Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon

Oscar Egu, Patrick Bonnel

Version of Record: https://www.sciencedirect.com/science/article/pii/S2214367X19302315
Manuscript_a838fa621a7767543458e15902a25e75

Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon.

Oscar Egu[1,*] and Patrick Bonnel[1]

[1] *LAET-ENTPE, Université de Lyon, CNRS, Rue Maurice Audin, 69518 Vaux-en-Velin Cedex, France*
[*] *Corresponding author: Oscar Egu, oscar.egu@entpe.fr*

Abstract

To examine the variability of travel behaviour over time, transportation researchers need to collect longitudinal data. The first studies around day-to-day variability of travel behaviour were based on surveys. Those studies have shown that there is considerable variation in individual travel behaviour. They have also discussed the implications of this variability in terms of modelling, policy evaluation or marketing. Recently, the multiplication of big data has led to an explosion in the number of studies about travel behaviour. This is because those new data sources collect lots of data, about lots of people over long periods. In the field of public transit, smart card data is one of those big data sources. They have been used by various authors to conduct longitudinal analyses of transit usage behaviour. However, researchers working with smart card data mostly rely on clustering techniques to measure variability, and they often use conceptual framework different from those of transportation researchers familiar with traditional data sources. In particular, there is no study based on smart card data that explicitly measure day-to-day intrapersonal variability of transit usage. Therefore, the purpose of this investigation is to address this gap. To do this, a clustering method and a similarity metric are combined to explore simultaneously interpersonal and intrapersonal variability of transit usage. The application is done with a rich dataset covering a 6 months period (181 days) and it contributes to the growing literature on smart card data. Results of this research confirm previous works based on survey data and show that there is no one size fits all approach to the problem of day-to-day variability of transit usage. They also prove that combining clustering algorithm with day-to-day intrapersonal similarity metric is a valuable tool to mine smart card data. The findings of this study can help in identifying new passenger segmentation and in tailoring information and services.

Keywords: Public transit, travel behaviour, smart card data, passenger clustering, day-to-day variability, user segmentation

Acknowledgements

This research was conducted as part of a research agreement between Keolis Lyon and the Urban Planning, Economics and Transport Laboratory (LAET). Their financial support is gratefully acknowledged. We would like to thank our colleagues from Keolis Lyon who provided the data and expertise that greatly assisted the research. The views expressed in this paper remain those of the authors.

Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon.

Abstract

To examine the variability of travel behaviour over time, transportation researchers need to collect longitudinal data. The first studies around day-to-day variability of travel behaviour were based on surveys. Those studies have shown that there is considerable variation in individual travel behaviour. They have also discussed the implications of this variability in terms of modelling, policy evaluation or marketing. Recently, the multiplication of big data has led to an explosion in the number of studies about travel behaviour. This is because those new data sources collect lots of data, about lots of people over long periods. In the field of public transit, smart card data is one of those big data sources. They have been used by various authors to conduct longitudinal analyses of transit usage behaviour. However, researchers working with smart card data mostly rely on clustering techniques to measure variability, and they often use conceptual framework different from those of transportation researchers familiar with traditional data sources. In particular, there is no study based on smart card data that explicitly measure day-to-day intrapersonal variability of transit usage. Therefore, the purpose of this investigation is to address this gap. To do this, a clustering method and a similarity metric are combined to explore simultaneously interpersonal and intrapersonal variability of transit usage. The application is done with a rich dataset covering a 6 months period (181 days) and it contributes to the growing literature on smart card data. Results of this research confirm previous works based on survey data and show that there is no one size fits all approach to the problem of day-to-day variability of transit usage. They also prove that combining clustering algorithm with day-to-day intrapersonal similarity metric is a valuable tool to mine smart card data. The findings of this study can help in identifying new passenger segmentation and in tailoring information and services.

Keywords: Public transit, travel behaviour, smart card data, passenger clustering, day-to-day variability, user segmentation

1 Introduction

Travel behaviour research has been prominently based on cross-sectional data where individuals are asked to report their travel behaviour on a single day [Gärling and Axhausen, 2003, Pas, 1986, Hanson and Huff, 1986]. However, using one-day observation is in general insufficient because individual needs and desires that generate travel, vary from day-to-day [Pas, 1987] and because classifications based on a single day are prone to be unstable [Hanson and Huff, 1986]. Multiday data are therefore needed to refine the understanding of travel behaviour and measure how it may vary from day-to-day [Hanson and Huff, 1981, Schlich and Axhausen, 2003]. This is an important area of research that has several practical applications such as assessment of policy impact [Jones and Clarke, 1988], implementation of travel demand management strategies and individualised marketing [Gärling and Axhausen, 2003], travel behaviour modelling [Pas, 1986, 1987] or even market segmentation [Hanson and Huff, 1986].

To increase patronage, public transit operators also need to develop and evaluate new strategies. This requires a comprehensive understanding of transit usage behaviours, but also the ability to measure the multiday dynamics of individual demand [Morency et al., 2007, Briand et al., 2017, Ma et al., 2013, Bhaskar et al., 2015, Zhao et al., 2018]. In large urban areas, travel patterns are heterogeneous [Goulet-Langlois et al., 2016] and for decades transportation researchers have been exploring this heterogeneity with data of active solicitation [Schlich and Axhausen, 2003, Chen et al., 2016]. Unfortunately, active multiday data are costly, difficult to collect and often limited in terms of sample size [Goulet-Langlois et al., 2016, Briand et al., 2017, Gärling and Axhausen, 2003, Chen et al., 2016, Schlich and Axhausen, 2003]. Thanks to recent advances in technologies, it is now possible to collect continuously and passively massive data about mobility [Chen et al., 2016] with less or no burden for respondent [Bagchi and White, 2005]. In the field of public transit, smart card data is considered to be one of the most promising passive data sources [Pelletier et al., 2011]. As opposed to extrinsic mobility data such as mobile phone data, it is an intrinsic mobility data that is generated by travel events and therefore it provides direct information about transit usage [Zhao et al., 2018]. This data sources can be used to measure variability [Morency et al., 2007] and has resulted in a multiplication of studies on transit usage pattern. However, as noted by Chen et al. [2016] studies based on passive data sources often lack the long used conceptual framework of transportation researchers familiar with active data sources.

In this context, the purpose of this study is to investigate day-to-day transit usage variability using the conventional concept of daily trip pattern. The application is done with 6 months of smart card data. Two dimensions of variability are measured in parallel and interpreted with available socio-demographic profile derived from the type of fare used by each card. More precisely, day-to-day interpersonal variability is examined with a clustering method designed for this specific analysis. Day-to-day intrapersonal variability is measured with a trip based similarity metric [Huff and Hanson, 1986] taking into account the daily trip rate and the spatiotemporal characteristic of trip pattern. This paper shows that combining interpersonal clustering with traditional intrapersonal similarity metric is a valuable approach to mine smart card data. It can extend our knowledge of day-to-day variability of transit usage. Results of this research can assist transit marketers in defining more meaningful passenger segmentation. They can also help in tailoring information and services.

The remainder of the paper comprises four sections organised as follows. The first part will review related works and clarify the research needs. The second part will describe the data and methods. The third part will present the main results of this research. Finally, in the fourth section, the empirical findings will be synthesised and future directions of research will be given.

2 Literature review

2.1 Studies based on active data collection

The first serious discussions about travel behaviour variability emerged during the 1980s. At that time, multiday data were difficult to collect due to high response burden [Schlich and Axhausen, 2003] and risk of deterioration of data quality over the survey period [Hanson and Huff, 1981]. In a series of work based on a 35 consecutive day travel survey known as the Uppsala survey, Huff and Hanson [1986], Hanson and Huff [1981, 1988] have found significant systematic intrapersonal variability and showed that daily travel behaviour is neither totally repetitious neither totally variable. With the same dataset, they have also identified five clusters of individuals based on multiday travel characteristics [Hanson and Huff, 1986]. Even if the five groups share distinctive travel behaviour and socio-demographic attributes, the authors noted that there is still substantial intragroup variance. They suggest that future classifications should recognize multiday patterns and recognize that individuals have more than one habitual daily travel pattern. Pas and Koppelman [1987] with a 5-day travel diary survey have examined the determinant of day-to-day trip rate variability and conclude that there are large intrapersonal variability and significant differences across socio-demographic groups. In a somehow related paper, Pas [1987] has investigated the effect of day-to-day variability on model goodness-of-fit. They divide the total sum of square of a standard least square trip generation models between interpersonal variability and intrapersonal variability and report that a substantial proportion of the total variability is due to intrapersonal variability. Jones and Clarke [1988] have discussed the importance of taking into account day-to-day variability from a policy perspective notably to assess the impact of measure that affects multiday behaviour. They proposed an activity-based measure of variability and noted that different measure can lead to different conclusions. Using the data from the mobidrive six-week travel diary, Schlich and Axhausen [2003] have compared various measures of day-to-day similarity. They noted that different measures may have different interpretations in terms of variability and conclude that day-to-day behaviour is more variable if measured with trip-based methods than with activity-based methods. They also found that travel behaviour is more stable on weekdays and recommend that surveys about travel behaviour cover period of at least two weeks. More recently, Susilo and Axhausen [2014] have used Herfindahl-Hirschman index to examine the degree of repetition of daily-activity-travel-location. Their results indicate that constraint activities such as work or school have more repetitive combination than leisure and private business trips. With a seven-day travel diary collected in Belgium, Raux et al. [2016] have proposed different measurement methods to analyse interpersonal and intrapersonal variability within different periods of the week. Their results confirmed the overall picture that emerged from those studies: there is an important intrinsic variability in daily travel behaviour.

