Clustering Smart Card Data for Urban Mobility Analysis

Mohamed K. El Mahrsi, Etienne Côme, Latifa Oukhellou, and Michel Verleysen

***Abstract*—Smart card data gathered by automated fare collection (AFC) systems are valuable resources for studying urban mobility. In this paper, we propose two approaches to cluster smart card data, which can be used to extract mobility patterns in a public transportation system. Two complementary standpoints are considered: a station-oriented operational point of view and a passenger-focused one. The first approach clusters stations based on when their activity occurs, i.e., how trips made at the stations are distributed over time. The second approach makes it possible to identify groups of passengers that have similar boarding times aggregated into weekly profiles. By applying our approaches to a real data set issued from the metropolitan area of Rennes, France, we illustrate how they can help reveal valuable insights about urban mobility, such as the presence of different station key roles, including residential stations used mostly in the mornings and work stations used only in the evening and almost exclusively during weekdays, as well as different passenger behaviors ranging from the sporadic and diffuse usage to typical commute practices. By cross comparing passenger clusters with fare types, we also highlight how certain usages are more specific to particular types of passengers.**

***Index Terms*—Smart cards, public transport, machine learning, unsupervised learning, clustering methods, generative models.**

I. Introduction

NOWADAYS, various digital traces are collected through different sources such as GPS trajectories, ticketing data of public transportation systems, mobile phone traces, etc. [1]–[3]. The availability of such traces led to the emergence of a new field of research named urban computing, which can be defined as the process of acquiring, integrating, and analyzing voluminous amounts of data coming from heterogeneous sources in urban spaces (sensors, vehicles, pedestrians, etc.) in order to help solve problems from which big cities suffer on a daily basis, such as air pollution and traffic jams [4]. Within this general context, data collected by Automated Fare Collection (AFC) systems in public transit networks of large cities are a valuable resource that can be harnessed to achieve a better understanding of human mobility and evaluate the performance of transportation systems.

AFC systems are currently widely adopted all around the globe to manage payments in public transit networks. Existing implementations include the Navigo pass in Paris, France, the Oyster card in London, UK, the Octopus card in Hong Kong, the Trajeta Bip! card in Santiago, Chile, and many others. At the center of AFC systems are contactless smart cards containing embedded microchips capable of storing and even processing data that passengers use when interacting with the system. Two types of transactions are collected through smart cards: (i) monetary transactions occurring when a cardholder adds credit or renews his travel pass and (ii) journey transactions made when passengers enter stations, board buses, etc. While the original purpose of AFC systems is to automate and manage the various billing operations involved in the fare collection process, the collected data (especially the journey transactions) present an unprecedented opportunity to extract valuable knowledge, which can be used for performance evaluation, transit planning, etc.

Compared to more traditional transport data sources (e.g., surveys and travel diaries), smart card data are:

- More extensive since all the transactions made by cardholders are registered in the system (in contrast with those reported by a small sample of passengers in the case of surveys).
- More accurate since the transactions are often timestamped and geotagged with their exact time and location.
- Traceable at an (anonymized) individual level since each transaction is paired with the card it was made with, making it possible to conduct longitudinal studies on traveler behavior over extended periods of time.

However, using smart card data to analyze human mobility raises challenges due to their:

- Big volume: depending on the size of the network, hundreds of thousands to tens of millions of transactions are registered per day.
- Incompleteness: while origin information is available for most AFC systems (passengers being required to validate their cards when boarding or entering in stations), trip destination information (particularly for trips involving multiple stages) are often missing. Evidently, data about trip purposes are also unavailable. These information play a key role not only in understanding travel behavior

Manuscript received December 10, 2015; revised April 14, 2016; accepted July 30, 2016. Date of publication September 1, 2016; date of current version February 24, 2017. This work was supported by the Predit (Programme de Recherche et d'Innovation dans les Transports Terrestres) Program. The Associate Editor for this paper was J. Li.

M. K. El Mahrsi, E. Côme, and L. Oukhellou are with the Engineering of Surface Transportation Networks and Advanced Computing (GRETTIA) Laboratory, French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR), 77447 Marne-la-Vallée, France (e-mail: mohamed-khalil.el-mahrsi@ifsttar.fr; etienne.come@ifsttar.fr; latifa.oukhellou@ifsttar.fr).

M. Verleysen is with the Machine Learning Group, Institute of Information and Communication Technologies, Electronic and Applied Mathematics, Université Catholique de Louvain, 1348 Louvain-La-Neuve, Belgium.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2600515

(which is believed to be activity-driven) but also in estimating travel demand distributions.
- Lack of socioeconomic data: due to the anonymization process aimed to protect passenger privacy, socioeconomic indicators (age, gender, revenue, etc.) are omitted, despite their usefulness in conducting detailed travel behavior analyses.

Novel mobility data mining methodologies based upon engineering and computer sciences are therefore needed in order to explore smart card data while taking into account the aforementioned aspects.

In this paper, we demonstrate how smart card data can be used to understand a public transportation system based both on how its stations are used and how passengers behave from a temporal standpoint. Identifying key station roles and passenger patterns can help transport operators better know the demand of their customers and propose targeted incentives, services, and tools accordingly. From a city perspective, this may also help redesign and improve existing transportation policies. The contributions of this work are the following:

- We identify key roles played by stations in the public transportation network by partitioning them into different clusters having a similar usage, thus highlighting the relationships between time of day, location and usage. To this effect, we construct count series describing each station's usage and apply a model-based count series clustering to partition the stations according to their usage.
- We study the extraction of passenger travel patterns from smart card data. To this effect, we construct temporal passenger profiles based on boarding information and apply a generative model-based clustering approach to discover groups of passengers who behave similarly with respect to their boarding times. The resulting clusters portray different travel behaviors that turn to be related to distinct trip purposes and can therefore be valuable in achieving a better characterization of travel demand.
- We study how passenger travel habits relate to socioeconomic characteristics. To this end, we cross the passenger clustering results with the fare types of their smart cards.
- We apply our approaches on a real smart card dataset covering four weeks of journey transactions registered in the metropolitan area of the city of Rennes (France).

Since the presented approaches use generative models, the various parameters estimated from the data can be used for simulation and transportation planning purposes by, for example, generating scaled synthetic data that take into account future growth of the population in order to study their repercussions on travel conditions, quality of service, etc. or that can be fed to demand forecasting models (such as the four stage model).

The rest of this paper is organized as follows. Related work is discussed in Section II. The real smart card dataset used for the study is presented in Section III along with a description of how we enrich it by inferring missing information as well as a preliminary analysis based on descriptive statistics. Our approach to clustering stations based on their count statistics and its main findings are detailed in Section IV. We detail our approach to clustering passengers based on their temporal profiles and present its results in Section V. Finally, we conclude the paper in Section VI with general remarks and future research directions.

II. Related Work

The availability of smart card data motivated a considerable amount of research that tackles different questions such as studying travel behavior, trip chains reconstruction and transfer detection, inference of destinations and OD (Origin/Destination) matrices, cluster analysis, etc. Apart from the differences in the finalities and the proposed approaches of all these works, they also consider heterogeneous public transport networks. Particularly, some studies focus on a single mode of transportation (e.g., buses only) whereas others consider multimodal transport networks that offer the possibility to travel using different modes (bus, subway, ferry, etc.). Additionally, in some cases the passengers are required to both tap in (when entering stations or boarding buses) and tap out (when exiting stations and alighting buses) which makes the information about the destination of each trip available for the study, whereas in other cases only tap ins are registered which results in the unavailability of destination information. In some cases, even origin information are missing.

In order to provide a complete picture of the context of our work, we start by briefly discussing travel demand forecasting in Section II-A. Early work on smart card data is presented in Section II-B. Preprocessing and enrichment techniques are presented in Section II-C. A brief tour of work studying smart card data through their descriptive statistics is presented in Section II-D. Propositions involving advanced knowledge extraction techniques are discussed in Section II-E. A positioning of our approaches with respect to these propositions is presented in the same section.

For extensive literature reviews on urban computing in general and on smart card technologies and their implications in public transportation (from the strategic, planning, and operational standpoints) in particular, we refer the reader to [4] and [5], respectively.

A. Travel Demand Forecasting

Demand modeling for personal travel has long been dominated by the trip-based four step models (4SM) approach [6] in which trip frequencies are first determined for a given set of trip purposes (e.g., home-based work, home-based non-work trips, etc.) based on trip production and zonal attraction models (that reflects characteristics such as land use, household demographics, and socio-economic indicators). These trip frequencies are then used to generate various trip tables (during trip distribution and subsequent steps) based on the underlying transport network's attributes (inter-zonal travel durations, etc.). Alternatively, activity-based models [7] constitute a second family of approaches that embrase the philosophy that travel decisions are activity-driven and form a complete agenda out of which they cannot be analyzed (i.e. the analysis cannot be conducted on an individual trip basis). Both types of models

rely essentially on survey data (household travel and activity-based surveys, travel diaries, etc.) for calibration. With the advent AFC systems, smart card data can be used to complement surveys and help provide better and more reliable data for these models to work with.

B. Early Work on Smart Card Data

Bagchi and White [8], [9] were among the first researchers to substantiate the potential of smart card data for transit planning. The authors apply rules-based processing to bus transactions in order to infer turnover rates, trip rates, and detect trip chains from the collected data. They also emphasize that since some information are not captured by smart card data (e.g., journey purpose, satisfaction with respect to the transport service, etc.), the latter cannot entirely supersede existing data collection approaches (mainly direct surveys) but rather complement them. Utsunomiya *et al.* [10] investigate the factors influencing the access distance (i.e., distance from the home address of a passenger to the station where he makes his first boarding of the day) and study usage regularity and consistency using bus and rail transactions. They also stress on how the presence of errors and missing information influence the quality of the data and suggest that the readability of the latter can be further improved by enriching them with socioeconomic information, destinations, trip purposes, etc.

C. Data Preprocessing and Enrichment

The inference of alighting locations when they are not directly captured in smart card transactions was discussed in [11]–[14]. Most of the proposed approaches rely on fairly similar assumptions. The distance to the next boarding is the main criteria for the assignment: the passengers are assumed to have rational behavior and alight at the station or bus stop (along the current line) that is closest and within reasonable walking distance to their next boarding location. The alighting location of the last trip of the day is estimated using either the first boarding of the same day or the one from the next day.

A trip chain (or linked trip) regroups two or more transactions that are part of a same "logical" journey (e.g., a passenger going from home to work). Reconstructing these chains requires identifying transfers. At the most basic level, this is conducted using a fixed time threshold [8], [9], [11], [15], [16]: validations that occur within a given timeframe (e.g., 30 minutes) are simply considered to belong to a same journey. Seaborn *et al.* [17] use separate thresholds to account for the nature of the transfer and whether alightings are available or not. A similar approach is also adopted in [18]. An alternative method that does not rely on time thresholds is reported in [19]. Instead, the authors use operations information to associate each boarding transaction to a bus run and a stop. Alighting locations are estimated with the assumption that the alighting for a given boarding is the closest bus stop (along the run) to the location where the next boarding occurs. If both stops are within reasonable walking distance and the boarded bus routes are not the same, then a transfer is detected and both transactions are considered as part of a same trip chain.

