Transportation Research Part C 79 (2017) 274–289

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Analyzing year-to-year changes in public transport passenger behaviour using smart card data ☆

Anne-Sarah Briand *, Etienne Côme, Martin Trépanier, Latifa Oukhellou

Université Paris-Est, IFSTTAR, COSYS-GRETTIA, 14-20 Boulevard Newton, F-77447 Marne-la-Vallée, France
CIRRELT/Polytechnique Montréal, Department of Mathematics and Industrial Engineering, P.O. Box 6079, Station Centre-Ville, Montréal, Québec H3C 3A7, Canada

ARTICLE INFO

Article history:
Received 14 March 2016
Received in revised form 27 January 2017
Accepted 28 March 2017
Available online 6 April 2017

Keywords:
Smart card
Passenger clustering
Mixture model
Public transit
Longitudinal analysis

ABSTRACT

In recent years, there has been increased interest in using completely anonymized data from smart card collection systems to better understand the behavioural habits of public transport passengers. Such an understanding can benefit urban transport planners as well as urban modelling by providing simulation models with realistic mobility patterns of transit networks. In particular, the study of temporal activities has elicited substantial interest. In this regard, a number of methods have been developed in the literature for this type of analysis, most using clustering approaches. This paper presents a two-level generative model that applies the Gaussian mixture model to regroup passengers based on their temporal habits in their public transportation usage. The strength of the proposed methodology is that it can model a continuous representation of time instead of having to employ discrete time bins. For each cluster, the approach provides typical temporal patterns that enable easy interpretation. The experiments are performed on five years of data collected by the Société de transport de l'Outaouais. The results demonstrate the efficiency of the proposed approach in identifying a reduced set of passenger clusters linked to their fare types. A five-year longitudinal analysis also shows the relative stability of public transport usage.

1. Introduction

The notion of mobility was developed near the end of the 1980s and gradually introduced individual practices and lifestyles into the analysis of transport demand, which made it necessary to redefine the meaning of the term. The aim of technically optimizing the straightforward spatial movement of goods and individuals (i.e., planning, flow, traffic, vehicle technology, etc.) has been supplemented, or even replaced, by the objective of obtaining a detailed understanding of the variation in individuals' ability to travel (accessibility), individuals' experience of daily travel conditions (comfort, sustainability), and/or even the role that mobility plays in individuals' lifestyles in terms of both actual and possible interactions. This paradigm shift should be understood as recasting the focus to concentrate more on individuals and less on vehicles and technologies. Consequently, mobility is now studied by economists, sociologists, urban planners, geographers and data scientists.

Traditionally, the analysis of mobility is based on travel surveys (household travel surveys (HTSs), origin-destination surveys, and cordon line surveys). These surveys have a number of advantages – for example, they cover all transport modes and trip purposes and generate meta-data regarding the respondents (gender, socio-occupational group, etc.). However, they are also expensive and consequently are undertaken fairly infrequently (typically, every five or ten years), which means that current developments and the public policies that aim to influence them are not closely monitored. In addition, HTSs are

☆ This article belongs to the Virtual Special Issue on "Smart cards, big data and travel behaviour".

* Corresponding author at: Université Paris-Est, IFSTTAR, COSYS-GRETTIA, 14-20 Boulevard Newton, F-77447 Marne-la-Vallée, France.

http://dx.doi.org/10.1016/j.trc.2017.03.021
0968-090X/

typically subject to statistical biases (due to sample size, representativeness, and non-responses), some of which can be partly compensated for by sampling theory techniques (such as correction and imputation). Another type of bias that is more difficult to quantify has been identified recently by Lathia and Capra (2011), who have shown that there are differences between the trips listed by participants in survey responses (when providing information about all their trips on the previous day) and the trips they actually made (as objectively shown by ticketing data).

Analyses of daily travel can rely on the growing number of digital records that are generated during trips. In many transportation networks, boarding logs are registered for all cards, which generates large-scale data. Anonymized ticketing data can indeed be used to implement new modelling approaches for urban mobility analyses even if they were not initially designed for this purpose. These data have a number of valuable intrinsic advantages, i.e., they are exhaustive (i.e., the data are distributed along the network on all transport modes), they are spatially and temporally precise, and they lack the response bias that plague survey data. However, to protect personal privacy, these data provide no socio-economic data on the user. Likewise, the data provide no information about trip purposes or the reasons behind the choice of a given travel mode or route. The use of smart card data to analyse urban mobility thus raises a number of issues and challenges that researchers in this field are working to address.

A great number of studies have been undertaken to examine passenger behaviour on transit networks, and a detailed review of related studies on this topic is provided in the next section. Transportation users' habits are of great interest to transit network operators because knowing these habits can help to predict affluence and enhance our understanding of the demand for transportation. Mobility patterns may also be used in urban modelling to provide simulation models with realistic public transport passenger behaviours. Many methods have been developed that identify such mobility patterns, most of which are based on clustering approaches. However, very few aims to analyse the evolution of cluster partitioning over a span of years (e.g. longitudinal analysis has been done in Morency et al. (2007)). The aim of this paper is to analyse the variability of passenger behaviour over time – in addition to passenger clustering – based on smart card data that spans a five-year period. Indeed, a transit network with few passenger activity changes from year to year will require less attention in terms of operator adjustments than a network with many passenger activity changes.

This paper makes the following contributions:

- It proposes that passenger cards should be clustered using temporal activities and should seek to partition passengers into small sets of clusters based on their usage habits of the transportation network, i.e., passengers with the same habits in terms of the hours in which they use the transport network will belong to the same cluster. In addition, this paper will demonstrate the replicability of the methodology initially proposed by Briand et al. (2016). The strength of this approach centres on its ability to take a continuous representation of time into account instead of having to employ the time binning used in most of the previous literature.
- In addition to offering a simple interpretation of cluster patterns, the aim here is to exploit the potential of the clustering methodology to perform a longitudinal analysis and, in particular, to study the evolution of passenger behavioural changes using their membership in different clusters over time. Thus, this study details the experimental results obtained using a real dataset covering a period of five years.
- Spatial characterization is also performed on the clusters identified. Shannon entropy is used to perform this analysis notwithstanding that socio-economic data on passengers are not available, except for fare type.

The remainder of this paper is organized as follows. Section 2 reviews related studies on ticketing data analysis and clustering. The real datasets used in our study and preliminary statistics are then described in Section 3. The approach to clustering passengers based on their temporal behaviour is introduced in Section 4. The experimental results are then presented and discussed in Section 5. Finally, concluding remarks and a discussion of future work are presented in Section 6.

2. Related studies

In recent years, the increased number of digital footprints that are collected reflecting our daily mobility has led to unprecedented approaches to innovative mobility studies (Zheng et al., 2014). From the use of GPS (Pang et al., 2013) to bike sharing data (Ahillen et al., 2015), growing amounts of information are available. Our study takes place in the public transportation field and focuses on mobility behaviours using smart card data collected via fare card collection systems. Although the information potential of smart card data has been attested to in previous studies (Bagchi and White, 2004; Park et al., 2008; Utsunomiya et al., 2006), their incomplete nature (lack of alighting locations, socioeconomic data, etc.) continues to motivate a substantial amount of research (Pelletier et al., 2011). Moreover, myriad topics have been addressed with respect to mining smart card data. We choose to categorize them three main topics, i.e., data completion and enrichment, prediction and passenger behaviour analysis.

2.1. Data completion and enrichment

Alighting data are of great interest for purposes of network analysis. Notably, the first estimations on Origin-Destination (OD) matrices based on assumptions of passenger mobility were undertaken in (Barry et al., 2002). Other studies have also focused on this topic by expanding prior hypotheses used to estimate destinations (Trépanier et al., 2007; Wang et al., 2011)

or by proposing different methodologies based on a probabilistic approach (Zhang et al., 2014). Some validation methods for estimated OD matrices are developed in (Munizaga et al., 2014). In addition to alighting locations, completion procedures are undertaken with respect to other missing data, such as transfer detection (Bagchi and White, 2005; Chu and Chapleau, 2008; Nassir et al., 2015; Seaborn et al., 2009), reconstruction of load profiles (Hofmann and O'Mahony, 2005), passenger route deduction using entry and exit validation data (van der Hurk et al., 2015) or trip purpose inferences (Lee and Hickman, 2013). Most studies combine two or more of these approaches in their datasets (e.g., estimating alighting locations and detecting transfers (Munizaga and Palma, 2012; Zhao et al., 2007)).

2.2. Smart card use for prediction

The second topic that has been investigated is related to the use of smart card data to generate insights into passengers' travel practices and to identify or predict travel patterns (Ceapa et al., 2012; Foell et al., 2014; Fuse et al., 2010; Lathia et al., 2010; Seaborn et al., 2009). Using smart card data, Ceapa et al. (2012) focus their work on traffic congestion and establish that there is some spatial and temporal regularity in congestion that makes it easier to predict. These authors also show that congestion does not last. Lathia et al. (2010) study individual travel data on London subways (personalized service offering) to estimate personal travel and develop a prediction method of personalized travel hours for passengers that ranks stations based on future mobility patterns. The approach proposed by Foell et al. (2014) is applied to a bus network and also used to predict trips.