2.2 Studies based on smart card data

The first research on transit usage variability with smart card data was initiated in Canada. Agard et al. [2006] have used smart card data from the Société de Transport de L'Outaouais (STO) to cluster card users over 12 weeks based on the weekly temporal characteristic of trips. Change in cluster composition was then measured to explore intrapersonal variability. Morency et al. [2007] with 10 months of data from the STO have investigated separately the spatial and temporal variability of transit users. Their sample included only cards that were observed travelling at least once in the first and last month of their study period. The spatial variability is measure through the frequency of usage of bus stops and the temporal variability is evaluated with a clustering method based on boardings times. Ma et al. [2013] have investigated the regularity of trips pattern using 5

days of data from Beijing. For each card, the data is aggregated into four scalar features: number of travel days, number of similar first boarding times, number of similar route sequence and number of similar stop ID sequence. The K-means ++ algorithm is then used to identify cluster with different level of regularity. A rough set approach is also proposed to enhance the performance of the algorithm for large dataset. Bhaskar et al. [2015] have used four months of working days data from the transit authority of SEQ (Australia) to segment passenger based on the spatiotemporal variability of their trips. The proposed methodology starts with the application of DBSCAN to identify independently regular origin-destination (OD) trips and habitual trips starting time. Then, each passenger is described with the percentage of regular OD trips and the percentage of regular trips starting time. Using a priori rules, users are then segmented into four categories. For example, transit commuters are defined as those who make more than 50% of trips within habitual time and between regular OD. Briand et al. [2017] have proposed a Gaussian mixture model to cluster typical trips temporal pattern and measure the evolution of cluster composition over multiple years to assess change in passenger behaviour. They noticed some changes in the cluster composition but conclude that the majority of cards move to cluster with similar characteristics. Manley et al. [2018] have processed three months of data from London Oyster smart card to identified clusters of travel event for each individual with DBSCAN. A bottom-up approach is then used to derive a system-wide spatiotemporal understanding of regularity and irregularity. The analysis reveals that there are more regularities in the origin of trips in the suburbs than in central London. Goulet-Langlois et al. [2016] have proposed a longitudinal representation of passenger activity based on the sequence of location (user area) infer for each card. Principal component analysis is then used to reduce the dimension of each sequence and serve as an input to cluster analysis. They apply this method on frequent users over 29 days of data from London and found 11 clusters with distinct sequence structure. The variability of each sequence is then measured with entropy rate to take into account the order of travel events and to detect individual with more variability [Goulet-Langlois et al., 2018]. Deschaintres et al. [2019] have focused on weekly variability of daily trip rate using smart card data from Montréal. Their sample includes cards observed with an amplitude of 12 months. Using the K-means clustering algorithm a week typology is created and each card is then represented as a sequence of week cluster. Those sequences are then used to cluster interpersonal variability and measure intrapersonal variability.

2.3 Research gap

The literature review shows that in studies based on smart card data researchers rely on clustering techniques to investigate and measure variability. To construct clustering variables, researchers often used scalar or vector aggregation of passenger's trips attributes [Goulet-Langlois et al., 2016] or generative model (e.g [Briand et al., 2017]). Clusters are then used: (1) to group passengers with similar travel behaviour i.e to study interpersonal variability, (2) to assess intrapersonal variability by studying cluster membership through time and (3) to identify events and sequence that are not regular. Those approaches are interesting but do not provide a metric of variability such as those proposed by authors using active data. Moreover, the clustering variables are often built aggregating more than one day of data. However, the conventional paradigm of travel research is based on the concept of daily trip pattern and many previous authors have argued that variability should be measured between days [Hanson and Huff, 1988, Pas and Koppelman, 1987, Schlich and Axhausen, 2003]. Finally, authors using smart card data often limit the scope of their analysis to a certain type of users (e.g frequent users), to a reduced temporal period (e.g a month), or to only one dimension of travel behaviour (e.g temporal pattern or spatial pattern).

For those reasons, we believe that the problem of measuring day-to-day public transit usage variability with smart card data has not been addressed completely. Therefore, there is a gap that

needs to be filled. To achieve this objective, two flexible methods that can be applied to any type of users, any type of days, and any temporal period, are implemented. The first one is a clustering algorithm that allows to visualize and identify the most common day-to-day usage pattern and explore intrapersonal variability in a straightforward manner. The second one is a similarity index [Huff and Hanson, 1986] designed to measure day-to-day intrapersonal variability taking into account three fundamental features of daily trip pattern: space, time and trip rate. Our approach is based on the assumption that the day is the fundamental period for travel behaviour analysis, thus the intrapersonal day-to-day variability of transit usage needs to be explicitly measured. The application is done with a rich dataset from the public transit networks of Lyon covering a 6 months period. Results are then cross-checked with the available fare profile to understand the potential determinant of day-to-day variability. To the best of our knowledge, no previous study based on smart card data has gone so far into the analysis of day-to-day variability. Furthermore, there is no study that combines clustering methods with day-to-day similarity measurement. This work contributes to the reconciliation of traditional methods with novel datasets and clustering techniques. Findings can help to better understand the dynamics of individual transit usage. They can also assist transit marketers and operators in defining meaningful passenger segmentation.

3 Materials and methods

This section starts by providing a brief description of the most important aspect of this case study. Then, the methods to measure variability are specified in detail.

3.1 Materials

3.1.1 Case study

TCL ("Transport en Commun Lyonnais") is the commercial name of the public transit network of Lyon. The network consists of 4 metro lines, 2 funicular lines, 5 tramway lines and more than 100 regular bus lines. On a working day, approximatively 1.2 million trips are done on the network. The fare transaction system of TCL is an entry only system. All transactions are anonymized and stored with boarding time and location. Smart card and magnetic paper tickets can be used by passengers. Cards cost 5€, they are strictly personal and require an identity photo of the owner. Thereafter, we will assume that there is an unambiguous relation between users, cards and individuals. The three terms will refer to a single person. We will also assume that there is no lost or stolen card. Finally, because paper tickets cannot be traced through time, the analysis is strictly restricted to cards (which represent 75% of the total number of fare transactions).

3.1.2 Fare profile determination

Smart card data often lack socio-demographic informations [Pelletier et al., 2011, Bagchi and White, 2005]. However, when there is a large spread of fare product, it is possible to define categories that make sense from a socio-demographic point of view. In Lyon, cardholders can access to a broad range of fares like annual pass, monthly pass, weekly pass but also access to standard single trip fare. They can also benefit from reduced prices if they can show proper justifications. To assign a socio-demographic profile to each card, we have aggregated fare product according to the table 1. During the lifetime of a card (5 years), users can purchase different types of fare products. In this research, the profile of each card is defined as the one that accounts for the biggest proportion of fare transactions. With this ad-hoc method we were able to add one socio-demographic dimension to the data, however, it should be acknowledged that this information may

not be available in other city or at least not as specific as in this case study.

Fare profile	Fare type	Required justification	Pricing (€)
Student	Monthly and annual pass	Proof of university enrolment and under 28 years old.	31.5
Young	Monthly and annual pass	Student up to high school (18 years old) that reside within the TCL network perimeter.	9-31.5
Elderly	Monthly and annual pass	More than 65 years old or more than 60 years old and retired.	9-31.5
Social	Monthly and annual pass	Dedicated to individuals that can justify low revenue such as unemployed people or people that benefit from state allowance.	9-31.5
General public	Monthly and annual pass	No justification needed can be half reimbursed by the employer of cardholder.	44.1-63.2
Short duration	Multi-day pass up to one week and single trip fare	May be available at a reduced price with justification (e.g. for young people, student or large family) .	1.7-19.3
Intermodality	Monthly and annual pass	Combine with other transportation mode such as rail, regional bus or public transit network from other city. May be available at a reduced price with justification (e.g. for young people or students).	47.5-205.8
Other	Monthly and annual pass	Mainly free pass for the public transportation operators agents and family or very specific passenger (blind people, policeman etc.)	0-6

Table 1: Fare product classification into fare profile and corresponding prices for 2017 fiscal year, source: Authors

3.1.3 Trip identification

Trips are the building blocks of human travel behaviour. They are defined as a movement through time and space between two locations where activities are carried out [Bonnel, 2002]. In the smart card transaction database, records include boardings that are the beginning of a trip but also transfers. To identify the beginning of trips, we implement the following rules : (1) the first transaction of a day is always the beginning of a new trip, (2) two boardings transactions that occur within 60 minutes and that are not made on the same line or on metro station, are considered as part of the same trip [Munizaga et al., 2014, Devillaine et al., 2012, Deschaintres et al., 2019]. The 60 minutes rule was defined according to the current fare policy that stipulates that a single ticket is valid up to 60 minutes from the previous validation. Sensibility test have shown that increasing the time threshold up to 120 minutes has no impact on the results.