Once both origin and alighting locations are known and trip chains are reconstructed, the data can be used to estimate OD matrices [11], [13], [20], assign anchor points (frequently visited locations) [21], etc. Recent studies [14], [22], [23] concluded that threshold-based transfer and destination inference approaches are well robust in different settings. In particular, they have shown that the underlying assumptions are well grounded, that varying the involved thresholds within an interval of reasonable values (e.g., increasing the allowable transfer time from 15 min to 90 min in [22]) has minimal effect on the produced results, and that such approaches tend to be accurate (a 79% success rate in inferring destinations using a 400 m reasonable walking distance threshold is reported in [23]).

Another equally important enrichment is the inference of trip purposes (which play a key role in characterizing travel behavior and consequently providing novel and well-adapted transportation services). Several approaches were proposed to this end using a pre-defined set of rules (based on travel time, fare type, activity location and duration, etc.) [24], Naïve Bayes classifiers [25], Continuous Hidden Markov Models [26], etc.

D. Studying Smart Card Data Through Descriptive Statistics

Smart card data were used to study different facets of mobility in public transportation. Morency *et al.* [27] analyze the variability of travel behaviors based on activity rates, the number of boardings per day and the number of different boarding stations observed through bus trip data. Fuse *et al.* [28] use bus smart card data to determine travel time and bus loads that can in turn be used for congestion spot analysis and the improvement of bus stops planning. Lathia *et al.* [29] conduct a comparison between smart card data and the results of an online survey in order to characterize the differences between the perceived and the actual behavior of passengers and their reaction to travel incentives. Various aspects are inspected such as trips per day frequency and regularity, atypicality of travel modality and origin and destination stations, as well as cash-fare purchasing habits. Tran [30] analyses the behavioral difference between travel card holder and pay as you go passengers based on journey duration and travel extent. The author also proposes a classification of passengers based on the average number of trips on each route and the number of unique journeys they made. Lathia *et al.* [31] study community well-being as captured by smart card data: stations are mapped, based on geographic proximity, to communities and IMD (Index of Multiple Deprivation) scores obtained from national census results; trip data are used to compute a station-by-station flow matrix representing locations visited by different communities. Different indices are then inspected to analyze the correlation between the IMD and the passengers' flow.

E. Advanced Knowledge Extraction From Smart Card Data

In order to extract further knowledge involving group behavior, frequent patterns, etc. more advanced data analysis techniques (e.g. clustering and classification) need to be used. Trépanier *et al.* [32] study the loyalty of public transport users by applying a hazard model to smart card data. In [18],

DBSCAN [33] is applied to individual trip chains in order to retrieve each passenger's recurrent travel patterns. Additionally, k-means++ is used to cluster passengers based on regularity. The latter relies on four descriptive features of the passengers (number of travel days, number of similar first boarding times, number of similar route sequences, and number of similar stop sequences). In a similar fashion, Kieu *et al.* [16] use k-means (based on the number of trip chains) to separate infrequent from frequent passengers and apply DBSCAN to the latter group in order to further divide it based on boarding and alighting time and location regularity. DBSCAN is also used in [34] in order to segment passengers (based on their habitual travel times and the origins and destinations of their trips) into four distinct groups. In [27], clustering is used to study individual travel regularity: the trips of a given passenger are aggregated into a daily profile indicating for each time bin (a given hour) if at least one boarding was registered. k-means is applied in order to identify clusters of similar days with respect to boarding times. A similar analysis of weekly travel behavior is conducted in [35]: bus trips are aggregated into weekly profiles that include the 5 weekday (thus excluding the weekend) activity of a passenger. Hierarchical Agglomerative Clustering (HAC) and k-means are applied to the transactions in order to study group behavior. Lathia *et al.* [36] apply hierarchical agglomerative clustering on passenger weekday profiles (trip counts over five time bins within the day) in order to uncover different travel behaviors (e.g., typical commutes, evening-only travel, etc.) and motivate the need for using smart card data to build user-tailored transport information services. The authors also contribute a number of predictive models aimed to take advantage of a passenger's travel history to provide personalized travel time estimates. Ceapa *et al.* [37] use smart card data to study station congestion patterns. Individual trips are transformed into per-station data where each observation contains the difference between entries and exits for a given station over 2-minutes intervals. Dynamic Time Warping (DTW) and hierarchical clustering are used to regroup stations based on their usage. The authors also contribute three classification techniques to predict station crowdedness. Recently, Poussevin *et al.* [38] used NMF (Nonnegative Matrix Factorization) to discover a dictionary of behavioral atoms to describe passengers based on their subway journey transactions. The distribution of these atoms over the stations is then used to conduct multi-scale clustering and retrieve groups of stations with similar behavior. Goulet-Langlois [39] study patterns in longitudinal representations of travel activity (spanning over 4 weeks). Each passenger is described through a sequence of activities (inferred from his smart card data) spanning 1 h each. The sequences are then projected into a low-dimensionality space using PCA (Principal Component Analysis) and clustered using k-means in order to discover working day clusters, homebound clusters, etc. Associations between these patterns and demographic attributes (age, income, occupation, etc.) are also studied.

Other advanced mining techniques that do not fall under the clustering umbrella include work on modeling the spatial distribution of passengers using preferential selection of visited locations based on their popularity [40], reconstruction of individual mobility history using collaborative space alignment and filtering with Conditional Random Fields (CRF) [41], using the flow-comap technique to visualize passenger flow patterns [42], predicting bus riderships and studying their influential factors [43], etc.

The two main contributions of this paper, which concern station clustering and passenger profile clustering, differ from existing literature on smart card data clustering on the following aspects:

- Adoption of a generative, model-based approach: most approaches proposed in the literature rely on classic clustering algorithms such as k-means, DBSCAN, etc. for which an appropriate distance measure (such as the Euclidian distance) need to be specified explicitly (the choice of said distance measure requires, generally, significant domain expertise). In contrast, our approach tries to maximize the likelihood of a statistical model describing the distribution of the data. Model-based approaches are considered to be more interpretable than similarity-based approaches [44]. Additionally, the use of generative models makes it possible to use the parameters of the estimated models for simulation purposes (e.g. generate trip data under different conditions and use them in conjunction with for travel demand models for forecasting purposes).
- Choice of representation: in the case of passenger clustering, existing approaches tend to adopt coarse representations in which weekdays are neglected, all weekdays are flattened into a single daily profile, etc. Our approach to clustering passengers with respect to their temporal behavior is based on temporal profiles that describe passengers using a finer granularity than the ones used in previous studies, which enables a more detailed analysis of the passengers' behaviors and detecting subtle changes between weekdays as we will show in Section V-C.

III. Case Study Data Set

In this section, we present the smart card dataset that we use for our study, discuss how we enrich it by inferring alighting locations and detecting transfers, and study some basic aspects of mobility through descriptive statistics of the data.

A. Data Set

We conduct our study on smart card data collected through the automated fare collection system of the Service des Transports en commun de l'Agglomération Rennaise (STAR). STAR operates over 70 regular bus lines (excluding school bus and complementary services) and 1 subway line serving the metropolitan area of Rennes, France. The operator established its automated fare collection system on March 1st, 2006 and offers the possibility to travel on its network using a KorriGo smart card.

The original dataset spans over a one-month period (April 2014) and contains a total of 5404096 journey transactions out of which 4325839 (80% of the data) were made by 134979 smart cards, whereas the remaining ones were made using traditional paper tickets. Each transaction contains an

anonymized passenger id (only for transactions made using smart cards), the timestamp when the transaction occurred (date and time rounded to the minute), the name and identifier of the subway station or bus stop where the transaction took place, the name and identifier of the boarded bus or subway line, as well as information about the fare type. Additionally, in the case of bus transactions the travel direction (inbound or outbound) is also indicated. The AFC system requires passengers to validate their smart cards only upon entering a subway station or boarding a bus. As a consequence, alighting locations are not collected. In order to protect the privacy of cardholders, no personal information regarding the passengers were made available to us.

B. Data Enrichment Methodology

In order to be able to conduct the passenger cluster analysis that we present in Section V, we first need to reconstruct trip chains from the transactions with smart cards. In fact, if raw transactions are used to characterize travel behavior, passengers engaging in multi-stage trips (i.e., trips involving transfers) can be perceived as more active than passengers with mainly single-stage trips. This might in turn introduce bias during the clustering process.

We reconstruct trip chains by using a two-step approach similarly to previous work [11]–[14]. The first step consists in inferring the alighting location of each transaction based on two assumptions: (i) closest-stop assumption: for a given transaction, the passenger presumably alights at the stop or station closest to where his next transaction takes place, and (ii) daily-symmetry assumption: for the last transaction of the day the passenger alights at the stop or station closest to the location where his very first transaction of the day took place. For each transaction, the distances from all the stations on the current route to the boarding location of the subsequent transaction are inspected and the closest station is retained as the candidate destination of the current transaction. If the candidate destination and the next boarding location are within a reasonable walking distance (fixed to 500 m for this study) then the candidate destination is assigned as the alighting location of the current transaction. Otherwise, the inference fails and no destination is assigned. The same process is applied to the last transaction of the day with the exception of using the boarding location from the first transaction of the day instead of the subsequent transaction. Additionally, an estimate arrival (alighting) time is assigned to each transaction for which an alighting location was estimated.

Determining whether transactions are transfers or not is done in the second step. Here again, each passenger's transactions are inspected sequentially. In order for a transaction to be marked as a transfer, the following conditions must be met: (i) the alighting location of the previous transaction was successfully inferred and (ii) the connection between both transactions (i.e., the elapsed time between the estimate arrival time in the previous transaction and the boarding time of the current transaction) occurs within a time threshold of 30 min. Otherwise, the transaction is marked as a first boarding in order to indicate the start of a new trip chain. Additionally, transfers along the same route are prohibited: if two consecutive transactions are made on the same route, the second is automatically marked as a first boarding even if both the aforementioned conditions are met.

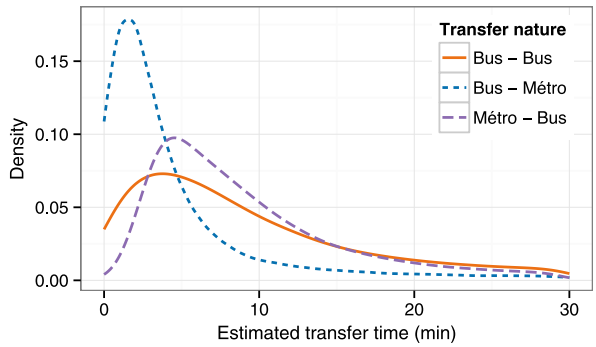

Fig. 1. Densities of the estimated transfer times for each type of transfer (bus to bus, metro to bus, and bus to metro).

Due to the absence of ground-truth information about trip destinations and transfer behavior in the dataset, the effectiveness of the applied data enrichment approach and its influence on the posterior analyses we conduct on the data (mainly the passenger clustering) cannot be evaluated directly and empirically. Nevertheless, as mentioned earlier in Section II-C, recent studies where such information is available [14], [22], [23] have demonstrated the robustness of such approaches. In our case, the assumed reasonable walking distance threshold (500 m) and time threshold (30 min) were fixed upon discussion with experts from the STAR public transportation operator and are in agreement with threshold values usually used in the literature. Fig. 1 illustrates the density distributions of our estimates of transfer times for each of the three possible transfer types in the STAR network. Most transfer times occur quite before the 30 min threshold (95% of transfers occur within 25 min) with transfers to metro occurring in a shorter span (90% occur within 10 min) compared to transfers to bus. In light of these results, which are coherent with those reported in the literature, we are confident in the chosen threshold values and we expect that changing them slightly would only have a marginal impact on the clustering results.