2.3. Analysis of passenger behaviour

Other studies aim to illuminate passenger behaviours on a transportation network. In particular, these studies seek to identify external factors that influence network use, such as weather conditions (Arana et al., 2014), spatio-temporal dynamics (Tao et al., 2014), and/or passengers' activity changes over time (Chu, 2015). Another issue of great interest to transport operators involves partitioning network passengers into groups based on their transportation network activity. Most studies on this topic aim to regroup the passengers using clustering approaches such as the Hierarchical Ascendant Classification (HAC) or k-means algorithm (Agard et al., 2006; Morency et al., 2006). More advanced data mining tools, such as DBSCAN, NMF (Nonnegativ Matrix Factorization) and a mixture of unigram models have also been applied to achieve the same goal (Ma et al., 2013; El Mahrsi et al., 2014; Poussevin et al., 2014).

The methodology adopted in this paper follows the line of research that the same authors initiated in (Briand et al., 2016). As in that study, we aim to identify groups of passengers with similar temporal usage of public transport using one month of their smart card data validations. Langlois et al. (2016) perform analyses of multi-week activity patterns using clustering, and these authors propose a representation of longitudinal activity sequences using temporal and spatial activity (area of validation). The groups created by the clustering are associated with distinct sequence structures, thus allowing for better knowledge of 4 weeks of passenger activity. In our paper, we also investigate the analysis of changes in public transport passenger behaviour by tracking passenger clustering assignments, but our analysis is driven by five years of available data. In the next section, we describe our data set.

3. Data description

3.1. Network

The data used in this paper were provided by the Société de Transport de l'Outaouais (STO) based in Gatineau, Canada. A medium-sized public transit authority, the STO operates 310 buses and services 291,000 inhabitants. The STO has been operating its smart card system since 2001. Today, more than 80% of all STO passengers use smart cards. An important feature of Gatineau's transportation network is that it is located near Ottawa (the capital of Canada) in Ontario. Many of Gatineau's bus lines consequently serve Ottawa, which hosts a more extensive population of 883,391 inhabitants (according to the 2011 census) and services a significant amount of activities, particularly work-related activities.

3.2. Datasets

Five different datasets were used in this study. The data were recorded between the 1st and 28th of February for the 2005–2009 period. Between 644,614 (2005) and 740,883 (2009) observations composed each of them for those years, which comes to a total of 3,492,310 validations made by 82,223 cards. These datasets were not filtered prior to the study, which means that they include all cards that engaged in at least one transaction during the study period.

Each observation consists of a ticketing log with an anonymized card ID, the card type, transaction date and time, stop location, transport line, method of validation and type of transaction (transfer or non-transfer). As this study is dedicated to temporal activity, the alighting locations do not have to be estimated. However, we can look at the validations' boarding locations to obtain a better idea of Gatineau City's organization and its interactions with nearby cities. The number of validations registered during February 2005 is presented in Figs. 1 and 2. Two types of validations were considered – morning validations before 12:00 PM and afternoon validations after 12:00 PM.

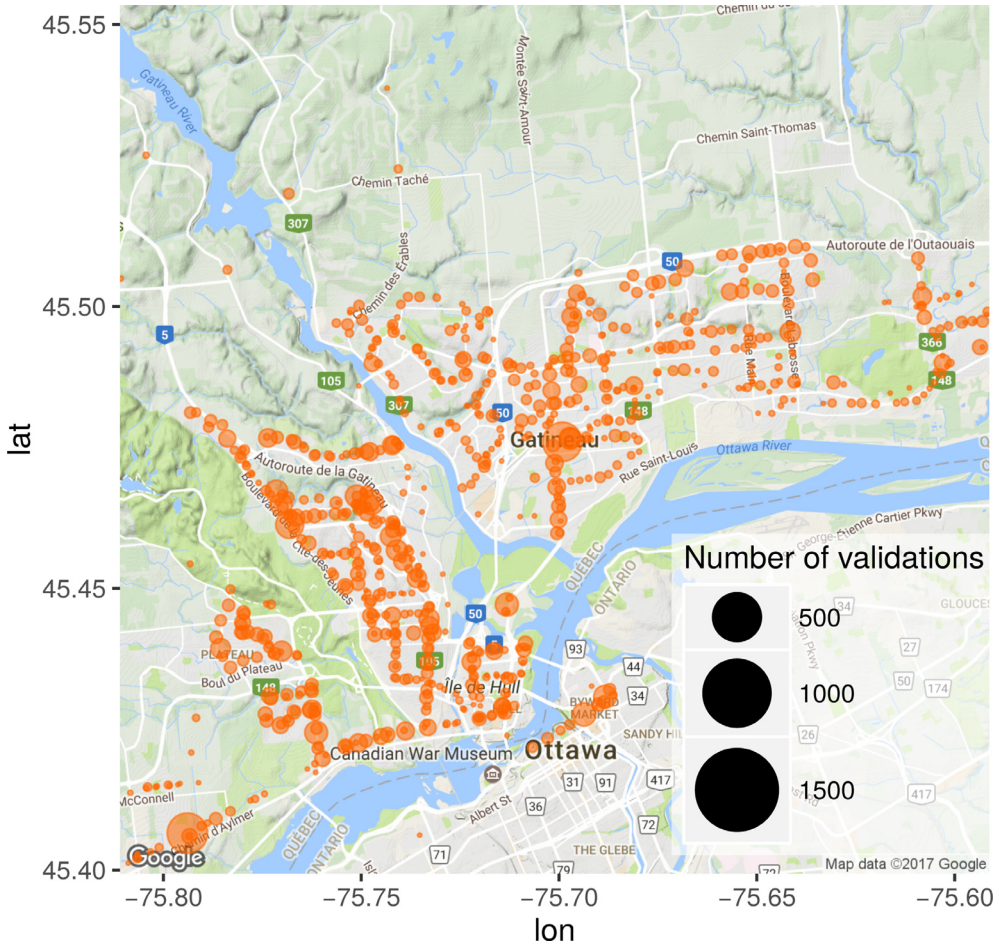


Fig. 1. Map representing the number of validations on the STO network during the month of February 2005, before midday.

The data reveal that there is a difference in the locations of these two types of validations. Afternoon validations mostly occurred in Ottawa, while morning validations mostly occurred in Gatineau, as most STO users live in Gatineau and work in Ottawa.

As discussed above, we focus our analysis on data collected over a five-year span. To obtain the same cluster parameters over the five years, these datasets have been grouped together to form one dataset. It is important to permit a card to change clusters from year to year. The card ID of a given card has been changed from one year to the next, thus allowing differentiation between cards based on the year. Separate analyses were also conducted for each dataset, but the results will not be described here.

The section addressing card tracking over the years indicates that only cards that are used continually from 2005 to 2009 are kept. Two datasets are thus considered: the first is called the all-card dataset and the second the reduced card dataset. The reduced card dataset consists of 391,783 validations made by 2,504 cards.

To better interpret the results, the card types are grouped into 11 card fare types: AP Express adult, AP Interzone adult, AP Regular adult, AP Senior, College/Univ., Express adult, Interzone adult, Other, Regular adult, Senior and Student. AP means that the cards are linked to a bank account for automated payment. Express users may use express routes, whereas the other users cannot. The same is true for those Interzone users who commute from the outer suburbs. Fig. 3 shows the proportion of each card type in the two datasets. The greatest difference in the composition of the two datasets is the absence of student cards for the reduced dataset. Students cannot keep the same card, they have to change it every year, which is not necessarily the case for other fare types. Thus, it will be inherently impossible to follow student activity in the second part of our study.

4. Passenger clustering methodology

One of the principal goals of transportation operators is to better understand their customers. One way to achieve this goal is to study the temporal habits of passengers on their networks. This type of analysis conducted on an individual scale is frequently too specific and does not offer an extensive overview of the activity on the network, which is why clustering