3.1.4 Study period

In this research, data from January 1st 2017 to June 30th 2017 were extracted from the fare collection database. This study period was chosen for two reasons. First, because it consists of 181 days which we believe is enough to investigate day-to-day variability as there are at least 25 days of observations for each day of the week; second, because this period includes two school holiday periods: winter break (from 2017-02-18 to 2017-03-05) and spring holidays (from 2017-04-15 to 2017-05-01) but also six bank holidays (2017-01-01, 2017-04-17, 2017-05-01, 2017-05-08, 2017-05-25, 2017-06-05). Those events can affect individual usage pattern and may be of interest in terms of variability. Throughout this paper, a day will be referred to as a holiday day if it is a weekday that is within the two periods of school holidays or if it is a bank holiday. The rest of the weekdays will be considered as working days. Saturday and Sunday are considered separately. Note also that a day (or service day) is defined from 4.30 a.m to 4.30 a.m of the next day when the activity of the network is null but also when the majority of people are asleep.

3.2 Methods

3.2.1 Clustering interpersonal variability

Clustering analysis is one of the most common techniques of data mining [Friedman et al., 2001]. It aims to group objects that are similar in the same cluster which makes it a valuable strategy to study interpersonal variability and give more semantic to raw data. The three main steps in clustering are the definition of a vector space, the definition of a metric distance and the grouping of objects based on their similarity in the vector space.

At the most basic level, the day-to-day transit usage pattern of a card k could be described using a boolean vector $X_k = [x_1, ..., x_i, ..., x_n]$ where x_i takes value one when there is at least one trip on the day i otherwise, it takes value zero. This simplistic representation has three important advantages. First, it is very straightforward to compute the vector X_k for each card without any enrichment of the data. Second, to be meaningful this representation doesn't require a minimum number of transactions per card. Third, the binary vector allows us to focus on the revealed choice to use public transit on a given day without taking into account the characteristics of this usage that are voluntarily incorporated later on in the investigation.

Having defined a vector space, we need a measure of dissimilarity. When studying public transit usage, it is as important to know on which day passengers do use the system than on which day passengers do not use the system. Thus, in each vector, X_k zero and one carry equivalent information. Two vectors are to be considered close in the vector space when there is mutual presence or mutual absence. The simple matching distance (SMD) is a measure of dissimilarity that has this property, and it can be expressed as follow for two users k and l,

$$D(X_k, X_l) = 1 - \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}} \tag{1}$$

where:

f_{00} = number of days where X_k is 0 and X_l is 0
f_{11} = number of days where X_k is 1 and X_l is 1
f_{01} = number of days where X_k is 0 and X_l is 1
f_{10} = number of days where X_k is 1 and X_l is 0

With the above dissimilarity measure, we can calculate a dissimilarity matrix M. In this matrix, each element M_{kl} corresponds to $D(X_k, X_l)$. This matrice is then used as an input for the clustering algorithm. Hierarchical clustering is a common approach that does not require that we commit to a particular number of clusters [Friedman et al., 2001]. It is very popular because it produces a dendrogram that illustrates how the objects are joined together. After some test, we decided to use an agglomerative approach with the Ward method [Ward Jr, 1963]. This method uses a criterion for choosing the pair of clusters to merge at each step by minimizing the change in the total sum of squares. It can be implemented recursively by a Lance Williams algorithms. As opposed to K-means, this method can be applied to dissimilarities measures that are not strictly Euclidean such as the SMD. This method also tends to produce compact clusters of approximate size.

To quantifies statistically the strength of the association between a given cluster and a given fare profile, we can use the odd ratio (OR) [Goulet-Langlois et al., 2018]. An OR bigger than 1 indicates a positive association and vice versa. For a sample of the population, OR can be estimated as follows:

$$\widehat{OR}_{a,b} = \frac{N_{a,b}.N_{a',b'}}{N_{a',b}.N_{a,b'}} \tag{2}$$

In the above formula, a refer to a single fare profile, b refer to a single cluster, a' refer to the aggregation of all clusters except a, b' refer to the aggregation of all fare profile except b, $N_{a,b}$ denotes the number of individuals with characteristic a and b. The log of the OR is normally distributed and can be used to statistically test whether an OR is significantly different from 1 at a given confidence level [Goulet-Langlois et al., 2018, Morris and Gardner, 1988].

3.2.2 Measuring intrapersonal variability

The previous method does not incorporate any information regarding how each passenger uses the network. Therefore, it does not address the question of how similar are each day for a given individual. From the perspective of transit usage, each day can be described in terms of trip rate but also considering the spatiotemporal characteristic of trips. More precisely, two days can be considered similar if they have the same number of trips and if trips share time and space attributes. Huff and Hanson [1986] have proposed a trip based similarity measure between two days i and j that can measure conjointly those aspects and can be expressed as follows,

$$S_{ij} = [1 - \frac{1}{2} \sum_{k} |P_{ic} - P_{jc}|] \frac{n_i}{n_j}, \quad nj \geq ni \tag{3}$$

where P_{ic} is the proportion of trips in days i that have the characteristic of the equivalence class c and n_i is the number of trips on day i. This measure of similarity ranges from 0 to 1. Two days having the same number of trips and identical trip pattern regarding the equivalent class c will result in a similarity of 1.

To define an equivalent class, there are several options because trips can be described with many attributes such as purpose, distance, mode, time of departure etc. With smart card data, the number of combinations is often reduced because not all attributes are available directly from the transaction database. In Lyon, two attributes are naturally available: transaction time and boarding stop. Since we now have identified the transactions that correspond to the beginning of trips, we can use the spatial and temporal features of those transactions to define the equivalent class. Both features have high cardinality as there are many stops in the network (more than 4000) and the timestamp is known with second precision. To reduce the dimension of those features, we have decided to use two grids that make sense both from a practical and behavioural point of view:

- **Temporal grid**. Trip starting time are grouped into the following time slot : before 7 a.m, 7 a.m to 10 a.m, 10 a.m to 12 p.m, 12 a.m to 2 p.m, 2 p.m to 4 p.m, 4 p.m to 8 p.m and after 8 p.m.
- **Spatial grid**. Trip origin stops are aggregated at the district level in the city of Lyon where the network is denser and at the communal level in the peripheral areas of the urban transit perimeters. This spatial aggregation is made up of 82 zones depicted in figure 1.

The equivalence class c are then built using a contingency table between the spatial zone and the time slot. Two trips are in the same equivalence class if they share both attributes. Each day i is then synthesised in the vector P_{ic} that transcribes the spatiotemporal distribution of trips of day i. Therefore, S_{ij} will be equal to one if, on two distinct days, a card makes exactly the same number of trips from the same spatial zone and in the same time slot. This way, three dimensions of daily trip pattern are considered: space, time and trip rate. Moreover, to infer trip destination in entry only smart card system, it is often assumed that the destination is close to the origin of the next trip (trip chaining model). Hence, considering only the origin of trips may provide a sufficient representation of the daily trip pattern and it does meet the goal of this study.