C. Preliminary Analysis of the Enriched Data

The enrichment approach detailed in the previous section was able to estimate the alighting locations for 75.11% of the transactions. The trip chain reconstruction step marked 81.56% of the transactions as first boardings which translates in 3528316 trip chains that we refer to as journeys hereafter. At this stage, we can already conduct a preliminary analysis on the enriched data to study basic aspects of the mobility of smart card passengers. Fig. 2 shows the hourly distribution of journeys made during the week from April, 7 to April, 13. Weekdays are characterized by three peaks occurring in the morning (7–8 am), midday (12–1 pm), and in the evening (4–6 pm). The midday peak is more important on Wednesday than other weekdays. This is mainly due to the fact that in France, course hours do not generally go past 12 pm in middle and high schools. During the weekend, the number of journeys is, as expected, lower than during weekdays. The three-peaks characterizing the latter disappear in favor of a steady increase in

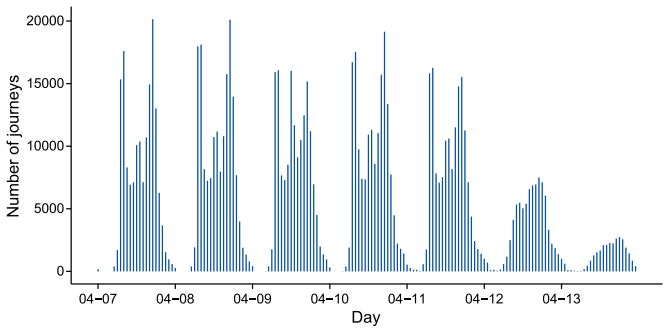

Fig. 2. Hourly distribution of journeys during the week from April 7 to 13.

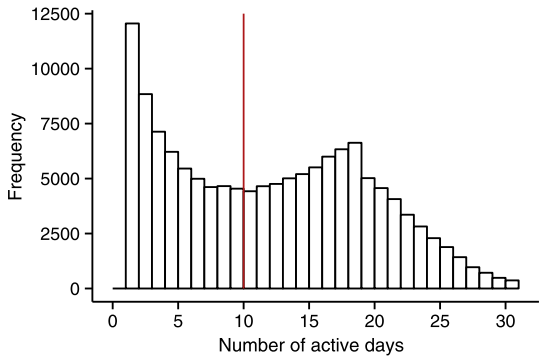

Fig. 3. Distribution of active travel days during which each smart card was used for the period from April 1 to 30. (Red Line) Ten-day mark at which the first infliction occurs and public transportation reliance starts to increase.

demand that starts slightly later (9–10 am), peaks in the evening (5–7 pm), then decreases until the end of the service. This aggregate view of the system depicts an average passenger as someone who travels regularly during the peak hours of all weekdays while relying less on public transportation during the weekends. In Section V, we will show how passenger clustering can help retrieve richer and more varied patterns.

The distribution of active travel days during the period of the study is depicted in Fig. 3. An active travel day is simply a day during which a given smart card was used to make at least one journey. A considerable number of cardholders travel using public transportation occasionally with around 25% of them having less than 5 active days. The number of passengers using their smart cards starts by decreasing before hitting an inflection point at around 10 active travel days, up from which we observe that passengers become more reliant on public transportation. This trend is again inverted once we hit the barrier of 19 active travel days where the number of smart cards starts decreasing as the number of active days increases.

The enriched data can also be used to study other aspects such as travel times, the nature and frequency of transfers, etc. However, since these are not particularly relevant to the main work exposed in this paper, we refrain from presenting them.

IV. Exploring Smart Card Data Through Station Clustering

A first portrait of the public transportation network can be drawn by analyzing how the different stations and bus stops are used through time. This can be done by clustering stations based on the number of validations they receive from passengers in order to reveal groups of stations with similar usage profiles, which is discussed in the present section.

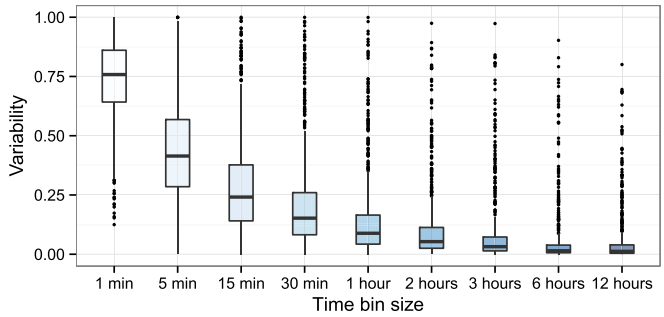

Fig. 4. Distribution of the variability of the number of transactions observed in stations for time bins ranging from 1 min to 12 h. When the bin size is increased, the variability is reduced, which suggests that mobility patterns become more apparent and relevant.

A. Station Clustering Approach

The clustering approach we use is an adaptation of the BSS (Bicycle Sharing System) station clustering approach reported in [45]. Therefore, we only give a brief overview of its main steps and refer the interested reader to the corresponding paper. First, we clean the data from erroneous transactions attributed to unknown locations (due to positioning errors, etc.). This leaves us with 5 294 672 transactions (i.e., 98% of the original dataset) made across 686 stations and bus stops. For each station, the raw data are transformed into transaction counts per one-hour bins over each day in the dataset. A given station's description for a given day $d \in \{1, \ldots, D\}$ (denoted $\mathbf{s}_d$) is expressed as follows:

$$\mathbf{s}_d = (s_{d1}, s_{d2}, \ldots, s_{dh}, \ldots, s_{dH})$$

with D the number of available days (30 in our case) in the dataset, $h \in \{1, 2, \ldots, H\}$ the hour of the day, and s_{dh} the number of transactions (both using tickets and smart cards) registered during hour h on day d. In order to decide the size of the time bins, we conducted a study of the variability of the number of validations observed in each station, similarly to the methodology described in [46]. We vary the size of the time bin from 1 min to 12 hours. Each time, we construct for each station a distinct temporal profile for each of the four weeks of the study, corresponding to transaction counts per daytype and time bin observed for that week. The correlation between the four profiles is then used in order to assess their variability (the mathematical details can be retrieved from [46]). The distributions of variabilities for each of the time bins we considered are illustrated in Fig. 4. As expected, the variability decreases as the size of the time bin is increased (indicating, as suggested in [46], that the mobility patterns become more predictable and apparent). A significant decrease is observed when increasing the size from 1 min (original granularity of the data) to 1 hour, up from which the decrease becomes less pronounced. Consequently, we consider one-hour time bins for our study.

Using a Poisson mixture, we build a model based on the station usage counts. The model relies on two sets of variables, $Z_{\mathbf{s}}$ which corresponds to indicator latent variables defining the memberships of stations to one of K clusters, and W_d which contains observed variables attached to days and encoding the differences between weekdays and weekends (which have considerably different usage profiles). The model is expressed as follows:

$$
\begin{aligned}
&Z_{\mathbf{s}} \sim \mathcal{M}(1, \pi) \\
&s_{d1} \perp\!\!\!\perp s_{d2} \perp\!\!\!\perp \cdots \perp\!\!\!\perp s_{dH} \mid (Z_{\mathbf{s}k} W_{dl} = 1) \\
&s_{dt} \mid (Z_{\mathbf{s}k} W_{dl} = 1) \sim \mathcal{P}(\alpha_{\mathbf{s}} \lambda_{klt}).
\end{aligned}
$$

In this simple model, we express station cluster memberships using a multinomial distribution of parameter $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ (that specifies cluster proportions) which is a classic way in statistics of representing that the result of a trial (here the cluster of a given station) belongs to exactly one of K categories (clusters).

Given a station's cluster and the type of the day, we suppose that transaction counts occurring at different time bins are conditionally independent. This is a simplification that is wildly assumed in the literature and that we adopt in order to facilitate the estimation of the model and keep it easy to interpret. In reality, riderships do correlate across hours of the same day.

We also assume that transaction counts follow a Poisson distribution of parameter $\alpha_{\mathbf{s}} \lambda_{klt}$. $\alpha_{\mathbf{s}}$ is a scaling factor that captures the station $\mathbf{s}$'s global activity which makes it possible to regroup stations with similar activity silhouettes even if the total volume of their transactions varies. The parameters λ_{klt} capture the temporal variations of transactions and vary depending on the station cluster and day type. Here again, the choice of Poisson distributions at this second level is made in order to keep the model parsimonious (i.e. it does not contain an excessive number of parameters) and consequently easily interpretable, since our aim in this work is to cluster the stations for exploratory purposes, rather than model their behavior with extreme finesse. Alternative statistical distributions such as negative binomials can be tested (in this particular case, this leads to doubling the number of parameters for the second layer of the model and complicates their estimation considerably).

The model's parameters are estimated using a custom EM (Expectation Maximization) algorithm which is fully detailed along with additional constraints imposed on the model's parameters in [45]. Estimating the model, however, requires fixing the number of clusters K beforehand. To select an appropriate number of clusters, penalized likelihood criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used and asymptotically consistent but they are also known to be less efficient in practical situations than on simulated cases. To overcome this drawback in real situations, Birge *et al.* [47] have recently proposed a data-driven technique, called the "slope heuristic," to calibrate the penalty involving penalized criteria. The slope heuristic was first proposed in the context of Gaussian homoscedastic least squares regression and was since used in different situations, including model-based clustering. Birge *et al.* [47] proved the existence of a minimal penalty and that considering a penalty equal to twice this minimal penalty allows to approximate the oracle model in terms of risk. The minimal penalty is estimated in practice by the slope of the linear part of the objective function with regard to the model's complexity. A detailed overview and advices for implementation are given in [48]. We set K by first running the EM algorithm while varying K from 2 to 25, then using the slope heuristic to pick an appropriate value and retrieve the model that best fits our data. Using this approach we retrieve 14 station clusters as indicated in Fig. 5.

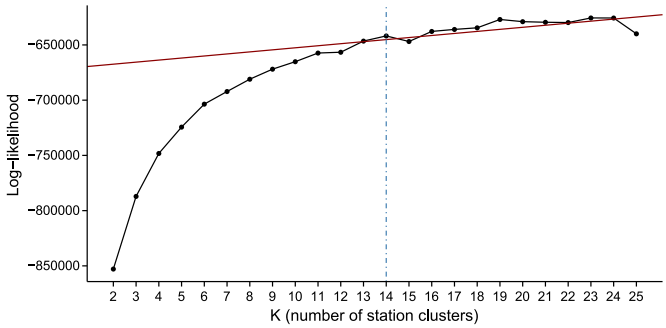

Fig. 5. Evolution of the log likelihood as a function of the number of station clusters $K = 2, \ldots, 25$. (Red Line) Linear model fitted to the linear part of the curve. (Blue Vertical Line) Suitable number of clusters $K = 14$.

B. Results

The retrieved clusters can be studied based on their temporal activity profiles (given by the λ parameters of the model). Under this angle, the clusters can be broken in two main categories:

- Stations with mostly "balanced" usage during the day with several peaks occurring during rush hours. Most station clusters fall under this category.
- Stations with unbalanced usage, in which the number of transactions during one half of the day drastically differs from the other half (e.g. stations heavily used in the morning but not in the evening).