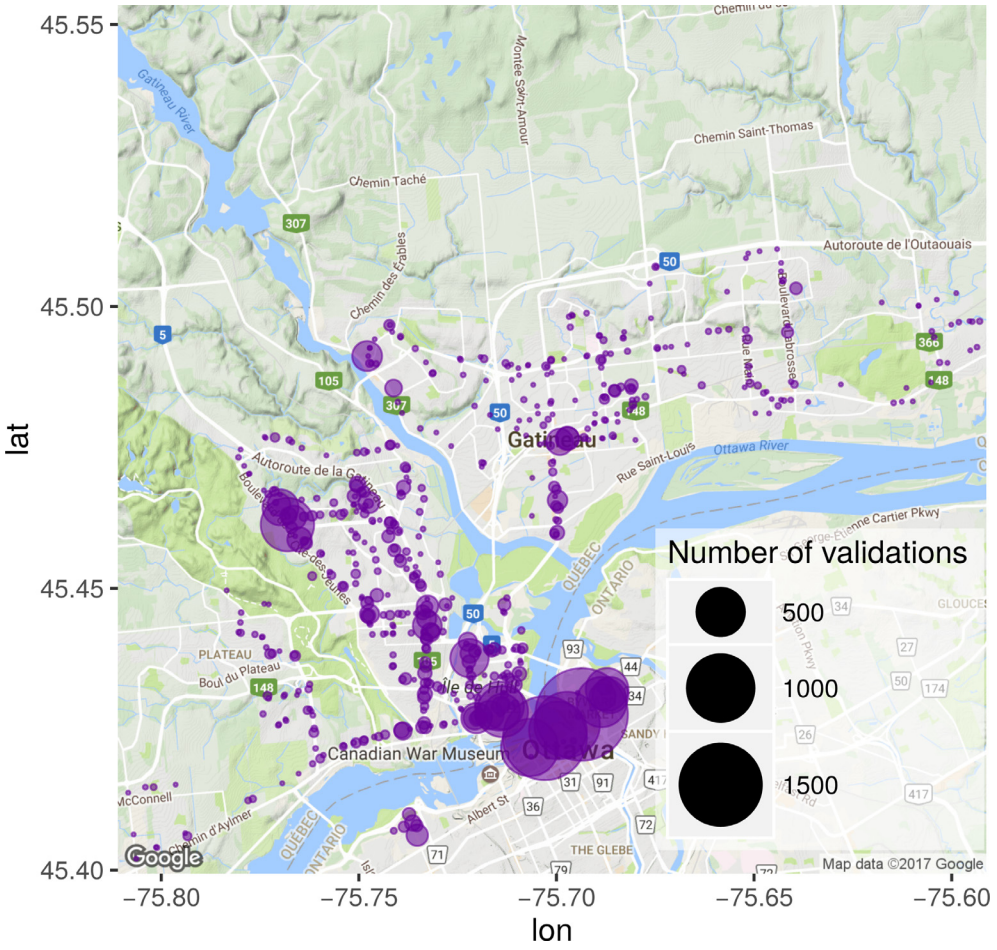


Fig. 2. Map representing the number of validations on the STO network during the month of February 2005, after midday.

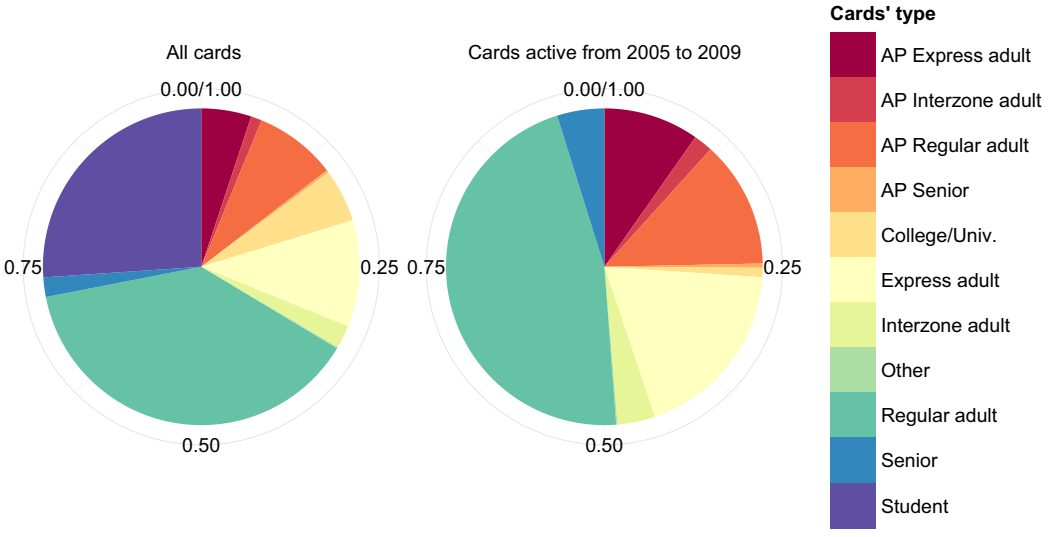


Fig. 3. Proportions of card types for the entire dataset and a reduced dataset that includes only cards used from 2005 to 2009.

methods are of great interest for use with such data. Indeed, they offer the opportunity to achieve an overview of passengers' habits by observing them through representative groups. Moreover, crossing clustering results with metadata allows us to characterize this cluster with additional specifications. In this section, an overview of the generative mixture model will be presented, as well as the algorithm used for clustering. As discussed above, in the following study, we will be working with data collected over five different years.

4.1. Gaussians mixture generative model

Profile clustering has been conducted on mobility data in recent years, and several methods have been developed to this end. The method that we propose in this study is a mixture model approach inspired by the work of El Mahrsi et al. (2014). The difference involves the use of a Gaussian mixture approach instead of a unigram mixture. In El Mahrsi et al. (2014), observations were aggregated into one-hour bins, whereas the mixture model uses non-aggregated data and fits a Gaussian mixture onto it. In other words, the model is a two-level generative model: the first-level model cards are partitioned into groups (card clusters), whereas the second takes all ticketing logs of the clusters' cards and uses them to represent temporal activity profiles of these groups as a Gaussian mixture model. The choice of Gaussian mixture to represent a passenger's cluster activity is a natural choice to describe the temporal habits of the passenger when the continuous nature of timestamps needs to be preserved. The Gaussian number in the mixture (H) will be explained in Section 4.3.

To model the cluster memberships of the cards, a latent variable with a multinomial distribution is introduced. Z^1 denotes membership of one of the K card's clusters. Similarly, the second level of the generative model is a Gaussian mixture model where membership of one of the H Gaussians is denoted by Z^2. Indeed, modelling the cards' temporal activity using a mixture of Gaussians is a natural choice in view of the continuity of data and also allows us to reveal the peaks of activity and characterize them using the variance and mean of the Gaussian model.

More formally, the model can be written as

$$
\begin{aligned}
&Z_i^1 \sim \mathcal{M}(1,\pi),\\
&Z_{ij}^2 | Z_{ik}^1 D_{ijl} = 1 \sim \mathcal{M}(1,\tau_{khl}),\\
&X_{ij} | Z_{ik}^1 Z_{ijh}^2 D_{ijl} = 1 \sim \mathcal{N}(\mu_{khl},\sigma_{khl}).
\end{aligned}
$$

Z_i^1 encodes the membership of card $i, i \in \{1,\dots,M\}$ onto one of the K cards' clusters and follows the multinomial distribution (denoted $\mathcal{M}$) of parameter $\pi = (\pi_1,\dots,\pi_K)$. Z_{ij}^2 encodes the membership of trip $j, j \in \{1,\dots,N_i\}$ (with N_i being the number of trips of cards i) to one of the H Gaussians, describing the temporal activity of cluster Z_{ik}^1 for the day D_{ijl} ($l \in \{1,\dots,7\}$ the set of the days of the week) and follows a multinomial distribution of parameter τ_{khl}. Finally, X_{ij} is the trip time, which is generated using the Gaussian distribution defined by $\mathcal{N}(\mu_{khl},\sigma_{khl})$. A graphical representation of the model is presented in Fig. 4.

The conditional density of X_{ij} can then be written as

$$
f(X_{ij} | \{ Z_{ik}^1 Z_{ijh}^2 D_{ijl} = 1 \}) = \sum_{h=1}^{H} \tau_{khd_{ij}} f(x; \mu_{khd_{ij}}, \sigma_{khd_{ij}}), \tag{1}
$$

where $f(.;\mu,\sigma^2)$ is the density function of Gaussian distribution of mean μ and variance σ. The Likelihood of the model is then given by

$$
L(\theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k \left(\prod_{j=1}^{N_i} \sum_{h=1}^{H} \tau_{khd_{ij}} f(x_{ij}; \mu_{khd_{ij}}, \sigma_{khd_{ij}}) \right)
$$

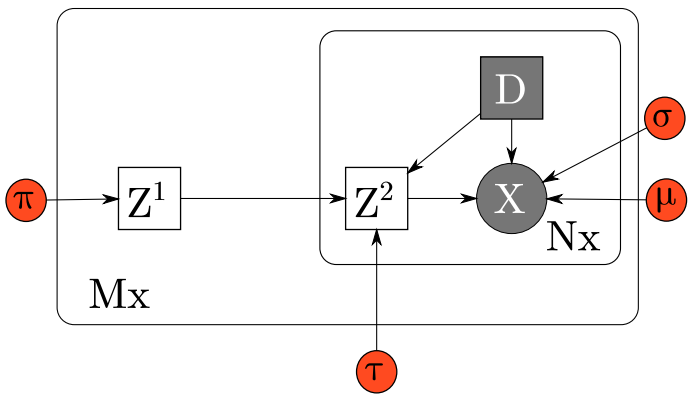


Fig. 4. Graphical model of the generative mixture model.

4.2. CEM and EM combined algorithmes

To estimate the likelihood parameters in the mixture models, the traditional method is to use an Expectation Maximization (EM) algorithm or a Classification Expectation Maximization (CEM) when including a classification step. The model presented in the final section consists of two levels, which recommends us to adopt a two maximization step process for the parameter estimation. The complete log-likelihood is used as a maximization criterion for the estimation. The idea is to work in a density estimation context for Z^2 and to work in a clustering context for Z^1. To this end, a CEM algorithm (McLachlan and Krishnan, 2008) is used because it includes a classification step that assigns each observation to its most probable cluster (rather than yielding a vector of membership probabilities, as in the classic EM). As our model is a two level generative model, it needs to be initialized in two steps. The first consists of randomly sampling the passengers into small groups. The parameters $\pi = (\pi_1, \ldots, \pi_K)$ of the multinomial distribution are then initialized using the proportions of these groups. The second step is then to use each of the small groups to initialize the remaining parameters. As is usual in mixture model initialization this step consists of using a k-means algorithm on each previously defined group. The k-means results are used in order to obtain a preliminary estimation of the remaining parameters of each gaussian in the mixture model, namely τ, μ and σ.