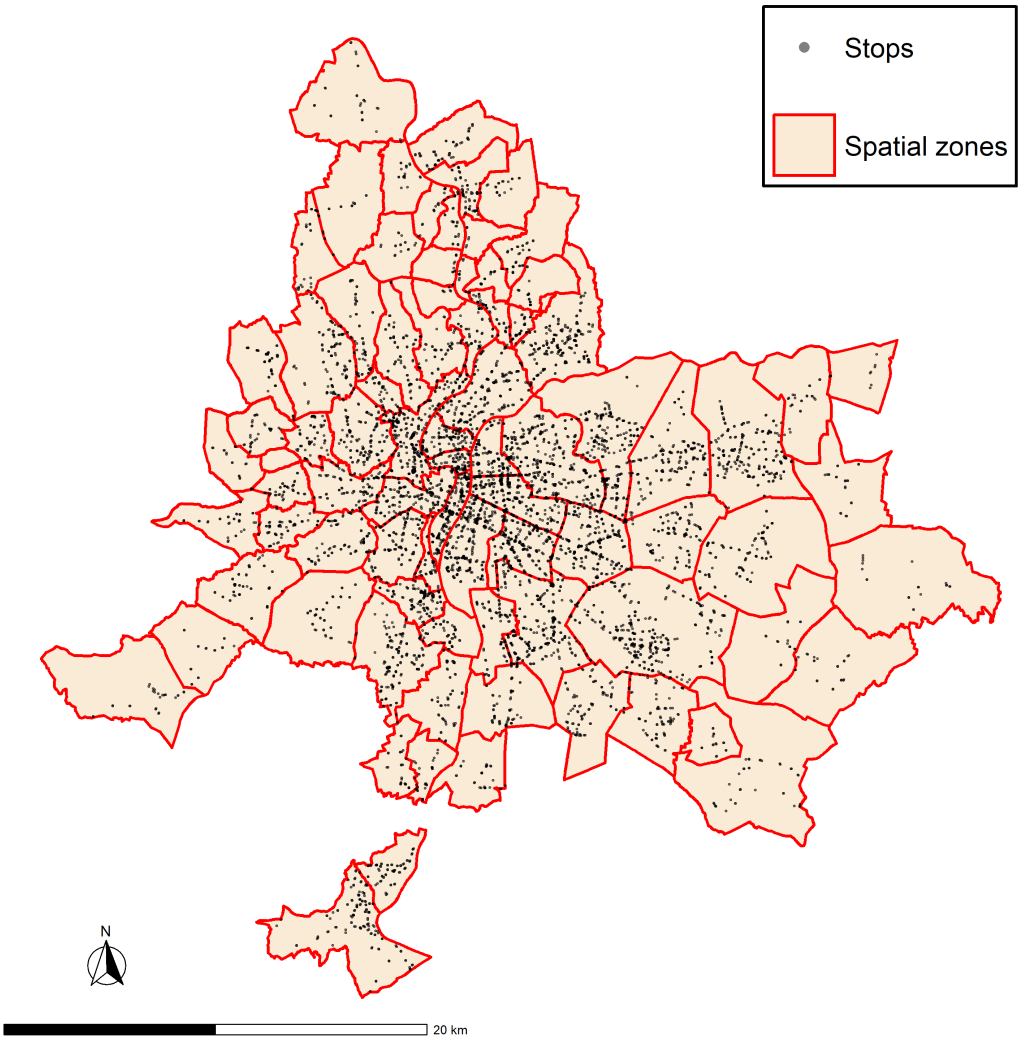

Figure 1: Study area and aggregation of stop into a spatial grid, source: Authors

4 Results

4.1 Data driven sample selection

To start this study with a holistic approach, a random sample of 40,000 cards among the 591,124 cards observed travelling at least once between January and June 2017, was drawn. The clustering method was then applied to a random subset of 10,000 cards to visualize day-to-day usage pattern and select a sample in a data-driven way. The results are represented on a heatmap in figure 2. Each row corresponds to a card, each column to a day and each cell can be either black (at least one trip) or white (no trip). Week numbers are indicated in the x-axis. The resulting dendrogram is shown on the left of the heatmap. Figure 2 demonstrates that even at the most basic level of days of usage, the interpersonal variability is considerable with a large diversity of pattern.

Some rows are entirely white which indicate that some cardholders rarely use the transit system. Weekends generate a strong and repetitive vertical white pattern that affects a large proportion of users. Nonetheless, some rows are almost entirely black i.e some individuals use the transit system almost every day. The two holiday period in weeks 8-9 and weeks 16-17 are also visible and can lead to episodic break of usage. Lastly, figure 2 reveals that some users exhibit clear changes in usage intensity over the six months.

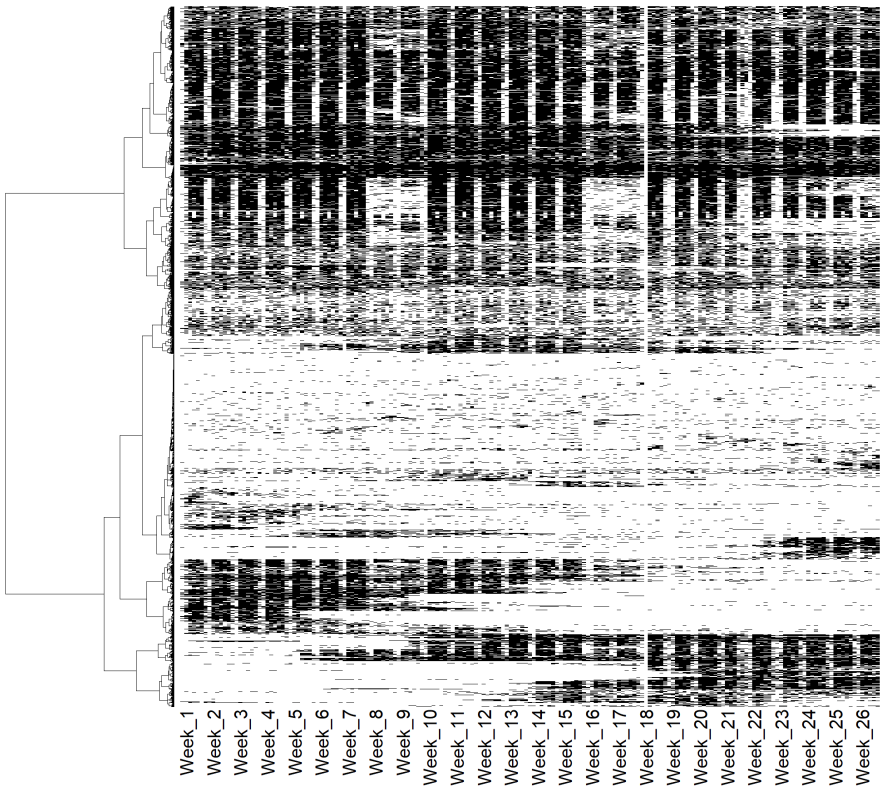

Figure 2: Dendogram resulting from the application of the clustering method to 10,000 randomly selected cards, source: Authors

A simple interpretation of this dendogram could be to classify users in three main groups:

1. **The low frequency users (LF)**, mainly located in the middle of the dendrogram. They almost never use the system on a multimonth scale. Our hypothesis is that public transit is something that is not part of their daily routine. Those people may actually use other transportation modes, or be present in the city only during a short period of time such as tourists visiting the city.

2. **The consistent transit users (CT)**, mainly located at the top of the dendrogram. Those are individuals that used the transit system consistently over the 6 months period. They may be subject to ruptures such as holiday or weekend and may not use the transit system

every day but they will not stop using the system over a long period of time. For those users, we can assume that transit usage was part of their daily routine from the beginning to the end of the study period.

3. **The intermittent transit users (IT)**, mainly located at the bottom of the dendrogram. Those are individuals that present characteristics of low frequency users but also characteristics of consistent transit users. For those users, transit usage was part of their routine but at one point of the study period, they exhibited a marked change in day-to-day usage intensity.

To discriminate between those three groups, two attributes were computed for each card k. N_k being the number of travel days and M_k the maximum number of consecutive days without transit usage. The distribution of both variables is given in figure 3. N_k is spread almost uniformly between 10 and 130 meaning that few cards use the system more than 140 days out of the 181 days of the study period. There is also a concentration of cards around small values of N_k i.e cards that are observed travelling only a few days. The distribution of M_k is characterized by peaks at each multiple of 30 correspondings to usage during only a subset of months. The distribution of M_k also presents a concentration of cards between 0 and 30 meaning that a high proportion of cards will not stop using the transit system for more than 30 consecutive days. To classify users into the three proposed groups, the following rules are implemented: (1) a user is considered as LF if the number of travel days is less or equal to 10 days i.e $N_k \leq 10$, (2) a user is considered as IT if the number of travel days is bigger than 10 but there is a usage interruption of more than 30 continuous days i.e $N_k > 10$ and $M_k \geq 30$, (3) otherwise users are classified as CT.

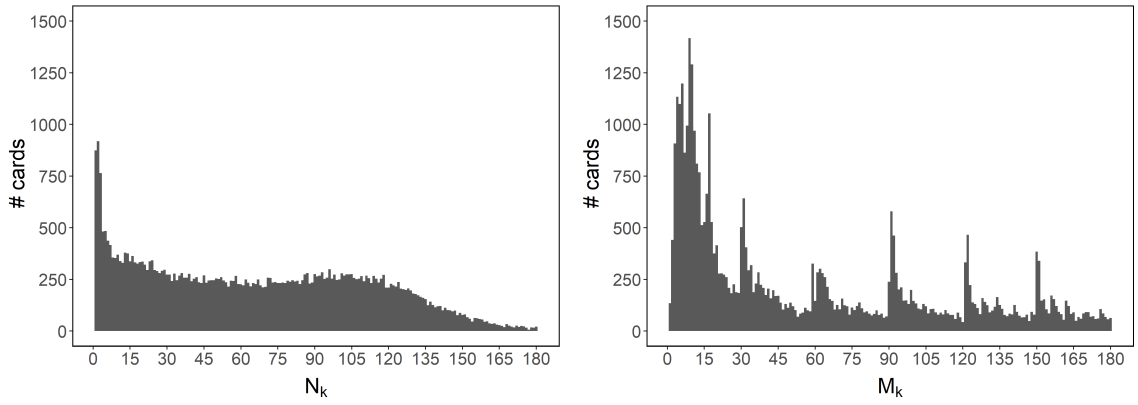

Figure 3: Distribution of N_k and M_k, source: Authors

We have applied those rules to the 40,000 cards of our initial sample. The number of cards and the number of trips by groups are given in table 2. The smallest group corresponds to the LF users with a total of 5456 cards (14% of the cards), but less than 1% of the total number of trips. 16,358 users are classified as IT (41% of the cards) and account for 30% of the trips. The CT users form the biggest group with 18,186 cards which correspond to approximatively 45% of the users and account for almost 70% of the trips.