In what follows, we discuss some of the most interesting station clusters that we retrieved from the data. Fig. 6 shows the activity profiles of three station clusters that are intensively used during the morning period (7–8 am) of weekdays. Comparatively to the clusters 1 and 14 [see Fig. 6(a) and (c)] which continue to register activity afterwards, cluster 9 [cf. Fig. 6(b)] is the least used during the other periods of the day. Weekend activity in this cluster is also lower than the other two. A map of the stations belonging to the three clusters is shown in Fig. 7 and shows that they are mainly located in residential areas of the metropolitan area of Rennes. This suggests that the three clusters regroup "housing" stations that are mainly used by passengers to commute to their work. The stations in clusters 1 and 9 are exclusively located in the outskirts of the city, especially in remote towns, whereas cluster 14 regroups stations both in and out of the city.

The opposite behavior occurs in the station cluster shown in Fig. 8. The cluster's profile [cf. Fig. 8(a)] indicates an activity

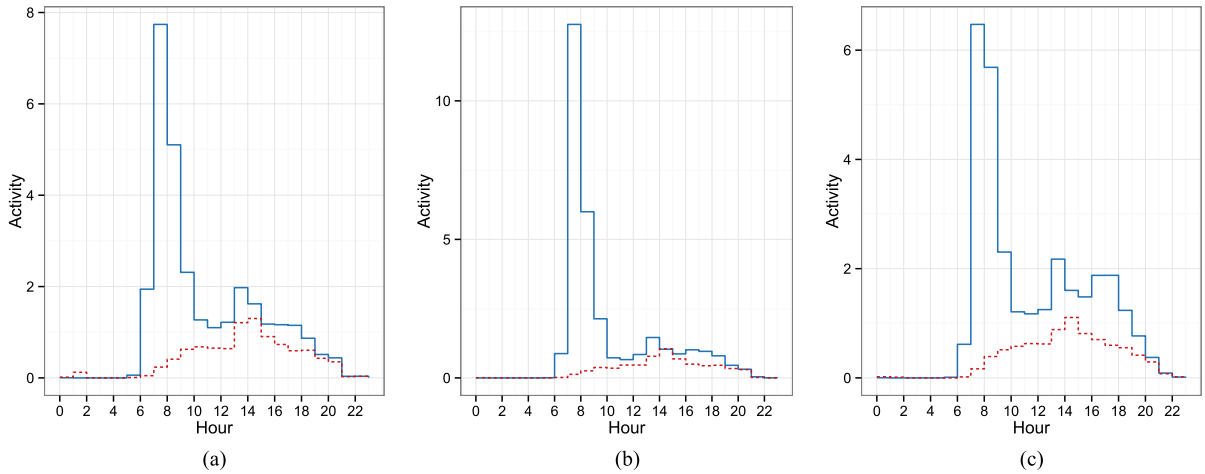

Fig. 6. Activity profiles of three station clusters characterized by an important peak in the morning during weekdays and comparatively low activity in the second part of the day. (Solid Blue Line) Weekday activity. (Red Dashed Line) Weekend activity. The scales of the activity axis are set independently for each cluster in order to make their respective temporal profiles more apparent. (a) Cluster 1: 52 stations (7.58%). (b) Cluster 9: 106 stations (15.45%). (c) Cluster 14: 115 stations (16.76%).

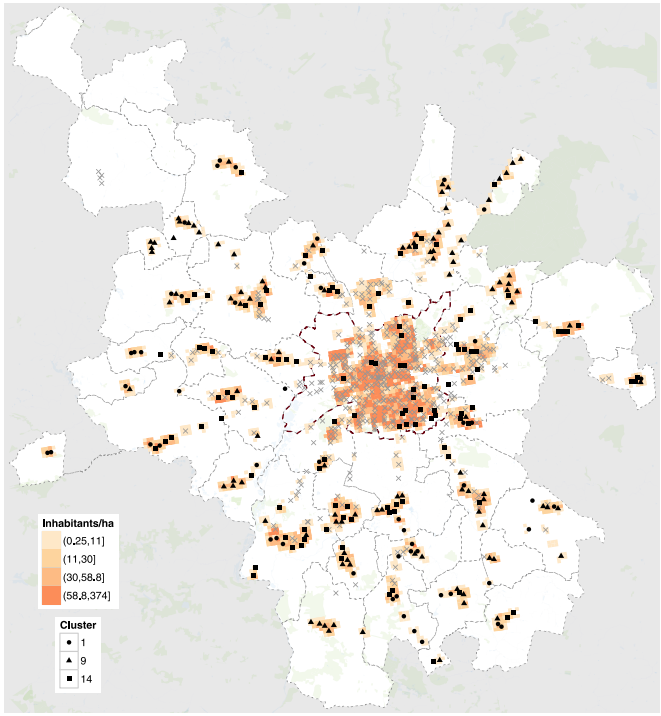

Fig. 7. Map of the stations belonging to clusters 1 (circles), 9 (triangles), and 14 (squares). (Gray Cross) Stations that do not belong to either of the three clusters. The background represents the metropolitan area of Rennes (with the city itself indicated by the red contour), water and riverbanks (light blue), and green areas (light green). (Orange Overlay) Density of the inhabitants in the area.

that is centered around the afternoon with a first peak occurring at 12 pm and a second more intense rush during the evening peak (4–6 pm). The cluster is also characterized by a very low activity during the weekend. This suggests that the involved stations are "work" stations located in industrial areas and activity zones of the city to which passengers travel in the morning to work then use in the evening to commute back home. This claim is confirmed in Fig. 8(b) which shows that most of the cluster's members are indeed located in such places.

The activity profiles of the remaining clusters are shown in Fig. 9. Multiple clusters show an activity that involves two or three peaks centered around rush hours, which is the case for cluster 3 [see Fig. 9(b)], cluster 5 [see Fig. 9(d)], cluster 13 [see Fig. 9(j)], etc. Among those, cluster 5 is particularly interesting since its stations are very active throughout the whole day (i.e., even outside of peak hours) during weekdays. Most stations in this cluster (cf. Fig. 10) are located in the city of Rennes itself which confirms the polycentric operation of the metropolitan area in which the center plays a key role as the backbone of the transportation system. Some clusters, such as cluster 2 [see Fig. 9(a)] and cluster 4 [see Fig. 9(c)], present important activities during the weekend which suggests that they might be located in leisure and recreational spots in the city. In contrast, cluster 10 (much like cluster 11 shown earlier) shows little activity during the weekend.

In some cases, the retrieved clusters can have quite similar activities. This is, for instance, the case of clusters 1 and 14 (see Fig. 6) or clusters 3, 7, and 8 (see Fig. 9). This is due to the choice of the number of clusters: when this number is increased, the approach starts detecting very subtle differences (e.g., the presence of late night time activity in clusters 3 and 7 during the weekend, contrary to cluster 8), whereas decreasing it results in fewer but more coarse clusters.

Extracting station roles and behavior using smart card data as shown in this section is an important step in the direction of characterizing the demand in the city's public transportation system and can be helpful for transport authorities in order to decide on future planing and restructuring of the network. In the following section, we complement this view of the system by studying public transportation usage from the passengers' perspective and extracting the latter's frequent travel patterns.

V. Clustering Passengers Based on Temporal Behavior

In this section, we present our approach to discovering groups of passengers who exhibit similar behaviors from a

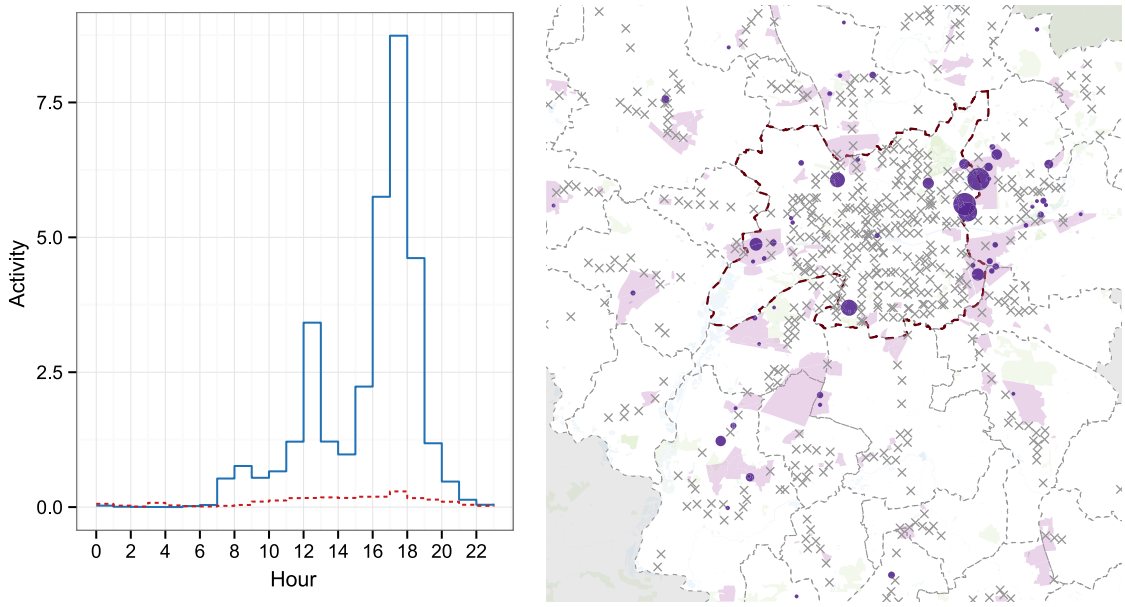

Fig. 8. (a) Activity profiles during weekdays and the weekend. (b) Map of station cluster 11. (Purple Overlay) Activity zones (industrial areas, offices, etc.). (Dots) Stations in the cluster (each dot's area is proportional to its station $\mathbf{s}$'s scaling factor $\alpha_{\mathbf{s}}$). The cluster contains 58 stations (8.45% of all the stations). (Gray Cross) Locations of stations that are not included in the cluster.

purely temporal standpoint (i.e., passengers taking public transportation at the same times without accounting for the boarding locations). Intuitively, the discovery of these groups can help identify frequent patterns in the way passengers use public transit and characterize the demand accordingly. Since we are interested in studying travel behavior over time, ticket transactions cannot be used since they lack the information about the passengers who made them. Consequently, we limit our scope to the 3 528 316 journeys made using smart cards only and retrieved by applying the enrichment approach presented in Section III-B.

A. Passenger Filtering and Temporal Profiles Construction

In order to discover meaningful clusters of passengers, the latter must be observed for a sufficient amount of time: occasional passengers with an insufficient number of active travel days are not very informative when looking for travel patterns. To address this issue, we filter passengers based on active travel days: based on the first inflection point noticed in Fig. 3, we consider passengers with ten or more active travel days during the one-month period of the study as frequent travelers and retain them for our cluster analysis: 3 096 146 trips (87.75% of the total number of trips with smart cards) made by 76478 passengers (56.65% of the total number of passengers) are retained whereas only 12.25% of the trips is discarded. Our early experiments have shown that including the unretained passengers degraded the clustering quality which is to be expected due to the fact that these passengers were not observed sufficiently in order to be able to extract any relevant patterns from their scarce trips.

For each passenger, we build an aggregate "weekly profile" view describing the distribution of all his journeys over each hour (0 through 23) of each day of the week (Monday through Sunday). Therefore, each passenger is an observation over 168 variables: the first variable is the number of trips he took on Monday 0 to 1 am, the second is the number of his trips on Monday 1 to 2 am, and so on. We denote one such profile by $\mathbf{u}$.