In summary, this algorithm takes the new user id (defined in Section 3.2) as entry data, the day (Monday,. . .,Sunday), and the hour of validation (using service hours, i.e., with no break at midnight). It returns the associated cluster for each user as well as the Gaussian mixture parameters and the complete log-likelihood.

4.3. Parameter choice

As shown above, the algorithm depends on a great number of parameters. Some are estimated using the algorithm itself (π, θ and τ) and some are chosen beforehand (H represents the number of Gaussians and K represents the number of clusters). In this section, we attempt to determine the best choice of K and H for the remainder of the study.

To identify the best parameters, we launched the algorithm for different values of H ($H = 2, \ldots, 5$) and for different values of K ($K = 2, \ldots, 15$).As the algorithm can converge to local minima, it is launched several times, and the best result is kept. We use the Integrated Completed Likelihood (ICL) criterion as the selection criterion (Biernacki et al., 2000).

We first examine the number of Gaussians, H. The number of Gaussians is important because of its role in temporal activity. More Gaussians will provide a better representation of temporal activity but will take longer to estimate, while too few gaussians are quicker to estimate but could lead to an erroneous representation of a cluster's activity pattern. As an example we can consider Fig. 5. This shows a two-peak activity pattern using a three gaussian mixture model. It can be seen that two gaussians are actually used to represent the peak activity while the attenuated third one is used to represent the remaining activity, which can be useful in different cases.

Fig. 6 shows a gap in the ICL criterion between H equal to 2 and H greater than or equal to 3, meaning that the model is better for H greater than or equal to 3. However there is no significant difference between H equal to 3, 4 or 5, while 4 and 5 gaussians need more computation time for parameter estimation. In view of the fact that the majority of trips are commuter trips (with validations in the morning and evening, leading to two activity peaks) and that no four- or five-peak patterns occur when $H = 4$; 5, we have decided to keep $H = 3$ gaussians.

Knowing that the majority of trips are commuter trips (with validations in the morning and evening, that means two-peaks activity) and that no four or five-peaks pattern occur when $H = 4, 5$, we choose to keep $H = 3$ Gaussians.

The curve of the number of clusters shows us that the more clusters there are, the better the model is. As the aim of this study is to highlight principal passenger behaviours and because the analysis of cluster results with different value of K

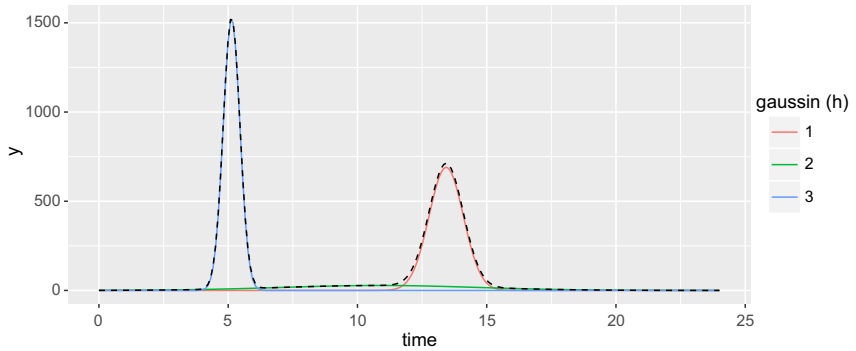


Fig. 5. Plot of a three gaussian mixture model with two activity peaks. The first gaussian (in red) fits the second activity peak. The second gaussian (in green) is flattened and represents the remaining activity. The third gaussian fits the first activity peak. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

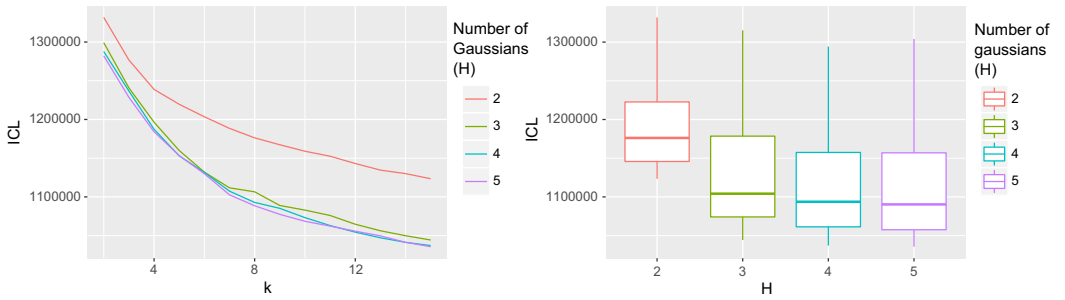


Fig. 6. ICL criterion and boxplot of the ICL criterion for 10 runs, with different numbers of gaussians, $H = 2, \ldots, 5$, and different numbers of passenger clusters, $K = 2, \ldots, 15$, in the mixture model.

showed us attractive results, we choose to fix the number of K clusters at 10. If the purpose of the study is to perform a more detailed analysis of passenger behaviour, the number of clusters can be increased accordingly.

The Boxplot shown in Fig. 5, shows that in addition to the performance of the clusters revealed by the ICL criterion, the results exhibit a degree of stability.

The results section will then detail the clustering obtained with $K = 10$ clusters and for each cluster a mixture of $H = 3$ Gaussians is considered.

5. Smart card clustering results

In the following section, the model presented in the last section is applied to the entire dataset. The clustering results are analysed first. We will then focus on the evolution of cluster composition year to year. In the latter, only cards that were active during the five-year period will be considered (reduced dataset).

5.1. Analysis of clustering results

For a general overview of all the clusters, the temporal activity profiles of all clusters for two different days of the week are shown in Fig. 6. Tuesday is chosen as the typical weekday and Saturday is chosen as the non-weekday. Analysing Tuesday activity (Fig. 6a) reveals the emergence of different patterns of activity. Most are two-peak patterns corresponding to home-to-work commutes. However, some differences can be noted.

First, two clusters (Clusters 7 and 10) seem to show regular afternoon activity, whereas the other clusters show more regular activity in the morning than in the afternoon. A quick look at the card composition (see Fig. 8) reveals the high percentage of student cards, which can explain this particular usage of public transport. Clusters 2, 3, 4, 5 and 8 present the typical two-peak pattern activity with varying morning (between 6:00 AM and 8:20 AM) and evening (between 3:40 PM and 5:00 PM) peak hours.

Even when presenting a two peak-pattern, Cluster 9 has more diffuse activity than other clusters with similar patterns. Finally, Clusters 1 and 6 both present diffuse three-peak pattern activity on Tuesday with more pronounced morning activity for Cluster 6. These different patterns can mainly be explained by the high percentage of cards presented in both clusters (13.4% and 13.6% respectively) and the highest proportion of low activity cards (cards active for fewer than ten days per month). Indeed, approximately 45% of all low activity cards are present in Clusters 1 and 6 (16.13% and 29.58%, respectively).

These clusters can be more specifically characterized by examining their weekend activity 7b. Apart from the fact that all the clusters present lower activity than during the week and that most of activity is concentrated in a diffuse manner in the afternoon, it can be observed that some clusters still present a morning activity peak on the weekend (Clusters 3, 4, 8, 9 and 10) whereas others also present a diffuse night activity (Clusters 1, 7, and 10).

Examining the cluster repartition of each card type reveals which clusters are the most linked to each card type (see Fig. 9). Cluster 1, which had the most diffuse profile, regroups the majority of Senior and College/Univ cards. Clusters 2 and 8 regroup a high proportion of Interzone cards. Student cards are most moved into Cluster 10. Finally, Express adult shows a high number of cards in Clusters 2, 4 and 5, whereas Regular adults seem to be distributed throughout the clusters.