To start this investigation, we have analysed the daily usage pattern of 40,000 cards selected randomly. In doing so, we were able to show the diversity of usage pattern on a multimonth scale. Based on the proposed clustering method, we have defined three groups of users with distinct

	# users	% users	# trips	% trips
Consistent users (CT)	18,186	45%	4,220,965	69%
Intermittent users (IT)	16,358	41%	1,840,412	30%
Low frequency users (LF)	5,456	14%	50,316	1%

Table 2: Distribution of users and trips by group, source: Authors

multimonth frequencies of usage. The rest of this study will focus on the behaviour of the 18,186 consistent transit users. This sample selection is justified by the fact that they account for the biggest proportion of trips done on the network. It also allows us to maximise the observation period and focus on a group that share common characteristics in terms of multimonth usage routine. While this can be seen as a limit for the rest of this study, the two methods could be applied in the same way to the rest of users (as long as there is more than one day with travel events). The next section of this paper examines day-to-day regularity and intrapersonal variability at an aggregate level.

4.2 Aggregated analysis of intrapersonal variability

As a first step, we may be interested in evaluating the regularity in an aggregate way for each chosen dimension of variability. For example, considering the spatial dimension, for each user, we can rank the spatial zone based on the number of trips they generate and calculate for all users the average share by rank. The same can be done with the time slot, combining both dimensions (i.e using the concept of equivalent class) and also for the daily trip number. The results of those calculations are given in figure 4.

In average, the two most important spatial zones generate 76% of the trips and the two most important time slots generate 67% of the trips. This indicates that overall there is a high degree of spatial and temporal regularity which is in line with previous work [Schlich and Axhausen, 2003, Hanson and Huff, 1988, Morency et al., 2007]. When combining spatial and temporal dimension, the most important equivalent class generates on average 27% of the trips which is close to the 30% obtained by Huff and Hanson [1986] for an equivalent class defined with time slots and city quadrant. The five most important equivalent class generate approximatively 65% of the trips. As indicated in figure 4, the concentration of trips in a few equivalent class is more important during workings days. However, the difference between each curve remains thin. For instance, on working days the two most important equivalent class will generate 49% of the trips but this number only decreases to 46% when considering all days. The bottom left plot indicates that on average, 55% of the days the daily number of trips will correspond to the most frequent trip rate. In other words, the most recurrent trip rate cover on average a bit more than half of the travel days. Again, there is no important difference when we focus on working days. This first analysis demonstrates that on average trips are repetitive in the sense that they share spatial and temporal characteristics, and that days are repetitive in the sense that on average users will make the same number of trips. However, it does not mean that all days are similar for each user, and that all users have the same level of variability. To measure those two aspects, S was calculated in a totally desegregate way i.e for each individual and each pair of travel days.

A first aggregation of S_{ij} could be to calculate for all users the mean similarity between any two days of the week. The results are given in table 3 and give rise to the following comments. First, we confirm the finding of Schlich and Axhausen [2003] that weekend days are less similar than other days of the week. The mean similarity within Saturday and within Sunday is equal to 0.18. The similarity of Saturday with other weekdays decrease to 0.11 and the similarity of Sunday

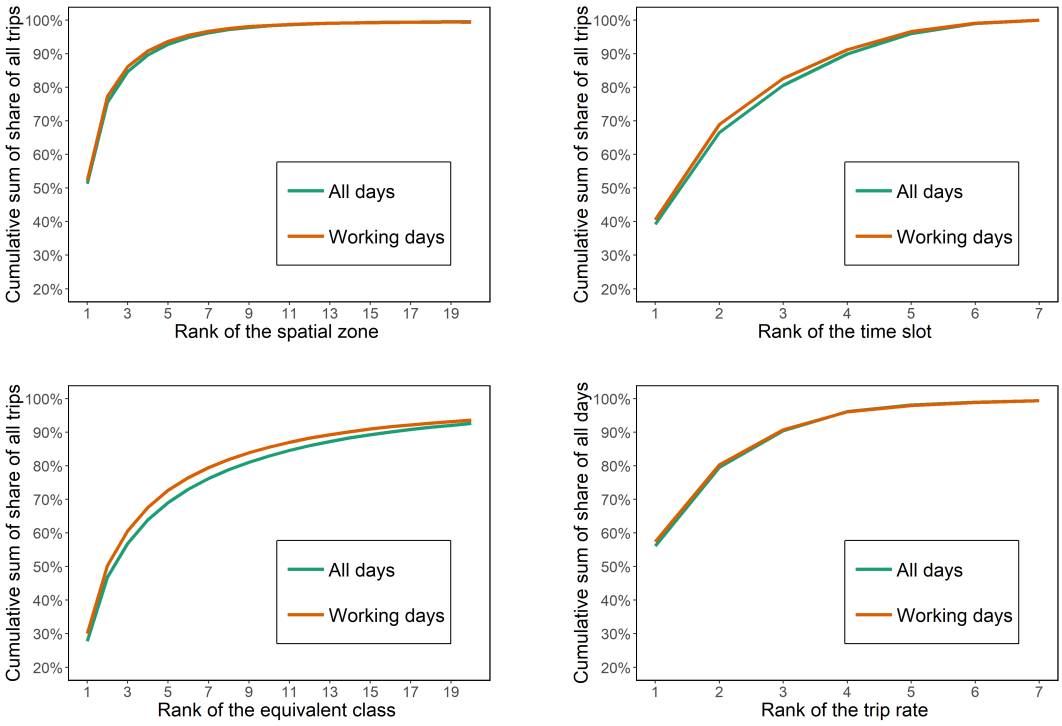

Figure 4: Regularity of transit usage on each chosen dimension of variability, source: Authors

with other weekdays decrease to 0.08. Second, even if weekdays are more similar within each other than with weekend days, there are more similarity within the same weekdays than between distinct weekdays. For instance, the similarity within Monday is equal to 0.33 but decrease to 0.27 when we compare Monday with Friday. Third, as found by Schlich and Axhausen [2003], Friday is the weekday that exhibits less similarity with the rest of the weekdays. Fourth, Tuesday is the weekday where the within-day similarity is the highest with a value of 0.36. This indicates that there is less intrapersonal variability and thus users have a higher tendency to repeat the same trip pattern every Tuesday than during other days of the week.

	Mon.	**Tue.**	**Wed.**	**Thu.**	**Fri.**	**Sat.**	**Sun.**
Mon.	0.33	0.32	0.28	0.29	0.27	0.10	0.08
Tue.		0.36	0.29	0.31	0.28	0.11	0.08
Wed.			0.33	0.27	0.25	0.11	0.08
Thu.				0.34	0.28	0.11	0.08
Fri.					0.31	0.11	0.08
Sat.						0.18	0.10
Sun.							0.18

Table 3: Mean similarity between days of the week, source: Authors

A second analysis will be to plot the distribution of similarity among users. For each user, we

compute the mean similarity for all pair of days $\overline{S}$ and the mean similarity focusing only on pair of workings days $\overline{S_{wd}}$. The distribution among users of both variables is given in figure 5. The median of $\overline{S}$ is equal to 0.18 and only increase to 0.22 for $\overline{S_{wd}}$ meaning that most users will have a high degree of day-to-day variability even when we focus only on working days. Both distributions are also very skewed toward higher values of similarity. In other words, there are large differences between users and some users exhibit lower levels of day-to-day variability than others. Another way to look at this problem could be to determine for each user and each travel day, the number of other travel days with the same daily trip pattern ($S_{ij} = 1$). We found that on average 62% of the daily trip pattern will reoccur at least once in the study period. We also looked for each user at the number of days where the daily trip pattern was completely unique (for i, $S_{ij} = 0 \ \forall \ j$). We found that more than 96% of the cards will not have one completely unique daily pattern. In other words, while there may be in average around 40% of user-day pattern that will not reoccur in the longitudinal records, there are very few users that have a travel day that does not share any attributes with other travel days in the longitudinal record.