B. Clustering Approach

The clustering step of the approach relies on estimating a mixture of unigrams model [49] from the retained passengers' temporal profiles. This approach is often used to cluster documents in the context of information retrieval. Under this perspective, each of the M passengers retained for clustering can be regarded as a "document" containing a collection of N "words." Each word, in our case, is a combination of a day and an hour (e.g. Friday 10 am). Therefore, the used vocabulary's size $D = 7 \times 24 = 168$. Consequently, the temporal profile of the ith passenger, denoted $\mathbf{u}_i$, is the vector of word counts (or frequencies):

$$\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iD}).$$

First, the membership of a passenger to one of K clusters is determined using a multinomial distribution:

$$z \sim \mathcal{M}(1, \pi).$$

$z = (z_1, z_2, \ldots, z_K)^T \in \{0, 1\}^K$ is the component indicator vector and $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ is the vector of cluster proportions. The N words (pairs of day and hour) of the passenger's profile $\mathbf{u}$ are then drawn from the conditional multinomial distribution relative to z, according to the following formula:

$$\mathbf{u} \,|(z_k = 1) \sim \mathcal{M}(N, \beta_k)$$

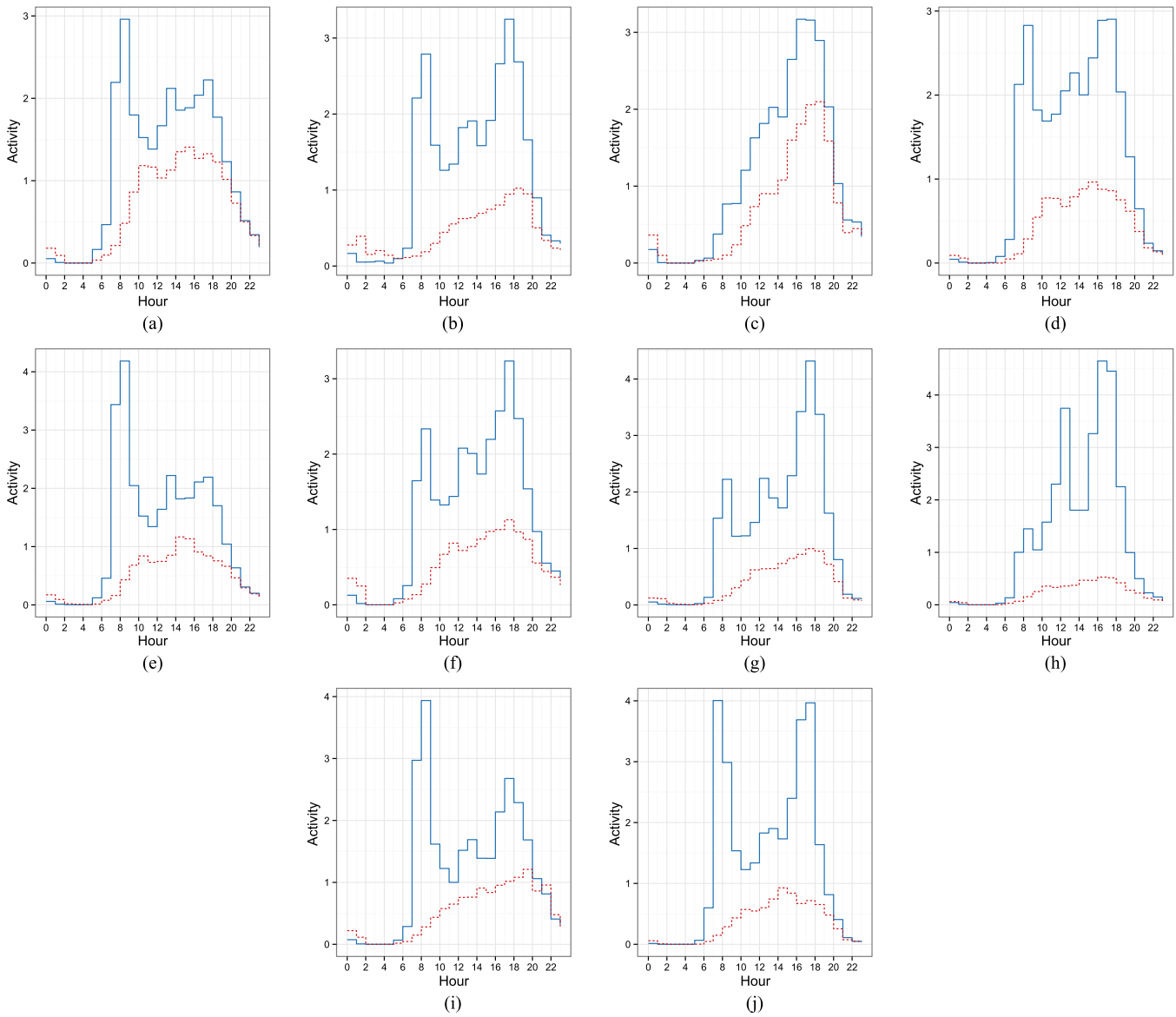

Fig. 9. Activity profiles of the remaining station clusters. The activity axis's scale is set independently for each cluster in order to make its profile more apparent. (a) Cluster 2 (2.19%). (b) Cluster 3 (2.04%). (c) Cluster 4 (2.48%). (d) Cluster 5 (9.77%). (e) Cluster 6 (10.45%). (f) Cluster 7 (2.48%). (g) Cluster 8 (8.02%). (h) Cluster 10 (5.68%). (i) Cluster 12 (0.73%). (j) Cluster 13 (7.73%).

with $\beta_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kd})$ the kth cluster's profile (word proportions) and N the total number of the journeys made by the passenger. The parameters π and β are estimated from the data using a classical Expectation-Maximization (EM) algorithm. This maximizes the likelihood of the profiles which is derived from their distribution given by:

$$p(\mathbf{u}_i) = \sum_{k=1}^{K} \pi_k \frac{\Gamma\left(\sum_{d=1}^{D} u_{id} + 1\right)}{\prod_{d=1}^{D} \Gamma(u_{id} + 1)} \prod_{d=1}^{D} \beta_{kd}^{u_{id}} \propto \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \beta_{kd}^{u_{id}}.$$

During the E phase, the probability of belonging to each cluster is calculated for each passenger:

$$p(z_k = 1|\mathbf{u}_i) \propto \frac{\prod_{d=1}^{D} \beta_{kd}^{u_{id}} \pi_k}{\sum_{k'=1}^{K} \prod_{d=1}^{D} \beta_{k'd}^{u_{id}} \pi_{k'}}.$$

During the M phase of the algorithm, the model's parameters are updated using the results from the E phase. Cluster proportions are updated as follows

$$\pi_k = \frac{1}{M} \sum_{i=1}^{M} p(z_k = 1|\mathbf{u}_i)$$

whereas word proportions are updated using the formula

$$\beta_{kd} = \frac{\sum_{i=1}^{M} p(z_k = 1|\mathbf{u}_i) u_{id}}{\sum_{d'=1}^{D} \sum_{i=1}^{M} p(z_k = 1|\mathbf{u}_i) u_{id'}}.$$

The number of passenger clusters K is set in the same fashion described in Section IV-A: we run an EM algorithm to estimate the mixture of unigrams models while varying K from 2 to 25 and select the most appropriate value of K using the slope heuristic.

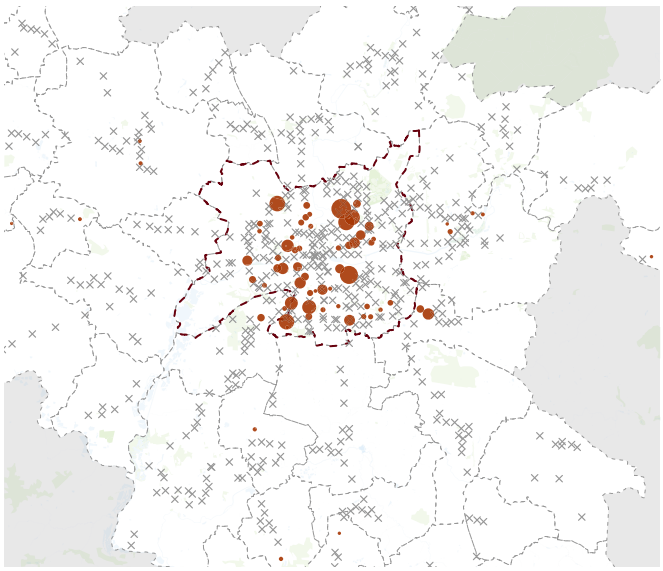

Fig. 10. Map of station cluster 5. (Dots) Stations in the cluster (each dot's area is proportional to its station $\mathbf{s}$'s scaling factor $\alpha_{\mathbf{s}}$). The vast majority of stations of the cluster are located in the city center itself, which explains why an important activity is observed all day long. (Gray Cross) Locations of stations that are not included in the cluster.

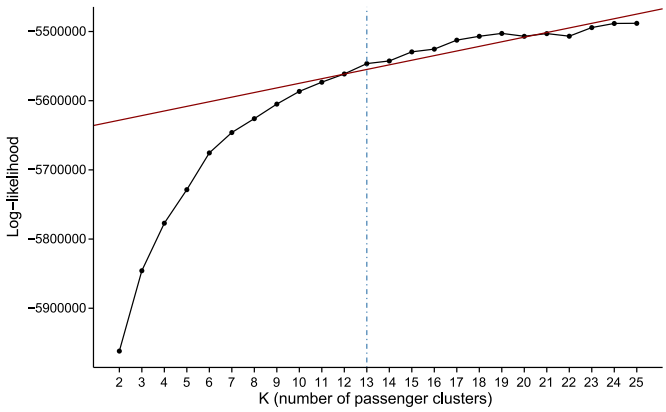

Fig. 11. Evolution of the log likelihood as a function of the number of passenger clusters $K = 2, \ldots, 25$. (Red Line) Linear model fitted to the linear part of the curve. (Blue Vertical Line) Suitable number of clusters $K = 13$.

C. Results

We now proceed to discussing the results of applying our clustering approach to the one-month dataset. Based on the slope heuristic, 13 passenger clusters are discovered (cf. Fig. 11). We study these clusters based on how their journey time (i.e., day and hour of the first boardings of their journeys) probabilities are distributed as well as on their fare type proportions. Originally, the public transport operator has an extensive grid with more than 90 fare types based mostly on pricing and operational considerations. We aggregate these fare types into seven categories: (i) Young subscribers (passengers aged 26 or less), (ii) Regular subscribers (passengers having a smart card with regular pricing and mainly aged 27 to 64), (iii) Elderly subscribers (aged 65 or more), (iv) Free travel (granted to citizens based on social considerations such as unemployment and income), (v) Short duration pass (unlimited travel during a short period ranging from one day to one week), (vi) Pay as you go (passengers paying per journey), and (vii) KR agents (the public transportation operator's agents).

