Tree and ANN. Out of them, Decision Tree was selected to infer intermodality for rail trips in the AFC dataset. Furthermore, the resulting inferred trips were clusterized in three groups using the k-means algorithm.

The methodology may certainly be transferred and tailored to study other mobility phenomena based on both HTS and AFC. As for intermodal practice, further research might address a deeper characterization of O-D situations in terms of not only transit conditions but also car conditions on the main trip leg, as well as deeper characterization of feeding conditions such as the transfer walk distance and the average speed of the feeder mode around the rail station. As for the Paris – Ile-de-France case, on-going research involves the more recent HTS (2018) and addresses the full set of railway stations, with ad-hoc characterization of the parking conditions around the station. Lastly, more sophisticated Machine Learning (ML) methods can also be explored to improve the process of IIM, yet a local sensitivity analysis about K for user classification.

Acknowledgements

This research was supported by the Research and Education Chair on Territorial Mobility operated by ENPC in partnership with IdFM, to which the authors are grateful.

References

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

IdFM, 2020. D'ici 2021, 80 Parcs Relais pour faciliter le stationnement aux abords des gares. https//www.iledefrance-mobilites.fr/l-innovation/parcs-relais/ (laste Visit. 4th Feb. 2020) 1 (in French).

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

Kimpton, A., Pojani, D., Sipe, N., Corcoran, J., 2020. Parking Behavior: Park 'n' Ride (PnR) to encourage multimodalism in Brisbane. Land use policy 91.

Leurent, F., Polacchini, A.-R., 1995. Quelques éclairages sur la mobilité des Franciliens. Rapp. sur Conv. DREIF. INRETS, Arcueil, Fance 1-35 (in French).

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

Rousseeuw, P.J.S., 1987. A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18, 267–276. https://doi.org/10.1007/BF02289263

Xie, X., Leurent, F., 2020. Chapter 5 Passenger trajectory generation using modern datasets, in: Book "Towards Sustainable and Economic Smart Mobility: Shaping the Future of Smart Cities", World Scientific Publishing Europe Ltd. pp. 81–98. https://doi.org/10.1142/9781786347862_0005

Zhou, Y., Xie, X., Leurent, F., 2019. Mining regional passenger mobility activity using multiple data sources: A case study in Paris area. TRANSITDATA2019, 5th Int. Work. Symp. 8-9-10 July 2019. Paris, Fr. 1–3.