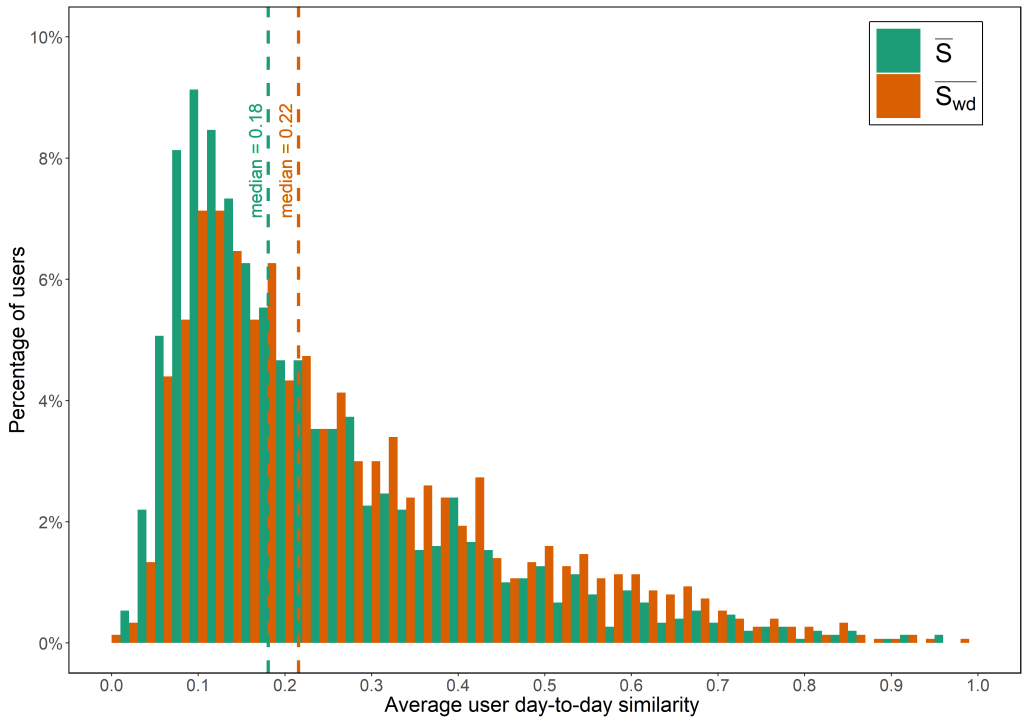

Figure 5: Distribution of average users day-to-day similarity ($\overline{S}$: for all pair of days, $\overline{S_{wd}}$: only for working days), source: Authors

Those aggregated results support the idea of Hanson and Huff [1988] that there is a high degree of regularity in transit pattern, but there is also systematic day-to-day variability. Thus, to characterize transit usage a single day is insufficient. Moreover, if we aggregate trips characteristics over periods longer than a day, we will eclipse the daily variability of transit usage. Finally, as indicated by the large skew in the distribution of $\overline{S}$, there are reasons to believe that some users are less variable than others. The next section shows how the combination of intrapersonal variability measurement, interpersonal clustering and fare profile can offer deeper insights into day-to-day

transit usage variability.

4.3 Combining clustering, intrapersonal variability and fare profile

The clustering method was applied to the 18,186 consistent transit users. With the help of the dendrogram, we have decided to retain 6 clusters (numbered from C1 to C6) which we believe is a good balance between the quality of the clusters and the interpretativeness of the results. To visualize the pattern of each cluster, 100 cards were selected randomly in each cluster and plot according to the convention of figure 2. The graphical results are given in figure 6. To further help the interpretation, descriptive statistics are given for each cluster and each dimension of variability in table 4.

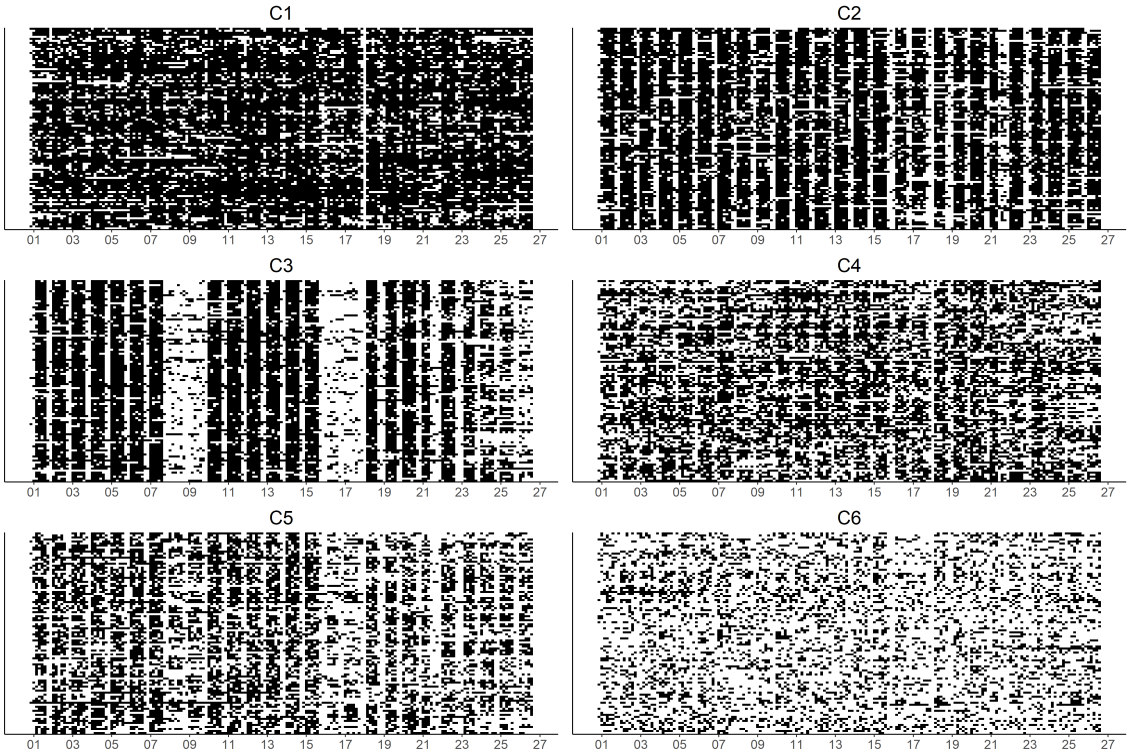

Figure 6: Visualization of the day-to-day usage pattern of 100 random users selected from each cluster, source: Authors

In table 4, it can be seen that cluster 1 is characterized by the highest percentage of days of usage and almost no calendar structure as the usage rate remain quite high on holidays and weekends. Cards in cluster 1 will in average use the transit system more than one Sunday out of two. On the day they use the system, they tend to make more trips as indicated by the mean number of trips per day (2.7), but also by a high proportion of user-day with more than three trips (24%). Cluster 2 is the biggest cluster in terms of size with 4592 cards (25% of the CT users). It is characterized by a very high usage rate on workings days, few travel days during the weekend, a large proportion of travel days with two trips and a high percentage of trips in the morning and evening peak period (31% and 35%). Like cluster 2, cluster 3 exhibits a concentration of trips in the two peak period but their day-to-day usage pattern differs. In fact, in figure 6, it is easy to see that

	C1	C2	C3	C4	C5	C6
# users	3,211	4,592	2,442	2,624	2,235	3,082
% users	18	25	13	14	12	17
Day of usage						
Working days (%)	86	87	84	64	62	31
Holyday days (%)	75	63	20	51	33	23
Saturday (%)	72	31	22	51	18	23
Sunday (%)	56	13	10	28	9	12
Trip rate						
Mean trip per travel day	2.7	2.3	2.2	2.2	2.0	1.8
One trip (% user-day)	16	16	25	27	29	39
Two trips (% user-day)	41	59	52	47	52	47
Three trips (% user-day)	19	14	13	14	12	10
Four trips or more (% user-day)	24	14	14	13	10	7
Trip temporal distribution (%)						
before 7 a.m	4	4	2	3	3	2
7a.m to 10 a.m	19	31	32	19	28	17
10 a.m to 12 p.m	11	7	7	12	8	13
12 a.m to 2 p.m	13	11	15	13	12	14
2 p.m to 4 p.m	12	8	11	13	10	16
4 p.m to 8 p.m	31	35	30	32	34	32
after 8 p.m	10	5	3	9	5	7
Mean spatial indicators						
# distinct spatial zone	13.7	10.4	9.5	11.6	9.3	8.7
% of trips in the two most frequent spatial zone	71	79	81	72	78	73
Mean users similarity measurement						
All days	0.16	0.34	0.25	0.17	0.26	0.16
Working days only	0.19	0.38	0.30	0.20	0.29	0.18

Table 4: Descriptive statistics for each cluster, source: Authors

users in cluster 3 almost do not use the transit system during the holidays period. On this same figure, it is also possible to notice a decrease in usage at the end of June just before the summer period. Cluster 4 does not exhibit such a clear calendar structure. Cards in this cluster use the transit system on working days a bit more than 3 days out of 5 (64%) and remain largely present during the holiday (51%), Saturday (51%) and Sunday (30%). The temporal trips distribution of cluster 4 is somehow related to the one of cluster 1, with no pronounced concentration of trips in the morning peak period. Cluster 5 can be seen as an intermediate between cluster 3 and 4. As in cluster 3, transit usage is impacted by holiday and weekend but as in cluster 4, the usage rate on working days is under 65%. The last cluster (C6), is formed by vectors that are very sparse i.e with lots of zero. In this cluster, we found users that despite the fact that they consistently use the system over the 6 months, their usage rate is less than 30% and doesn't vary much within the type of day. Cluster 6 is also characterized by a lower trip rate and a very high percentage of one trip day (39%). Finally, table 4 indicates that the spatial diversity of trips vary between clusters. The concentration of trips inside the two most frequent zone is higher for cluster 2, 3, and 5.