Fig. 12 regroups passenger clusters in which no particular routine patterns are detected. Instead, the clusters are rather characterized by a diffuse usage of public transportation that appears at different times of the day. For example, in cluster 3 [cf. Fig. 12(c)] the diffuse usage appears mostly during the evening period, whereas in cluster 4 [cf. Fig. 12(d)] it starts in the morning and spans until the end of the evening. Except from cluster 1, the other three clusters are majorly composed of free travel passengers. In fact, the free travel ratios in the diffuse usage clusters shown in Fig. 12 are the highest among all clusters and add up to 65% of the total number of free travel passengers. This suggests that passengers benefiting from free travel (mainly due to unemployment or their unstable financial situation) do not have tight daily schedules around which their trips revolve, hence the absence of clear temporal travel patterns. Additionally, elderly passengers also seem to fall under the category of diffuse usage since the four clusters regroup 90% of this fare type cardholders (with 51% in cluster 4 and 35% in cluster 3). In total, 41.45% of cardholders are considered to have a diffuse usage of public transportation.

Typical home-work commute behavior is clearly visible in the clusters shown in Fig. 13 with a first peak in the morning and a second, more diffuse peak appearing during the afternoon for weekdays. Subtle differences exist between the two clusters. For instance, both peaks are slightly shifted in cluster 6 [cf. Fig. 13(b)] compared to those in cluster 5 [cf. Fig. 13(a)]. Additionally, the commuting pattern appears also on Saturday in the case of cluster 6, contrary to cluster 5. Both clusters are mainly composed of regular subscribers and contain almost 8% of the passenger population.

Similar commute patterns are also apparent in the clusters in Fig. 14 with the exception that the second peak on Wednesdays is shifted and occurs midday (around 12 pm) rather than in the evening. As mentioned earlier in Section III-C, this behavior is mostly related to students (especially in middle and high school) since course hours on Wednesdays end midday in France. Expectedly, three out of the four clusters (clusters 7, 8, and 9) are mainly composed of young subscribers. Exceptionally, cluster 10 [cf. Fig. 14(d)] mostly contains regular subscribers (adults). This suggests that these passengers are accompanying parents that align to their children's schedule. Notice that retrieving the clusters shown in Fig. 14 and separating them from those in Fig. 13 would be hard to achieve using clustering approaches that aggregate weekdays into a single daily profile since the subtle change in behavior on Wednesdays would be engulfed by the standard behavior occurring in the other weekdays. This is one of the reasons why we opt for a weekly temporal profile to describe passengers instead.

Besides from the two morning and evening peaks, a third peak appears in the two clusters shown in Fig. 15 (almost 17% of the passengers) which suggests that passengers in these groups also rely regularly on public transportation during their lunch breaks. Finally, the last cluster (see Fig. 16) contains "early-bird" passengers for which a single peak occurs very early in the morning (6 am) and usage in the afternoon is diffuse. We observe that for all passenger clusters in which travel

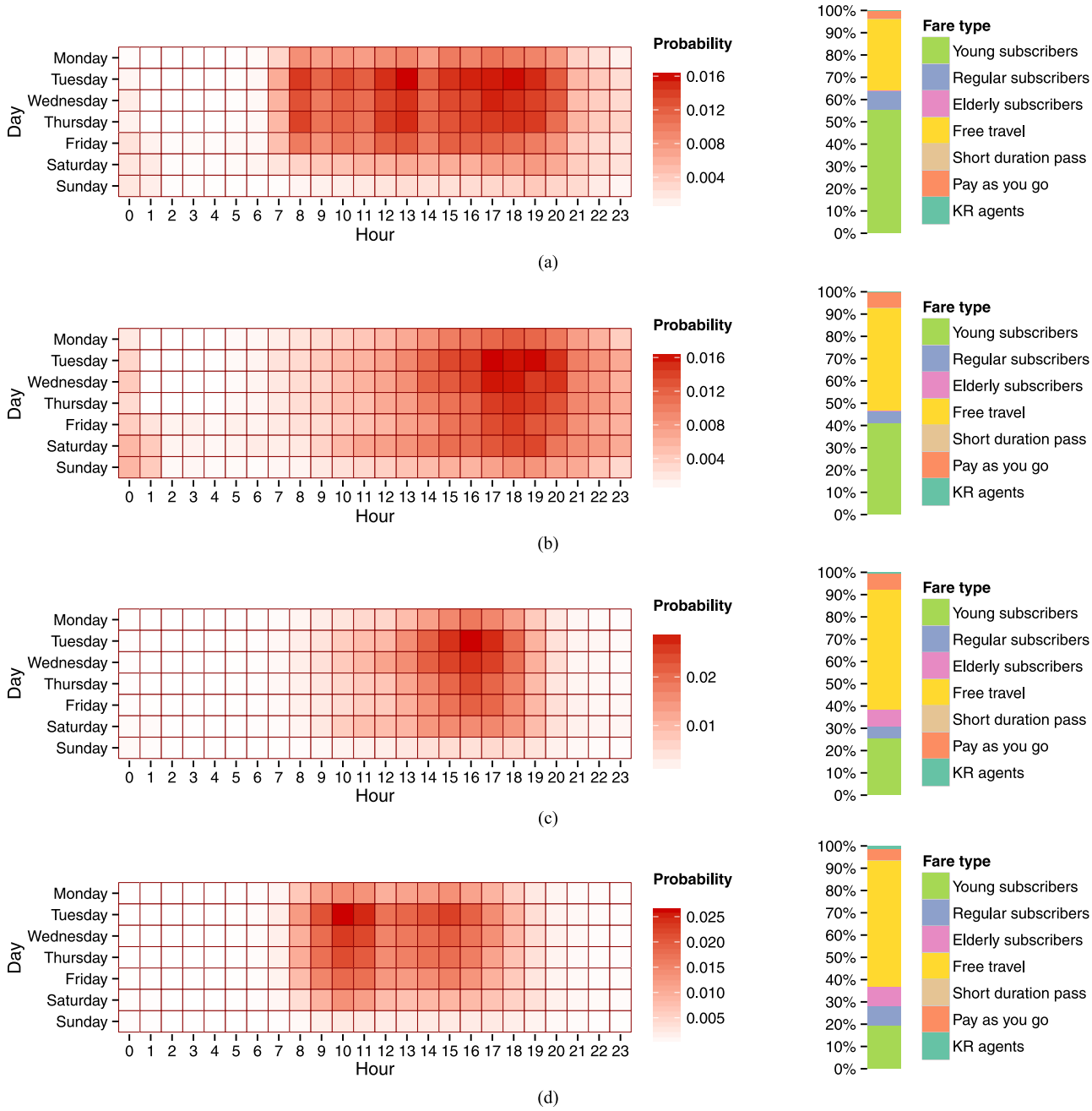

Fig. 12. Passenger clusters with diffuse public transportation usage appearing in various times during both weekdays and weekend. The heatmap on the left shows how journey time probabilities are distributed across the week, whereas the cluster's composition with respect to fare types is shown on the right. Color coding of the journey time probabilities is set locally for each cluster in order to make patterns more apparent. (a) Cluster 1: 11572 passengers (15.13%). (b) Cluster 2: 4924 passengers (6.44%). (c) Cluster 3: 6606 passengers (8.64%). (d) Cluster 4: 8600 passengers (11.25%).

patterns exist, the morning behavior is more regular and consistent than in the evening. This can be explained by the fact that in these clusters that mostly contain active adults and students, passengers have strict obligations with respect to the starting hours of their work or classes, whereas their leaving times are more flexible.

VI. Conclusions and Future Work

Smart card data present a unique opportunity to study passenger travel behavior in public transportation systems. In this paper, we started by applying a model-based clustering approach to transaction count statistics of stations. The retrieved clusters make it possible to distinguish between the different usage types of the stations. While usage is balanced during the day for many stations, the activity in others centers only on specific parts of the day. Housing stations located in remote residential parts of the city are mainly used in the morning during weekdays by passengers who commute to work, whereas the majority of the activity in work stations located in activity zones takes place in the evening when they are used by passengers to commute back home.

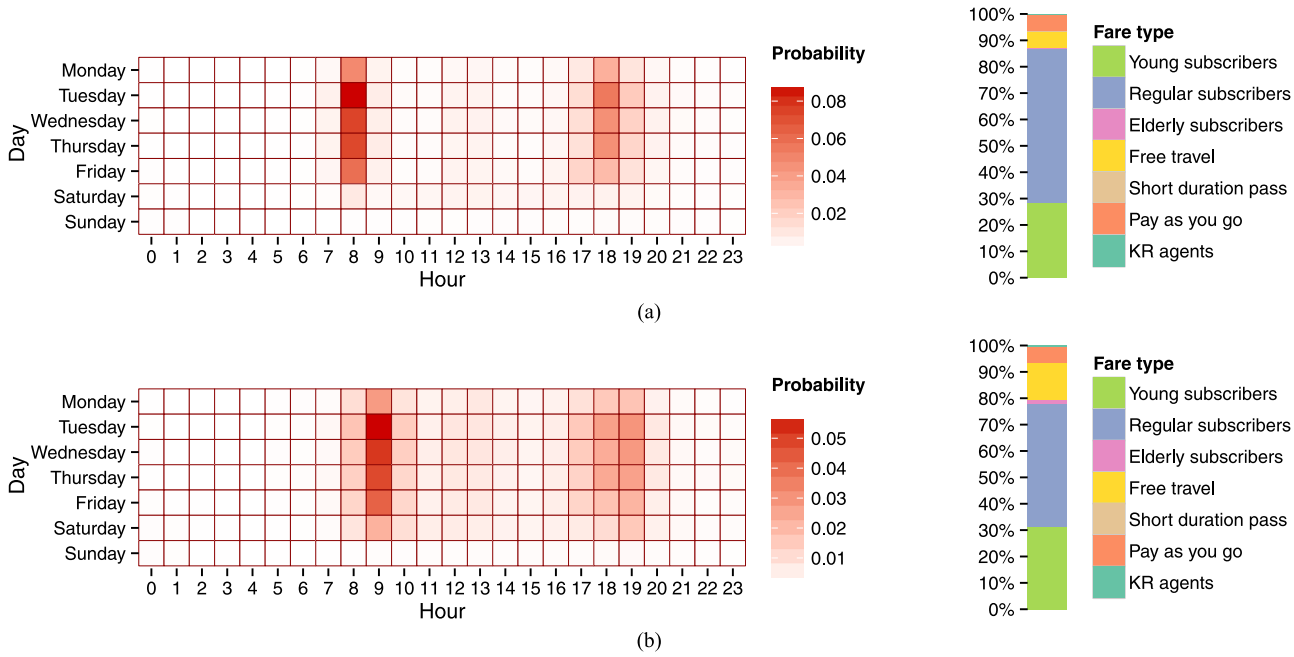

Fig. 13. Passenger clusters exhibiting typical commute behavior with the first peak occurring in the morning and the second one in the evening. (a) Cluster 5: 3479 passengers (4.55%). (b) Cluster 6: 2585 passengers (3.38%).

We also introduced an approach to clustering passengers based on their temporal habits. By estimating a mixture of unigrams model from journeys captured through smart card data, we retrieve weekly profiles (or temporal clusters) depicting different public transportation demands. Inspecting these profiles shows that other behaviors exist outside from the classically presumed home-work commute pattern. Some passengers use public transportation sporadically in a diffuse fashion, whereas the routines of other passengers revolve around a single time of the day (e.g., the morning period). Even in the presence of commute behavior, some differences exist between passengers. For instance, commute behavior of young subscribers (mainly students) differs from that of regular subscribers (mainly working adults). Additionally, commute routines also appear at various slightly-shifted times of the day for different passenger groups exhibiting such patterns.