For a more detailed analysis of cluster activity, we now focus on three particular clusters. As previously shown, different patterns emerge from our clustering results. We decide to study Cluster 1 in more detail as a diffuse pattern cluster, Cluster 8 as a classic two-peak pattern and Cluster 10 as a two-peak pattern with regular activity in the afternoon. Despite its diffuse pattern, the morning and afternoon peaks of Cluster 1 (Fig. 10b) are viewable, and the morning peak has lower variance than in the afternoon. A third peak appears during the evening, which can be understood as secondary travel. An interesting feature of this cluster is that weekend activity is higher than in the other clusters. This cluster mainly consists of Regular adults (48.47%), Students (29.71%) and College/Univ (13.74%). Cluster 10 (Fig. 12b) presents a very regular pattern consisting of two peaks around 8 AM and 3:30 PM. The afternoon peak is more concentrated than the morning peak, which can be explained,

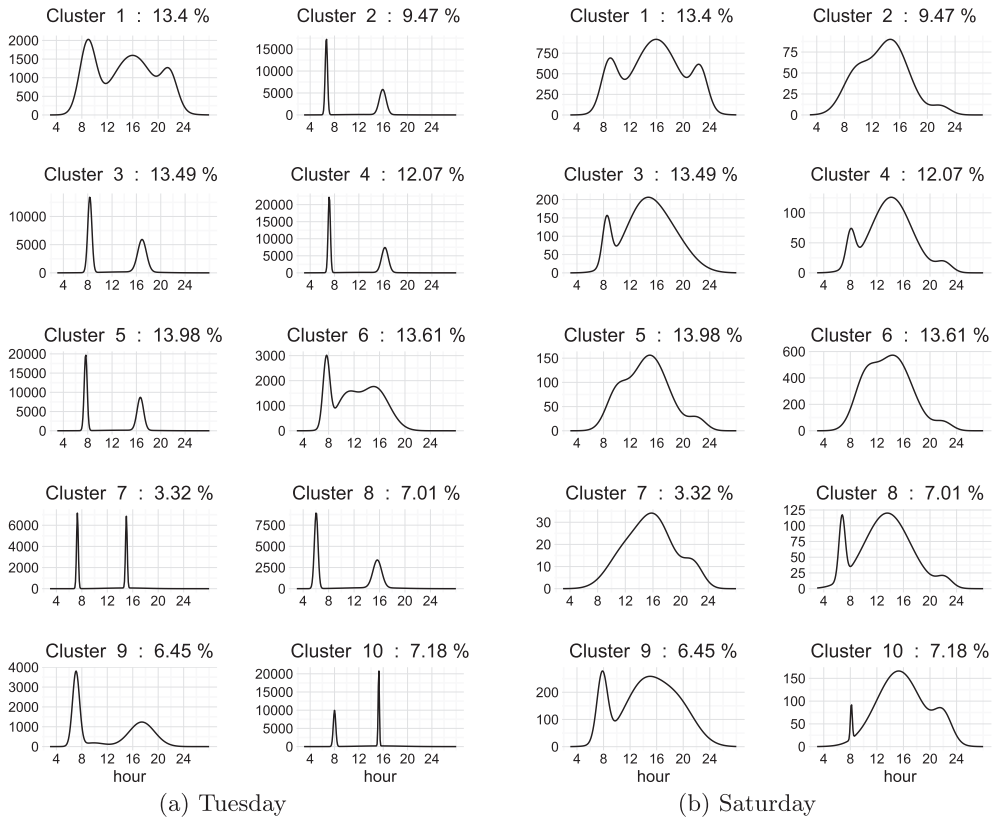


Fig. 7. Tuesday and Saturday temporal activity profiles for all ten clusters.

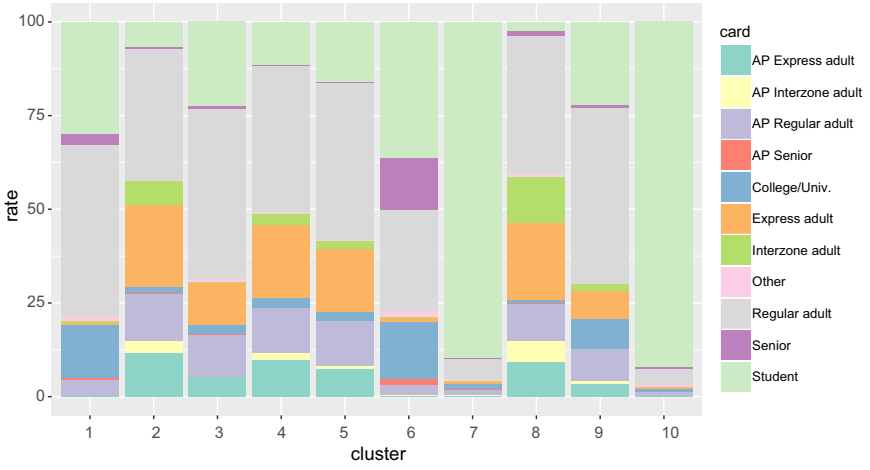


Fig. 8. Distribution of each passenger cluster across card types.

as discussed above, by the high proportion of students (93.5%) that likely have very regular school schedules (high school). Contrary to Cluster 1, it does not show much activity on weekends. Cluster 8 (Fig. 11), along with Cluster 10, present a regular two-peak pattern. However, the variance of the peaks is greater than in Cluster 10, particularly in the afternoon. The morning peak is centred at 6 AM and the afternoon at 3:40 PM. This cluster is more mixed than the other two. Although the cluster has 39.76% Regular adults, it also has 20.09% Express adults, 12.03% Interzone adults and almost no students. The earlier trips can be explained by the distance between passenger homes and workplaces.

Now that we have better knowledge of our passenger profiles, we can extend the analysis and follow the changes over the years.

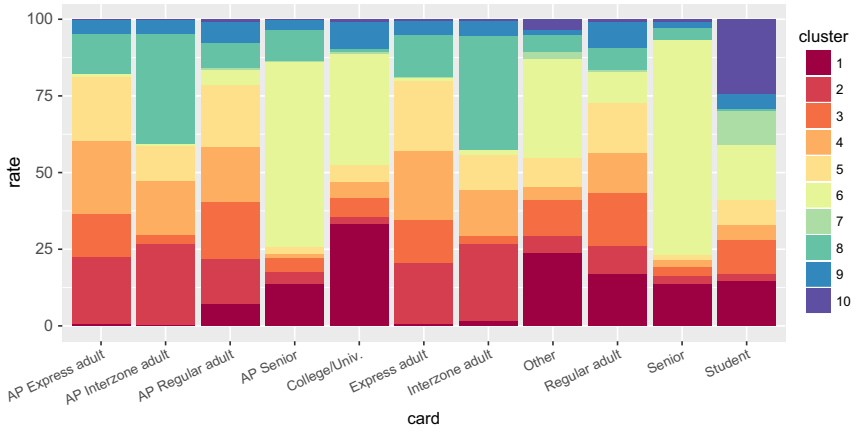


Fig. 9. Distribution of each card type across passenger clusters.

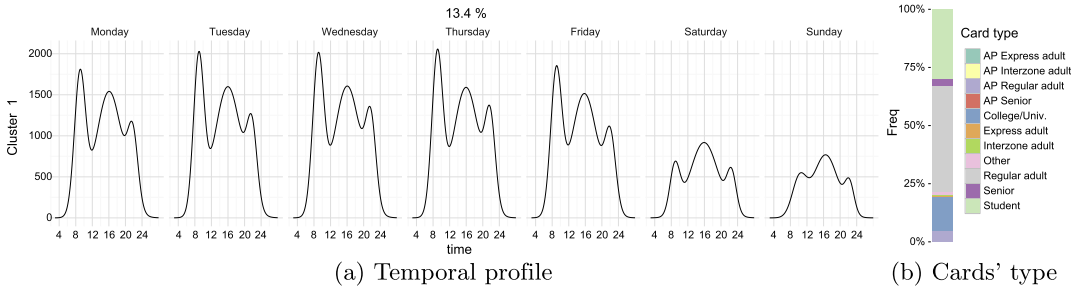


Fig. 10. The temporal activity profiles for each day of the week and card types of Cluster 1 passengers.

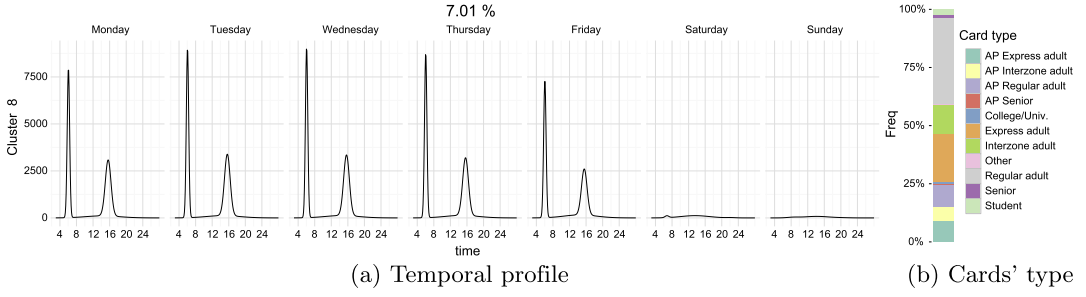


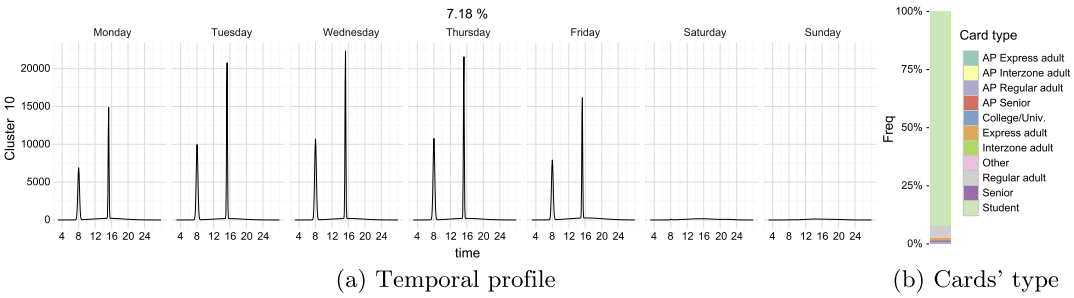
Fig. 11. The temporal activity profiles for each day of the week and card types of Cluster 8 passengers.