To extend this analysis, the contingency table between fare profile and clusters is computed. The results are given in table 5 with the corresponding odd ratio. The distribution of mean user day-to-day similarity on working days ($\overline{S_{wd}}$) for each cell of the contingency table is also given as a boxplot in figure 7.

		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Total**
Elderly	# users	317	135	11	375	130	549	1,517
	OR	**1.26**	0.27	0.04	**2.11**	0.65	**3.17**	
General public	# users	1,175	2,747	235	700	684	518	6,059
	OR	**1.19**	**4.62**	0.18	0.69	0.87	0.35	
Intermodality	# users	28	370	74	46	205	108	831
	OR	0.16	**2.5**	0.62	0.34	**2.47**	0.72	
Other	# users	59	77	25	90	60	199	510
	OR	0.6	0.52	0.33	**1.28**	0.95	**3.28**	
Short duration	# users	2	11	2	30	39	366	450
	OR	0.02	0.07	0.03	0.42	0.67	**24.1**	
Social	# users	815	315	63	425	149	365	2,132
	OR	**3.53**	0.48	0.18	**1.57**	0.5	1.01	
Student	# users	616	609	326	618	518	299	2,986
	OR	**1.26**	0.72	0.76	**1.72**	**1.65**	0.5	
Young	# users	199	328	1,706	340	450	678	3,701
	OR	0.22	0.23	**15.97**	0.54	0.98	1.13	
Total		3,211	4,592	2,442	2,624	2,235	3,082	18,186

Table 5: Contingency table between clusters and fare profiles and associated Odd Ratio, bold indicate superior to 1 and statistically different from 1 at 99% confidence level, source: Authors

General public fare profile is mainly aimed at people in employment and thus work-related trip will probably shape their transit usage. Table 2 indicates that general public users are strongly associated with cluster 2 (OR of 4,62) and to a lesser extent with cluster 1 (OR of 1.19). Cluster 2 exhibits a usage pattern that seems to be more work-oriented than cluster 1 where users probably make more diverse use of the network. In figure 7 it can be seen that the median of $\overline{S_{wd}}$ for general users in cluster 2 is 0.43, but it decreases to 0.21 for general users in cluster 1 which confirms that general public users from cluster 1 tend to be more variable than those of cluster 2. Holders of

intermodality pass combined public transit with other services such as trains which can constrain their usage patterns. Table 2 indicates that they are mostly found in cluster 2 and 5 where they tend to show a high level of day-to-day similarity. 13% of intermodality profile are also assigned to cluster 6 where their usage of the transit system is on average more variable with a median of $\overline{S_{wd}}$ equal to 0.19 compared to 0.49 and 0.43 for intermodality users in cluster C2 and C5. As anticipated, young people dominate cluster 3 and rarely use the transit system during holidays. Young users in cluster 3 present a mean similarity between working days that is relatively high with a median equal to 0.28. They are also found in cluster 6 where their usage is less intense and with lower day-to-day intrapersonal similarity. People that are using social pass, exhibit a high level of day-to-day variability in their transit usage. Table 2 indicates that there is a positive and significant association between cluster 1 and social users (OR of 3.53). As pointed before, cluster 1 presents the highest intensity of usage both in terms of number of travel days and number of trips per day. Thus, users in cluster 1, may cover an important proportion of their urban travel needs with public transit. Table 2 also indicates that a non-marginal part of social users is assigned to cluster 4 and 6, so we can not consider all social users as very intense users. Like social users, elderly users are mostly found in clusters that do not exhibit clear calendar structure such as cluster 1, 4 and 6. In figure 7, it can be seen that the median of $\overline{S_{wd}}$ for elderly users in those three clusters is between 0.14 and 0.17, almost three times less than general public users of cluster 2. Student users is an interesting population because university constraints are very heterogeneous both spatially and temporally. Table 2 shows that this population is almost equally spread within the six clusters. The level of day-to-day similarity of students is in general low, but it can vary between clusters. For instance, students in cluster 2 and 3 have a median similarity of respectively 0.22 and 0.2 compared to median similarity below 0.15 in other clusters (see figure 7). Finally, as expected, short duration users are almost all assigned to cluster 6 where they exhibit a high level of variability with a median of $\overline{S_{wd}}$ equal to 0.13. Those are users who use the network from time to time with varying spatiotemporal patterns.

Our empirical results demonstrate that we can find tangible links between the intrapersonal day-to-day variability, the multimonth pattern of usage synthesise by cluster membership and sociodemographic inferred from fare profile. Those three elements are essential to carry such a fine-grained investigation of day-to-day transit usage behaviour. They can extend our knowledge of transit usage variability and are useful to define interpretable user segmentation. This is what we have done in this section. The following section takes a step back from the strict description of numerical results to discuss the lessons to be learned from this investigation.

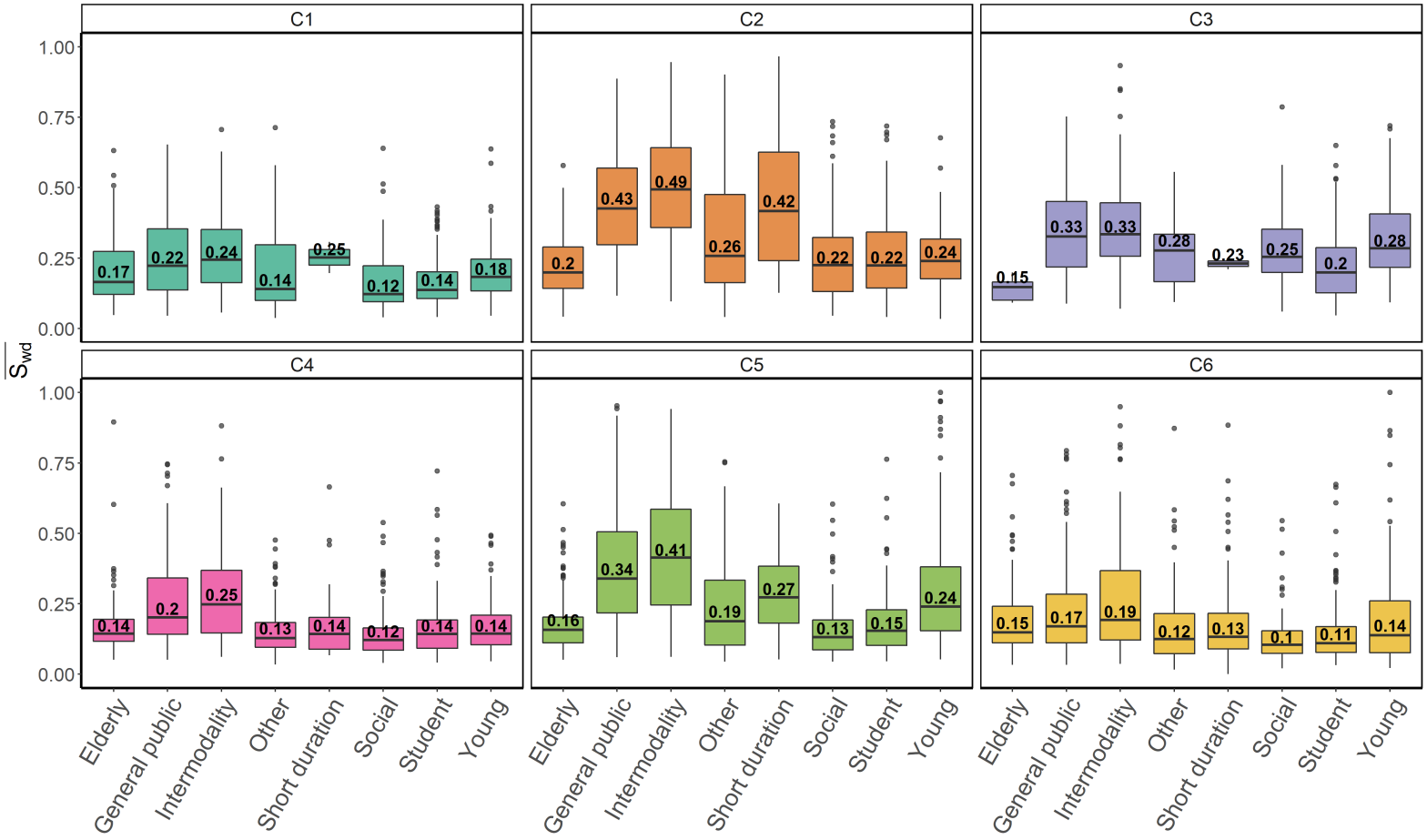

Figure 7: Boxplot of users working days mean similarity ($\overline{S_{wd}}$) by cluster and fare profile, source: Authors

5 Discussion and conclusions

Researcher and transportation planners have a strong interest in understanding the day-to-day variability of transit users. Smart card data give us the opportunity to undertake this analysis over longer periods and provide a deeper understanding than the current state of the art [Manley et al., 2018]. As opposed to studies based on travel surveys, most of the research on smart card data have relied on clustering techniques to analyse variability. In this paper, we complement this approach with a traditional day-to-day similarity measure [Hanson and Huff, 1988] and prove that the combination of both technics is a valuable tool to mine smart card data. It can be used to expand traditional research on travel behaviour variability using similar paradigm and concepts.