We believe that knowledge extracted through station and passenger clustering can be very useful for public transportation planning as they can help both operators and authorities adapt the existing offer and propose adapted tools and services tailored to their customers' needs. For example, based on the extracted stations' roles, the operator can decide to increase the number of busses and subways serving housing stations in the morning and those serving work stations in the evening (which are the main peak hours of these station types). The passenger temporal patterns can help avoid the pitfall of naively believing that all passengers are commuters and can be used to scale the offer appropriately to adapt the most to the various demands and expectations of the different customer groups.

As such, the two approaches we introduced in this work constitute a first step towards developing decision support tools destined towards the different stakeholders involved in defining the city's public transportation offer and strategy (transport authorities, transport operators, local and regional authorities, etc.). The two approaches we designed to offer a two-layered "passenger-network" view of the transportation system and are to be considered as complementary. They can be applied to smart card data individually or conjointly in no specific order (e.g., start with passenger clustering then conduct station clustering or vice versa). One of our future research directions is the inclusion of socio-economical variables describing the passengers (age, income, sex, etc.) and the territory (e.g., employment indicators, presence of touristic attractions, etc.) in order to be able to link both the passenger and network layers.

Further work can be conducted based on the work presented herein. In the station clustering approach we used, the only distinction made between days is based on them being weekdays or weekend days. This can be extended in order to account for special days (e.g., holidays) or even to consider each day of the week (Monday, Tuesday, etc.) separately.

When considering passenger travel patterns, we chose to cluster the passengers based solely on the boarding time of the journeys they made. In order to do so, we had to reconstruct trip chains using a threshold-based destination and transfer inference approach. While we remain confident about the threshold values we used, it would be wise to further investigate how the produced trip chains compare to reality (e.g., by comparing them to household survey results). It would be interesting to include the spatial dimension (i.e., boarding locations, inferred alighting locations) either during the clustering process or during the interpretation step (e.g., to identify the stations that are impacted by a given type of demand).

The number of clusters discovered using our approach can be overwhelming to analyze. This issue can be addressed by trying to regroup similar clusters and aggregating them into a hierarchy that is more suitable for multi-level exploration

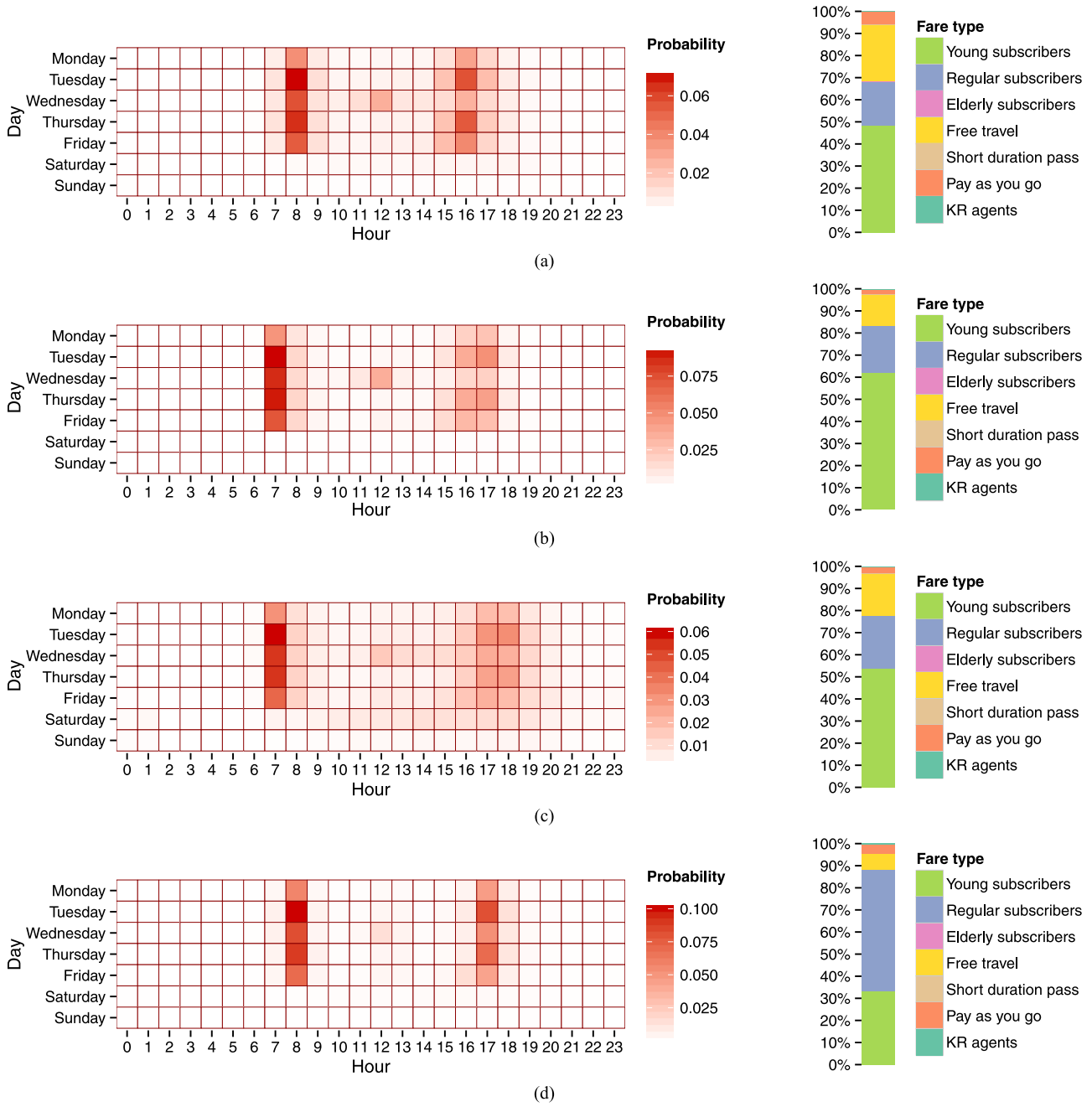

Fig. 14. Passenger clusters exhibiting commute behavior with a shifted second peak on Wednesdays. (a) Cluster 7: 3195 passengers (4.18%). (b) Cluster 8: 9587 passengers (12.54%). (c) Cluster 9: 6237 passengers (8.16%). (d) Cluster 10: 4018 passengers (5.25%).

(i.e. start with a small number of coarse clusters to quickly understand the macro structure of passenger behavior, then expand interesting clusters to reveal more refined patterns).

Another limitation of the presented models is that they make the assumption that behavior (making trips for passengers, trip counts for stations) is independent between different time bins of the same day. It would be interesting to address this shortcoming either, for example, by considering extensions that integrate additional parameters that capture these correlations (which will result in more complex models) or by using functional approaches in which the data are treated as functions (in the case of station clustering).

Also, we focused only on smart card data involving trips made by bus and subway. It would be interesting to study how these modes compare to and complement other transportation modes such as Bike Sharing Systems (BSS), etc.

Finally, one important aspect of travel behavior characterization that we didn't tackle in this work is that of identifying trip purposes. It is clear that some of the passenger clusters we discover pertain to particular trip purposes (e.g., school trip, home-work commutes, etc.) and can potentially be used to this end (e.g., by cross-comparing the clusters with the territorial characteristics of the locations they visit). Recently, a set of approaches ranging from identification of home and work

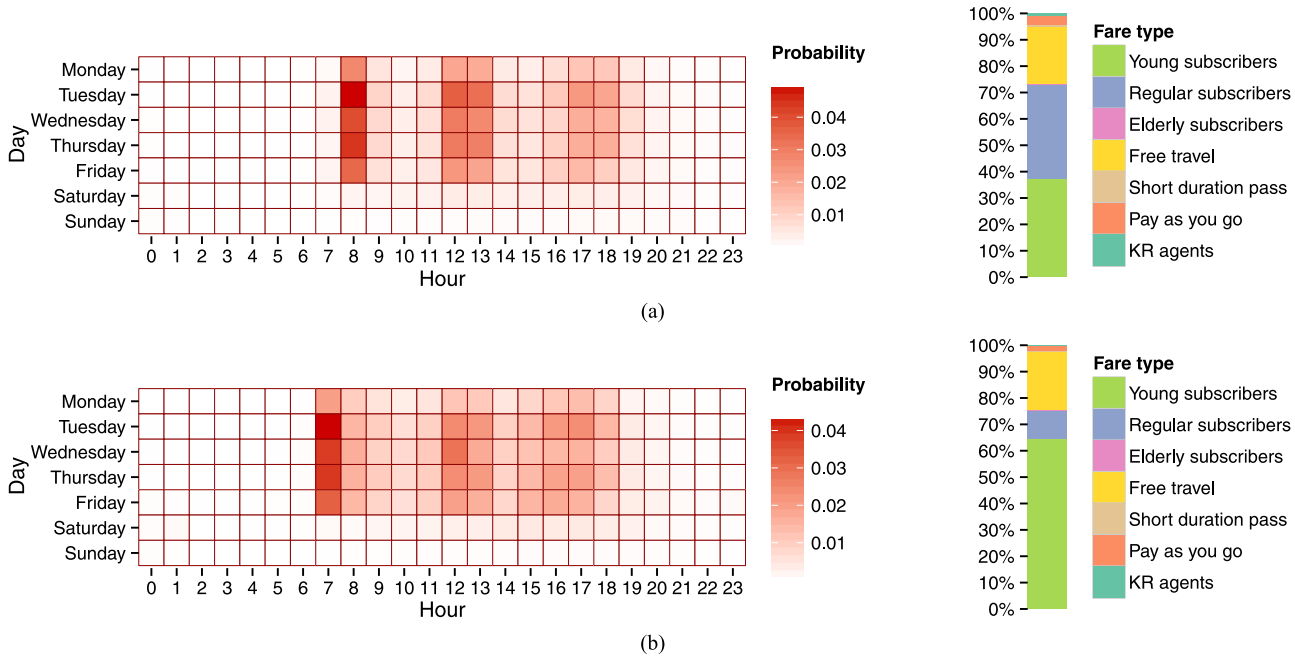

Fig. 15. Passenger clusters with three usage peaks occurring in the morning, midday, and evening, during weekdays. (a) Cluster 11: 4437 passengers (5.8%). (b) Cluster 12: 8486 passengers (11.1%).

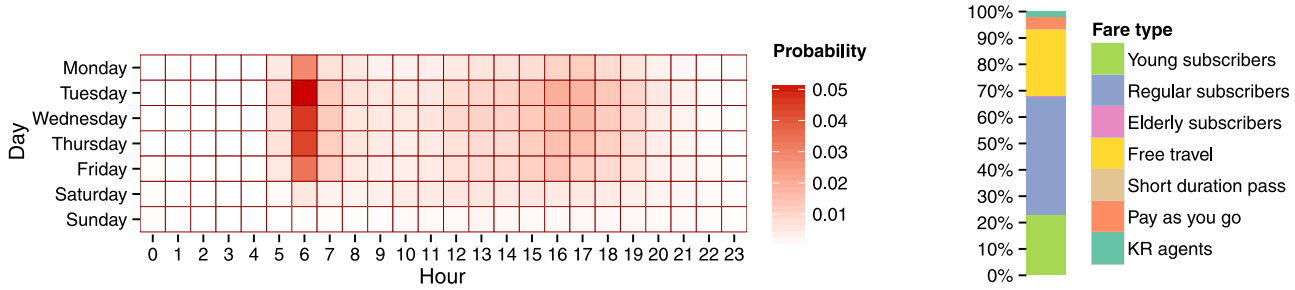

Fig. 16. Passenger cluster 13 regroups 2752 (3.6% of the total passengers) that exhibit a single peak occurring very early in the morning (6 am) and diffuse usage during the evening.

locations [50] to a more elaborate range of activities [25], [26] were proposed in the literature. In future work, we intend to extend our approach in order to include these aspects.