5.2. Tracking of cards

This section focuses on the analysis of passenger behaviours over 5 years based on their year-to-year cluster membership changes.

5.2.1. Focus on one cluster (Cluster 8)

First, we shall focus on one cluster in order to ascertain more information about the changes and see whether cards sometimes return to their original cluster. For this part of the study, Cluster 8 has been chosen because of its typical two-peak activity pattern. Fig. 13 shows the cluster membership of cards that were assigned to Cluster 8 for 2005 between 2005 and 2009. Each color corresponds to a different cluster and the year can be read from left to the right (from 2005 to 2009 respectively). It can be seen that most of the cards that left Cluster 8 have been assigned to Cluster 2. Clusters 2 and 8 have closer patterns, which may explain why Cluster 2 received the majority of the lost cards from Cluster 8 (6.6% in 2006). However, we can also see that many cards returned to their original clusters after changing. The proportion of cards leaving Cluster 8 for more diffuse activity clusters (such as Cluster 1 or Cluster 6) was very low (approximately 1%). Such changes can be



(a) Temporal profile (b) Cards' type

Fig. 12. The temporal activity profiles for each day of the week and card types of Cluster 10 passengers.

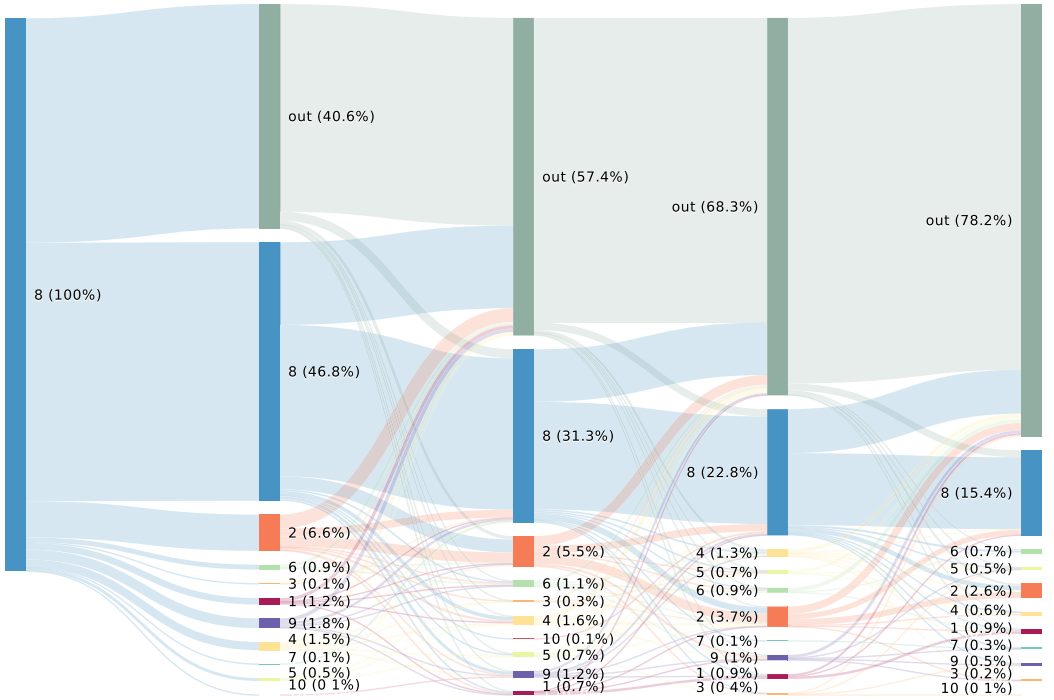


Fig. 13. Cluster membership of cards belonging to Cluster 8 (in blue) in 2005. The assigned cluster of each card over the years is presented using a flowchart. The size of the flow corresponds to the proportion of cards it contains. The "out" group is made up of cards that are no longer active.

explained by activity or location modifications for validation purposes. Further analysis on an individual basis is required to better interpret these modifications. If we now examine the leaving cards ("out" cluster), we can observe that the turnover rate of Cluster 8 cards was on average 60% over the 5 years.

5.2.2. Cluster change over the years

The overall flow of cards between 2005 and 2009 through the different clusters is shown in Fig. 14. For each year and cluster, the proportion of cards belonging to each cluster – as well as the clusters to which the cards are assigned – is shown. The incoming ("nin") cards, that is cards that were not yet active, as well as the leaving ("out") cards, that is cards that will no longer be active, are also shown on the figure. By looking at their flow we can see that a great number of cards were not active for more than one or two years. Some clusters seem to have lost all their cards rapidly from one year to the next.

In order to quantify those cards moving from one cluster to another, Table 1 depicts the mean transition probabilities between clusters estimated from one year to another. These probabilities have been calculated for all the cards and averaged over 5 years. Each line of the table allows us to observe more clearly both the probability that a card will change its cluster from one year to the next and the candidate clusters where the cluster modification could occur. When no cluster modification takes place, the diagonal elements are equal to 100%.

Each cluster mainly exchanges cards with one, two or three clusters. Other card changes are residual. For example, Cluster 8 generally gives cards to Cluster 2 and receives cards from this cluster. We observe a lack of stability for Clusters 7, 9 and 10

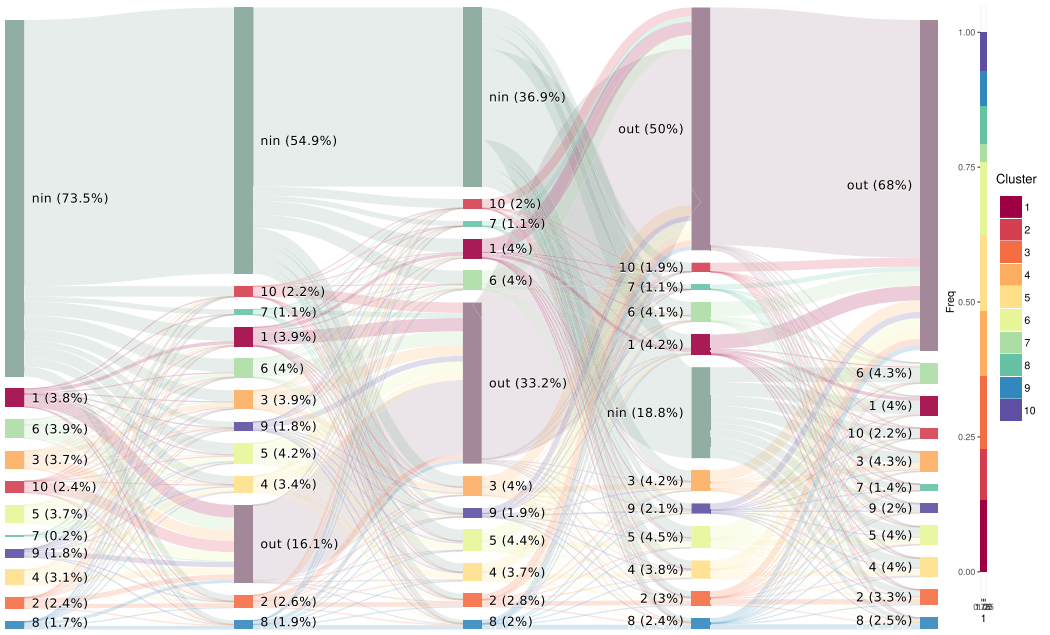


Fig. 14. Proportion of cards per cluster and cluster membership for the entire dataset during the years in which the cards were active (from 2005 to 2009). The assigned cluster of each card throughout the years (2005,. . .,2006) is shown by a flowchart. The size of the flow corresponds to the proportion of cards it contains. The "out" group is made up of cards that were no longer active and the "nin" group of cards that were not yet active.

Table 1
Table of the mean year-to year transition probabilities between clusters.

	1	2	3	4	5	6	7	8	9	10
1	**62.01**	1.24	9.06	1.41	2.68	**14.28**	0.21	2.04	6.29	0.78
2	0.86	**62.63**	1.03	**13.42**	3.18	1.32	0.33	12.92	3.97	0.33
3	5.83	1.31	**66.28**	2.97	**15.69**	3.48	0.07	0.33	3.33	0.71
4	0.95	**14.72**	3.00	**56.47**	**15.54**	1.69	0.61	2.12	4.51	0.38
5	1.59	2.96	**13.28**	**15.82**	**57.75**	2.58	0.26	0.76	4.49	0.51
6	**13.29**	2.17	5.15	3.49	3.68	**65.46**	0.75	1.56	3.05	1.39
7	5.83	5.27	3.17	**20.51**	**11.46**	**16.35**	**24.82**	3.05	6.11	3.43
8	1.73	**11.56**	0.42	2.02	1.24	1.61	0.25	**77.65**	3.25	0.28
9	9.83	8.62	7.48	**11.42**	**10.55**	5.77	0.76	5.91	**39.02**	0.65
10	**11.49**	2.69	**12.30**	5.47	**14.92**	**19.59**	2.18	4.95	3.43	**22.98**

Bold values shows the highest values of transition probabilities.

(the diagonal elements are respectively equal to 25%, 39% and 23%). This can be explained by their large proportion of student cards, whose number, as previously stated, does not stay the same from year to year and is not taken into account when computing the transition probabilities.

The above analysis leads us to incorporate the relationship between the clusters (i.e., how far they are from one another) in the tracking year to the year of the clustering change. We therefore calculate a divergence between all the clusters using the Kullback-Leibler (KL) divergence. The divergence between the two clusters is thus defined as the sum for each day of the week of the KL divergence of the two Gaussian mixtures. More formally,

$$div_{Clust}(cluster_i, cluster_j) = \sum_{l=1}^{7}(div_{KL}(g(.;i,l), g(.;j,l)) + div_{KL}(g(.;j,l), g(.;i,l))) \quad (2)$$

where $g(.;i,l)$ is the density function of the Gaussian mixture distribution of i on day l and $div_{KL}()$ is the Kullback-Leibler divergence.

By using a Hierarchical Ascending Clustering (HAC) on the divergence, it is then possible to obtain a hierarchical view of the different clusters by looking at the associated dendrogram (Fig. 15). Thus, the clustering seems to be ramified in three different groups with similar patterns. The first group, group A, which includes Clusters 1, 6 and 9, is the diffuse activity group. The second group, group B, which includes Clusters 3, 8 and 2, 4, 5, is a two-peak pattern group. Finally, the last group, group C, consists of Clusters 7 and 10, which themselves mostly consist of students, and exhibit very regular patterns during the afternoon.

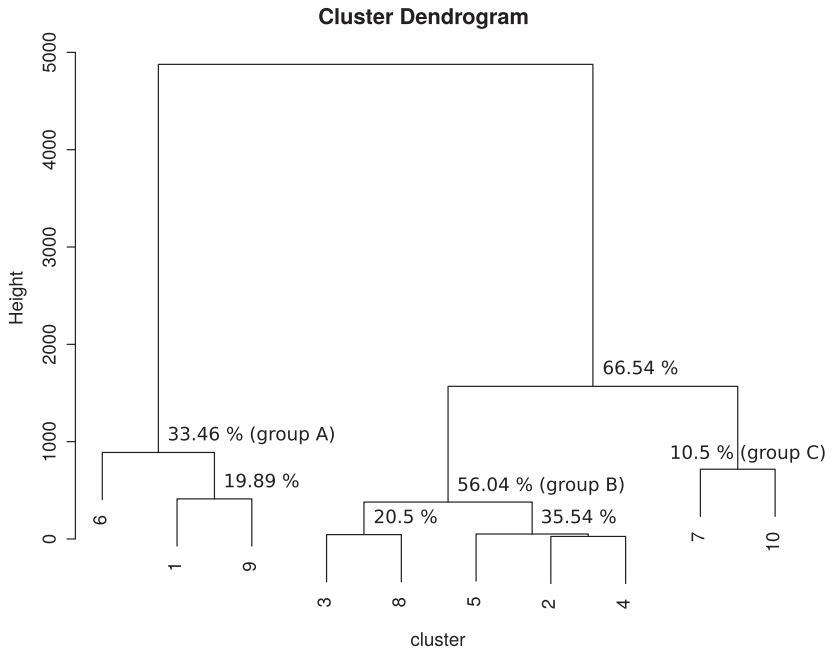


Fig. 15. Dendrogram based on cluster divergence clustering using Hierarchical Ascending Clustering.

Returning to our cluster tracking, we can better understand the cluster changes. Indeed, we can confirm that clusters mostly exchange their cards with clusters with similar patterns. For example, Clusters 2 and 4, which are the most closely situated in the dendrogram, exchange a large number of cards, which is likely because the clustering is performed by taking every day of the week into account, and just a few changes in user habits can move the user into a different close activity cluster. Moreover, 53.5% of the first group cards (diffuse group) and 80.2% of the second group cards (two-peak pattern) stay in the same group during the five years of study. The third group (mostly composed of student cards) being almost empty, it doesn't make sense to look at its stability during the years.

To provide a better overview of which groups stay stable and which do not, Fig. 16 shows the same flow chart as earlier using our larger groups A, B and C. It may be recalled that group A is characterized by a diffuse activity pattern, group B exhibits a two-peak activity pattern whereas group C mostly consists of students with very regular activity patterns during the afternoon. This figure confirms that a large number of the cards belonging to group B stay in this group, while the cards belonging to group A change clusters more easily. Cards belonging to group C which are mostly owned by students are potentially volatile since they can be renewed every year. Overall, we can conclude that the clustering exhibits significant stability from one year to the next for cards that stay active.

5.3. Spatial analysis using entropy

Thus far, the analysis conducted has been only temporal. However, as discussed above, the original data also include boarding locations. In this section, we present a spatial characterization of our clustering results using the reduced dataset.

To conduct this analysis, we investigate the spatial entropy of each card. This entropy is based on the probability of each card being validated at one of its most frequently used stations (i.e., the stations that are most active for the card). Each card's spatial activity is described using a multinomial distribution for its different stations. The entropy used here is Shannon entropy (Shannon, 1948), which is defined as:

$$H(X) = -\mathbb{E}[\log P(X = x_i)] = -\sum_{i=1}^{n} P_i \log P_i.$$

We then use a violin plot for the entropy of each cluster in Fig. 17. Three different types of violins can be observed. Moreover, these three types correspond to the three types of clusters that we had previously observed in Section 5.2. The two student clusters (Clusters 7 and 10) cannot be further examined because of the very small number of smart cards they contain (between 19 and 32 cards depending on the year, i.e. 0.8% and 1.3% of the total number of cards). Indeed, as discussed above, they do not have enough cards to present significant results. Clusters 1, 6 and 9, which present more diffuse temporal activity, seem to also present more diffuse spatial activity. The mean entropy of Clusters 1, 6, and 9 is higher than the entropy of the other clusters.

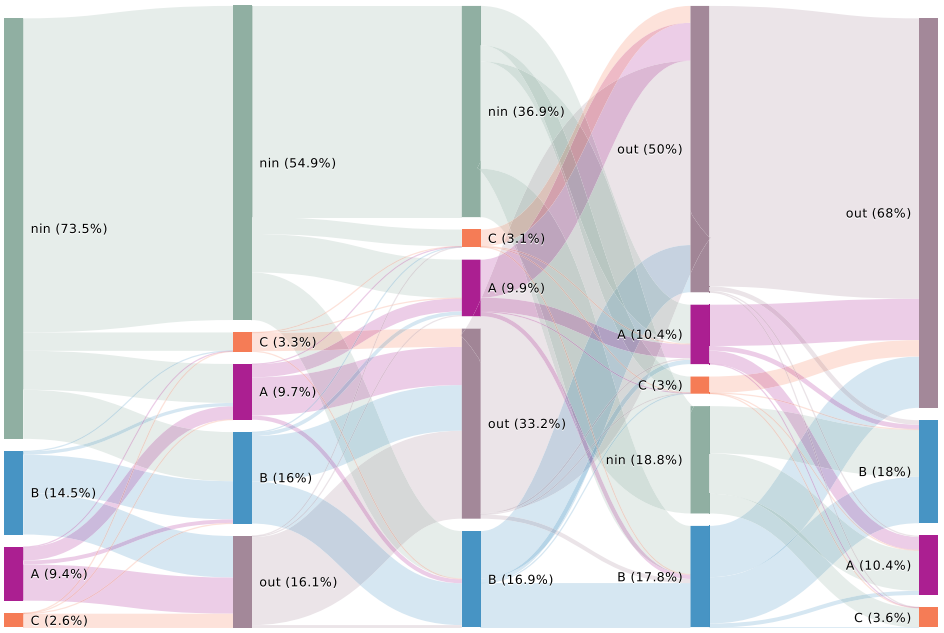


Fig. 16. Percentage of cards belonging to each group for the entire dataset and group memberships during the years in which the cards were active (from 2005 to 2009). Group A (purple) is composed of diffuse activity cards, group B (blue) of regular two-peak activity pattern cards and group C of student cards. The group to which each card is allocated over the years is presented using a flowchart. The size of the flow corresponds to the proportion of cards it contains. The "out" group is made up of cards that were no longer active and the "nin" group of cards that were not yet active. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

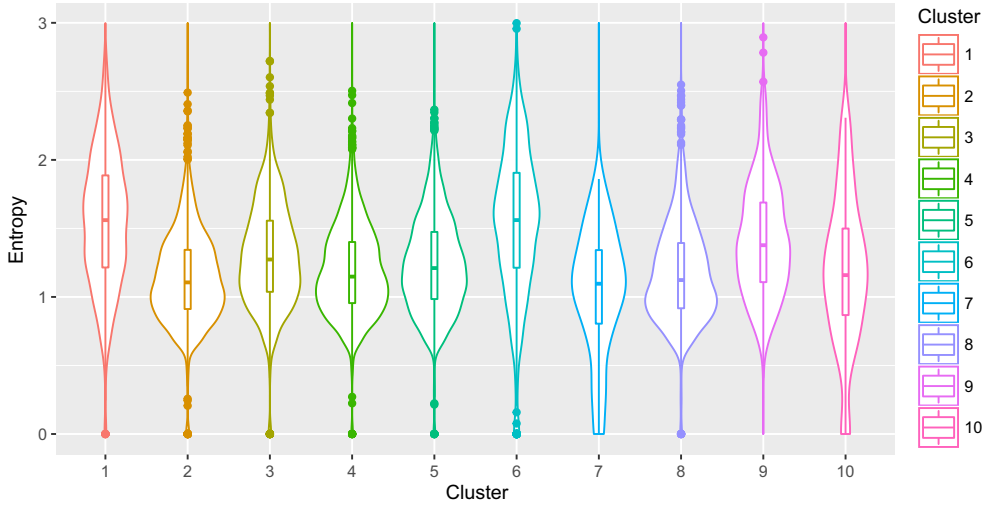


Fig. 17. Measurement of the entropy of validation location for each card in the different clusters for April of each year between 2005 and 2009.