Acknowledgment

This research is undertaken as part of the Mobilletic Project. The authors extend their gratitude to Keolis Rennes who generously provided the data used in this study and especially to Mr. Sebastien Leparoux for answering our numerous questions and directing us to additional data related to the city.

The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions that greatly helped improve the quality of this paper.

References

[1] Q.-J. Kong, Z. Li, Y. Chen, and Y. Liu, "An approach to urban traffic state estimation by fusing multisource information," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 499–511, Sep. 2009.

[2] L. Chen *et al.*, "Container port performance measurement and comparison leveraging ship GPS traces and maritime open data," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 5, pp. 1227–1242, May 2015.

[3] P. A. Laharotte, R. Billot, E. Come, L. Oukhellou, A. Nantes, and N. E. E. Faouzi, "Spatiotemporal analysis of Bluetooth data: Application to a large urban network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1439–1448, Jun. 2015.

[4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[5] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerging Technol.*, vol. 19, no. 4, pp. 557–568, 2011.

[6] M. G. McNally, *The Four-Step Model*. Bingley, U.K.: Emerald, ch. 3, pp. 35–53. [Online]. Available: http://www.emeraldinsight.com/doi/abs/10.1108/9780857245670-003

[7] M. G. McNally and C. R. Rindt, *The Activity-Based Approach*. Bingley, U.K.: Emerald, 2007, ch. 4, pp. 55–73. [Online]. Available: http://www.emeraldinsight.com/doi/abs/10.1108/9780857245670-004

[8] M. Bagchi and P. R. White, "What role for smart-card data from bus systems?" in *Proc. ICE—Municipal Eng.*, 2004, vol. 157, pp. 39–46.

[9] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[10] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1971, no. 1, pp. 119–126, Jan. 2006.

[11] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 22, no. 5, pp. 376–387, 2007.

[12] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *J. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 1–14, 2007.

[13] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile," *Transp. Res. C, Emerging Technol.*, vol. 24, pp. 9–18, 2012.
[14] J. Gordon, H. Koutsopoulos, N. Wilson, and J. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2343, no. 1, pp. 17–24, 2013.
[15] M. Hofmann and M. O'Mahony, "Transfer journey identification and analyses from electronic fare collection data," in *Proc. IEEE Intell. Transp. Syst.*, Sep. 2005, pp. 34–39.
[16] L. M. Kieu, A. Bhaskar, and E. Chung, "Mining temporal and spatial travel regularities for transit planning," in *Australasian Transport Research Forum 2013*. Brisbane, QLD, USA: Queensland Univ. Technol., Oct. 2013.
[17] C. Seaborn, J. Attanucci, and N. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2121, no. 1, pp. 55–62, Dec. 2009.
[18] X.-L. Ma, Y.-J. Wu, Y.-H. Wang, F. Chen, and J.-F. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. C, Emerging Technol.*, vol. 36, pp. 1–12, 2013.
[19] K. K. Alfred Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transp. Res. Rec.*, vol. 2063, pp. 63–72, 2008.
[20] E. Nasiboglu, U. Kuvvetli, M. Ozkilcik, and U. Eliiyi, "Origin-destination matrix generation using smart card data: Case study for Izmir," in *Proc. 4th Int. Conf. PCI*, Sep. 2012, pp. 1–4.
[21] K. Chu and R. Chapleau, "Augmenting transit trip characterization and travel behavior comprehension," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2183, no. 1, pp. 29–40, Dec. 2010.
[22] A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, "Use of smart card fare data to estimate public transport origin–destination matrix," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2535, pp. 88–96, 2015. [Online]. Available: http://dx.doi.org/10.3141/2535-10
[23] L. He, N. Nassir, M. Trépanier, and M. Hickman, "Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems," Centre Interuniversitaire de Recherche sur les Reseaux d'Entreprise, la Logistique et le Transport (CIRRELT), Montreal, QC, Canada, Tech. Rep., 2015.
[24] F. Devillaine, M. Munizaga, and M. Trépanier, "Detection of activities of public transport users by analyzing smart card data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2276, no. 1, pp. 48–55, Dec. 2012.
[25] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transp. Res. C, Emerging Technol.*, vol. 46, pp. 179–191, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X14001612
[26] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transp. Res. B, Methodol.*, vol. 83, pp. 121–135, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191261515002593
[27] C. Morency, M. Trépanier, and B. Agard, "Analysing the variability of transit users behaviour with smart card data," in *Proc. IEEE ITSC*, 2006, pp. 44–49.
[28] T. Fuse, K. Makimura, and T. Nakamura, "Observation of travel behavior by IC card data and application to transportation planning," in *Proc. Special Joint Symp. ISPRS Commiss. IV AutoCarto*, 2010.
[29] N. Lathia and L. Capra, "How smart is your smartcard? measuring travel behaviours, perceptions, and incentives," in *Proc. 13th Int. Conf. UbiComp*, New York, NY, USA, 2011, pp. 291–300.
[30] W. Tran, "Analysis of the differences in travel behaviour between pay as you go and season ticket holders using smart card data," in *Proc. 1st Civil Environ. Eng. Student Conf.*, 2012, pp. 1–6.
[31] N. Lathia, D. Quercia, and J. Crowcroft, "The hidden image of the city: Sensing community well-being from urban mobility," in *Proc. 10th Int. Conf. Pervasive*, Berlin, Germany, 2012, pp. 91–98.
[32] M. Trépanier, K. M. Habib, and C. Morency, "Are transit users loyal? Revelations from a hazard model based on smart card data," *Can. J. Civil Eng.*, vol. 39, no. 6, pp. 610–618, 2012.
[33] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996. pp. 226–231.
[34] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015.
[35] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *Proc. 12th IFAC Symp. INCOM*, 2006, pp. 1–14.
[36] N. Lathia, C. Smith, J. Froehlich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," *Pervasive Mobile Comput.*, vol. 9, no. 5, pp. 643–664, 2013.
[37] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: Understanding tube station congestion patterns from trip data," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 134–141.
[38] M. Poussevin, N. Baskiotis, V. Guigue, and P. Gallinari, "Mining ticketing logs for usage characterization with nonnegative matrix factorization," in *Proc. SenseML—ECML Workshop*, Nancy, France, Sep. 2014, pp. 147–164.
[39] G. G. Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. C, Emerging Technol.*, vol. 64, pp. 1–16, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X15004283
[40] S. Hasan, C. Schneider, S. Ukkusuri, and M. González, "Spatiotemporal patterns of urban human mobility," *J. Statist. Phys.*, vol. 151, no. 1/2, pp. 304–318, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10955-012-0645-0
[41] F. Zhang, N. Yuan, Y. Wang, and X. Xie, "Reconstructing individual mobility from smart card transactions: A collaborative space alignment approach," *Knowledge Inf. Syst.*, vol. 44, no. 2, pp. 299–323, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10115-014-0763-x
[42] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatialtemporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *J. Transp. Geogr.*, vol. 41, pp. 21–36, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966692314001689
[43] S. Foell, S. Phithakkitnukoon, G. Kortuem, M. Veloso, and C. Bento, "Predictability of public transport usage: A study of bus rides in Lisbon, Portugal," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2955–2960, Oct. 2015.
[44] S. Zhong and J. Ghosh, "A unified framework for model-based clustering," *J. Mach. Learn. Res.*, vol. 4, pp. 1001–1037, Dec. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=945365.964287
[45] E. Côme and L. Oukhellou, "Model-based count series clustering for bike sharing system usage mining: A case study with the Vélib System of Paris," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 39:1–39:21, Jul. 2014.
[46] C. Zhong, M. Batty, E. Manley, J. Wang, Z. Wang, F. Chen, and G. Schmitt, "Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data," *PLoS ONE*, vol. 11, no. 2, pp. 1–17, 2016.
[47] L. Birgé and P. Massart, "Minimal penalties for Gaussian model selection," *Probability Theory and Related Fields*, vol. 138, no. 1/2, pp. 33–73, 2007.
[48] J.-P. Baudry, C. Maugis, and B. Michel, "Slope heuristics: Overview and implementation," *Statist. Comput.*, vol. 22, no. 2, pp. 455–470, 2012.
[49] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, no. 2/3, pp. 103–134, May 2000.
[50] G. Li, L. Yu, W. S. Ng, W. Wu, and S. T. Goh, "Predicting home and work locations using public transport smart card data by spectral analysis," in *Proc. IEEE 18th ITSC*, Sep. 2015, pp. 2788–2793.

Mohamed K. El Mahrsi received Bachelor's degree in computer engineering from National School of Computer Science, Manouba, Tunisia, in 2008 and the Ph.D. degree in computer science from Télécom ParisTech, Paris, France, in 2013.

He is a Postdoctoral Researcher with the Engineering of Surface Transport Networks and Advanced Computing (GRETTIA) Laboratory, French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR), Marne-la-Vallée, France. His research interests include machine learning in general and exploratory data analysis and unsupervised learning techniques with applications to digital traces in particular.

Etienne Côme received the Master's and Ph.D. degrees from Université de Technologie de Compiègne, Compiègne, France, in 2005 and 2009, respectively.

He is a Researcher with the Engineering of Surface Transportation Networks and Advanced Computing (GRETTIA) Laboratory, French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR), Marne-la-Vallée, Paris. His research interests include random graph models, mixture models, and more generally probabilistic graphical models and their use to solve transportation problems.

Latifa Oukhellou received the Ph.D. degree in automatic and signal processing from Université Paris-Sud, Paris, France, in 1997 and the "Habilitation à diriger des Recherches" from Université Paris-Est, Marne-la-Vallée, France, in 2010 .

She is currently a Senior Researcher with the French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR), Marne-la-Vallée, where she is the Head of the Data and Mobility Group, Engineering of Surface Transport Networks and Advanced Computing (GRETTIA) Laboratory. She has been an Assistant Professor with Université Paris-Est Créteil, Paris. She has published several papers in international scientific journals and conference proceedings and she is involved in many research projects. Her research interests include pattern recognition, machine learning and information fusion applied to diagnosis problems, and sensor networks, as well as spatio-temporal data analysis or supporting driving behavior and mobility.

Michel Verleysen received the M.S. and Ph.D. degrees in electrical engineering from Université Catholique de Louvain, Louvain, Belgium, in 1987 and 1992, respectively.

He was an Invited Professor with Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 1992; with Université d'Evry Val d'Essonne, Evry, France, in 2001; and with Université Paris 1 Panthéon-Sorbonne, Paris, France, from 2002 to 2011. He is currently a Full Professor with the Machine Learning Group, Institute of Information and Communication Technologies, Electronic and Applied Mathematics, Université Catholique de Louvain, and an Honorary Research Director with the Belgian National Fund of Scientific Research. He is the author or coauthor of more than 250 scientific papers in international journals and books or communications to conferences with reviewing committees. His research interests include machine learning, feature selection, artificial neural networks, self organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.