6. Conclusion

This paper presented an analysis involving a clustering of transportation smart cards from Gatineau City, in Canada. Using a Gaussian mixture generative based model, a clustering of passengers was performed based on passengers' temporal activities. A follow-up on the cards' cluster memberships over a span of years was also performed.

A detailed analysis of the clusters using their statistical characteristics (peak means and variance) shows that in addition to common temporal patterns corresponding to the pendular home to work/study commute patterns, some groups make their trips earlier or are more regular than others, while other passengers have more diffuse activities, leading to a three-peak pattern. Even so, these latter passengers still exhibit a morning activity peak, particularly for the cluster with the greatest number of cards. Analysing the evolution of cluster assignments of the same card over a period of multiple years reveals

that even if some changes in the cluster emerge over the years, the clusters seem to maintain the same proportion of cards, and the majority of cards that move from clusters, move to clusters with temporal profiles that are similar to their original clusters.

Additional analyses must be performed to extend this work. We chose the number of clusters based on a selection criterion, which results in a trade-off between interpretation and fitting the model to the dataset. It would be interesting to investigate the choice of the number of clusters if we intend to highlight passengers with atypical behaviours. Moreover, as can be seen, the proposed model takes into account the days of the week, which could increase the number of clusters combinatorially. In Gatineau the main difference in transport activity is linked to belonging to a weekday or a weekend. Our approach was first developed in Briand et al. 2016 to analyse a dataset collected on the public transport of Rennes in France and has revealed some differences in activity in the city of Rennes depending on the day, especially in the case of Wednesdays and Fridays, in addition to the inherent difference between activity during the weekend and on weekdays. It would be interesting in future work to compare two models, one taking into account the day of the week and the second taking into account only the type of day (weekday or weekend). Experiments using the two models would be carried out on two datasets collected in two different cities namely, Rennes in France and Gatineau in Canada. This kind of study could help to highlight the close link between the type of statistical modelling and temporal habits in public transportation usage in a city. Furthermore, some additional data should be considered on the card. The data used in our analysis are incomplete (lost and stolen cards don't keep the same ID when they are replaced). And finally, an in-depth analysis is required to better understand the motivations behind the cards' cluster changes. To this end, a more dedicated model should be developed. Furthermore, some additional data should be considered on the card. The data used in our analysis are incomplete (lost and stolen cards don't keep the same ID when they are replaced). And Finally, an in-depth analysis is required to better understand the motivations behind the cards' cluster changes. To this end, a more dedicated model should be developed.

Acknowledgements

The authors wish to acknowledge the Société de transport de lOutaouais, in Gatineau, Quebec, for their support of the work. Part of this research was also funded by a mobility grant of Paris-Est University awarded to the author.

References

Agard, B., Morency, C., Trépanier, M., 2006. Mining public transport user behaviour from smart card data. In: The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM), pp. 17–19.
Ahillen, M., Mateo-Babiano, D., Corcoran, J., 2015. The dynamics of bike-sharing in Washington, D.C. and Brisbane, Australia: implications for policy and planning. Int. J. Sustain. Transp. 10 (5), 441–454.
Arana, P., Cabezudo, S., Pealba, M., 2014. Influence of weather conditions on transit ridership: a statistical study using data from smartcards. Transp. Res. Part A: Policy Pract. 59, 1–12.
Bagchi, M., White, P.R., 2004. What role for smart card data from bus systems. In: Proceedings of the Institution of Civil Engineers. Municipal Engineer. March Issue ME1.
Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–474.
Barry, J., Newhouser, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York city with automated fare system data. Transp. Res. Rec. 1817, 183–187.
Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 22 (7), 719–725. Jul.
Briand, A.-S., Côme, E., El Mahrsi, M.K., Oukhellou, L., 2016. A mixture model clustering approach for temporal passenger pattern characterization in public transport. Int. J. Data Sci. Anal., 1–14
Ceapa, I., Smith, C., Capra, L., 2012. Avoiding the crowds: understanding tube station congestion patterns from trip data. In: Proceeding of the 1st ACM SIGKDD International Workshop on Urban Computing. ACM press, pp. 134–141.
Chu, K., Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling. Transp. Res. Rec.: J. Transp. Res. Board 2063, 63–72.
Chu, K.K.A., 2015. Two-year worth of smart card transaction data extracting longitudinal observations for the understanding of travel behaviour. Transp. Res. Procedia 11, 365–380. Transport Survey Methods: Embracing Behavioural and Technological Changes Selected Contributions from the 10th International Conference on Transport Survey Methods 16–21 November 2014, Leura, Australia.
El Mahrsi, M.K., Côme, E., Baro, J., Oukhellou, L., 2014. Understanding passenger patterns in public transit through smart card and socio-economic data. In: 3rd International Workshop on Urban Computing (UrbComp), ACM SIGKDD Conference. New York, USA, p. 9. August.
Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., Bento, C., 2014. Catch me if you can: predicting mobility patterns of public transport users. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, pp. 1995–2002. Oct.
Fuse, T., Makimura, K., Nakamura, T., 2010. Observation of travel behavior by ic card data and application to transportation planning. In: Special Joint Symposium of ISPRS Commission IV and AutoCarto 2010.
Hofmann, M., O'Mahony, M., 2005. Transfer journey identification and analyses from electronic fare collection data. In: Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE. IEEE, pp. 34–39.
Langlois, G.G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. Part C: Emerg. Technol. 64, 1–16.
Lathia, N., Capra, L., 2011. How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. In: Proceedings of the 13th International Conference on Ubiquitous Computing. UbiComp '11. ACM, New York, NY, USA, pp. 291–300.
Lathia, N., Froehlich, J., Capra, L., 2010. Mining public transport usage for personalised intelligent transport systems. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 887–892. Dec.
Lee, S.G., Hickman, M., 2013. Trip purpose inference using automated fare collection data. Public Transp. 6 (1), 1–20.
Ma, X.-l., Wu, Y.-J., Wang, Y.-h., Chen, F., Liu, J.-f., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C: Emerg. Technol. 36, 1–12.
McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions. Wiley.
Morency, C., Trépanier, M., Agard, B., 2006. Analysing the variability of transit users behaviour with smart card data. In: The Ninth International IEEE Conference on Intelligent Transportation Systems, Toronto, Canada, September.
Morency, C., Trpanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14 (3), 193–203.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. Transp. Res. Part C: Emerg. Technol. 44, 70–79.
Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smart card data from santiago, chile. Transp. Res. Part C: Emerg. Technol. 24, 9–18.
Nassir, N., Hickman, M., Ma, Z.-L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation 42 (4), 683–705.
Pang, L.X., Chawla, S., Liu, W., Zheng, Y., 2013. On detection of emerging anomalous traffic patterns using GPS data. Data Knowl. Eng.
Park, J.Y., Kim, D.-J., Lim, Y., 2008. Use of smart card data to define public transit use in Seoul, South Korea. Transp. Res. Rec.: J. Transp. Res. Board 2063, 3–9.
Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C: Emerg. Technol. 19 (4), 557–568.
Poussevin, M., Baskiotis, N., Guigue, V., Gallinari, P., 2014. Mining ticketing logs for usage characterization with nonnegative matrix factorization. In: SenseML 2014 – ECML Workshop. Nancy, France. Sep..
Seaborn, C., Attanucci, J., Wilson, N.H.M., 2009. Analyzing multimodal public transport journeys in London with smart card fare payment data. Transp. Res. Rec.: J. Transp. Res. Board 2121, 55–62.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3–4). 379–423 and 623–656.
Tao, S., Rohde, D., Corcoran, J., 2014. Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J. Transp. Geogr. 41, 21–36.
Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. Intell. Transp. Syst. 11, 1–14.
Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transp. Res. Rec. 1971, 119–126.
van der Hurk, E., Kroon, L., Maroti, G., Vervest, P., 2015. Deduction of passengers' route choices from smart card data. IEEE Trans. Intell. Transp. Syst. 16 (1), 430–440. Feb.
Wang, W., Attanucci, J.P., Wilson, N.H., 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. J. Public Transp. 14 (4).
Zhang, F., Yuan, N.J., Wang, Y., Xie, X., 2014. Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach. Knowl. Inf. Syst., 1–25
Zhao, J., Rahbee, A., Wilson, N., 2007. Estimating a rail passenger trip origin?destination matrix using automatic data collection systems. Comput.-Aided Civil Infrastruct. Eng. 22, 376–387.
Zheng, Y., Capra, L., Wolfson, O., Yang, H., 2014. Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. October.