Our empirical finding suggests that there is no "one size fits all" approach to the problem of day-to-day variability of transit usage. The simple view of transit users as solely commuting passengers from Monday to Friday is inadequate. Likewise, the classification of transit users solely based on their fare product or one-day data is incomplete. Very distinct level of intrapersonal variability can be found within each fare profile and within each usage pattern synthesised with cluster membership.

By selecting randomly our initial sample among all cards, we were able to visualize the diversity of day-to-day transit usage pattern and we have defined three main groups of users. The high proportion of intermittent transit users indicate that there is porosity between high frequency usage and low frequency usage. In other words, it is not uncommon that public transit usage habits and routines change drastically over time. This confirms the old idea that the apparent stability at the aggregate level is in reality compensated by changes at the individual level [Jones and Clarke, 1988]. Although it is not possible to understand the leading cause of ruptures using only smart card data, we believe that public transit operators should show more interest in those users and those changes if they want to influence travel behaviour and increase loyalty. This requires that they move from a purely accounting approach such as the number of pass sold per month to a customer-centric approach where each individual pattern is mined. This approach can offer a large range of new opportunities. For instance, operators could track individual pattern to identify "unsuccessful" new users, to design and evaluate new fare structure, to define more targeted marketing actions or to detect commercial opportunities.

Our results confirm that over long periods, users show spatial and temporal repeatability in their trip pattern but there is also a systematic intrapersonal variability in day-to-day transit usage [Schlich and Axhausen, 2003, Hanson and Huff, 1988]. In large urban areas such as the metropolis of Lyon, people have access to a wide variety of activities. They can move with several modes such as walking or cycling. They are not all constrained by scheduled and fixed activities such as work or school (e.g social profile or retired people). Moreover, as noted by [Schlich and Axhausen, 2003] there is a trending decline in general constraint but also greater flexibility over working hours and places of work [Manley et al., 2018, Goulet-Langlois et al., 2016]. As a result, it is not surprising to observe important intrapersonal day-to-day variability. What is more interesting is that it is possible to correlate this intrapersonal variability to interpersonal variability using cluster analysis. In doing so, we found that there are important differences between clusters. In cluster 1, users probably cover most of their travel needs with public transit and tend to be more variable in their usage. In clusters with calendar structure such as cluster 2, 3, and 5, it can be assumed that the transit usage is driven by work or school trips which can explain the lower intrapersonal variability. Finally, in cluster 4 and 6, public transit usage is erratic and probably complementary to other modes of travel leading to higher levels of intrapersonal variability.

The analysis conducted in this paper also shows that the conventional distinction between working day, weekend and holiday may be relevant for some users but not for all. Even if on average, there are more stability in behaviour during weekdays, we have seen that the intrapersonal

variability decrease only marginally when computed solely on working days. Similarly, we have observed that holiday periods have an influence only on a reduced proportion of users. This clearly shows that the traditional paradigm of transit planning focused on a set of typical days such as working days, Saturday or Sunday is not perfectly valid from the individual point of view. Thus, a more disaggregated understanding of which days individuals may or may not use the transit system is an important task. With our interpersonal clustering method, it is possible to do so. In fact, this method is designed to find homogeneous groups of passengers based only on the day they used public transit. The resulting clusters can then be used by transit planners and marketers, to inform or promote specific usage in a targeted manner. The combination of interpersonal clustering and pre/post intrapersonal similarity measurement could also assist in identifying groups of users that are more prone to change their travel behaviour after the introduction of new services or in case of specific disruption (e.g on week-end services).

This case study also demonstrates that cross-checking results with socio-demographic information derive from the fare profile, add a lot of value to the investigation. This is because socio-demographic attributes are important factors affecting intrapersonal variability [Pas and Koppelman, 1987, Susilo and Axhausen, 2014]. Unfortunately, smart card data and more generally passive data are often very incomplete on those aspects which can generate ambiguous explanation [Manley et al., 2018]. Moreover, there are increasing privacy concerns with smart card data that could jeopardise the full valuation of these data. In Lyon, the card is, for now, individual and the unique id number that identifies each card is changed every 12 months. In other cities, cards can be shared between passengers, cards id are changed more regularly and sometimes passengers have the option to pay with contactless bank cards. In those cases, it is not possible to ensure individual traceability, to perform long term analysis or to determine profile with fare product. These are additional challenges that must be addressed otherwise the potential of these data to conduct longitudinal analysis will be strongly limited.

Research around individual day-to-day transit usage variability is an important area of investigation that has several practical implications. Automatic systems such as fare collection collect lots of data about lots of users over long periods of time and thus can be very useful to analyse the variability of transit usage. To make the best of those data, we need to develop new methodologies but also adapt the existing ones. This is what we have done in this paper, and while the results are specific to the city of Lyon, the two methods can easily be applied to other smart card datasets. It would especially be interesting to replicate this analysis in cities of distinct size, and distinct country to better understand the link between variability, city structure, and cultural context. Another direction of future research would be to further investigate the day-to-day variability of inconsistent users, and to try to understand the motivations behind the changes of transit usage intensity using targeted surveys. There are also some limits to the research presented in this paper that require additional analyses. First, representing a trip as a departure in a given time slot from a given zone is straightforward and easy to conceptualize but it is an oversimplification of the true characteristics of trips. More detail representation of trip patterns should be investigated and other measures of day-to-day similarity must be developed. Second, the calculation of a large dissimilarity matrix can be computationally expensive. More research is therefore needed to adapt this method to large datasets so it can be deployed in real-world applications. Third, our analysis is only based on smart card data but other data such as weather data or land-use data could provide further insight into the causes of variability.

References

B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, 39(3):399–404, 2006.

M. Bagchi and P. R. White. The potential of public transport smart card data. *Transport Policy*, 12(5):464–474, 2005.

A. Bhaskar, E. Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3):1537–1548, 2015.

P. Bonnel. *Prévision de la demande de transport.* Presses de l'École Nationale des Ponts et Chaussées, Paris, 425p, 2002.

A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79:274–289, 2017.

C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 68:285–299, 2016.

E. Deschaintres, C. Morency, and M. Trépanier. Analyzing transit user behavior with 51 weeks of smart card data. *Transportation Research Record*, page 0361198119834917, 2019.

F. Devillaine, M. Munizaga, and M. Trépanier. Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, (2276):48–55, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

T. Gärling and K. W. Axhausen. Introduction: Habitual travel choice. *Transportation*, 30(1):1–11, 2003.

G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.

G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao. Measuring regularity of individual travel patterns. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1583–1592, 2018.

S. Hanson and J. Huff. Classification issues in the analysis of complex travel behavior. *Transportation*, 13(3):271–293, 1986.

S. Hanson and J. O. Huff. Assessing day-to-day variability in complex travel patterns. *Transportation Research Record*, 891:18–24, 1981.

S. Hanson and O. J. Huff. Systematic variability in repetitious travel. *Transportation*, 15(1-2): 111–135, 1988.

J. O. Huff and S. Hanson. Repetition and variability in urban travel. *Geographical Analysis*, 18 (2):97–114, 1986.

P. Jones and M. Clarke. The significance and measurement of variability in travel behaviour. *Transportation*, 15(1-2):65–87, 1988.

X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

E. Manley, C. Zhong, and M. Batty. Spatiotemporal variation in travel regularity through transit user profiling. *Transportation*, 45(3):703–732, 2018.

C. Morency, M. Trépanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

J. A. Morris and M. J. Gardner. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal (Clinical Research Edition)*, 296(6632): 1313–1316, 1988.

M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva. Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, 44:70–79, 2014.

E. Pas. Multiday samples, parameter estimation precision, and data collection costs for least squares regression trip-generation models. *Environment and Planning A*, 18(1):73–87, 1986.

E. I. Pas. Intrapersonal variability and model goodness-of-fit. *Transportation Research Part A: General*, 21(6):431–438, 1987.

E. I. Pas and F. S. Koppelman. An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, 14(1):3–20, 1987.

M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

C. Raux, T.-Y. Ma, and E. Cornelis. Variability in daily activity-travel patterns: the case of a one-week travel diary. *European transport research review*, 8(4):26, 2016.

R. Schlich and K. W. Axhausen. Habitual travel behaviour: evidence from a six-week travel diary. *Transportation*, 30(1):13–36, 2003.

Y. O. Susilo and K. W. Axhausen. Repetitions in individual daily activity–travel–location patterns: a study using the herfindahl–hirschman index. *Transportation*, 41(5):995–1011, 2014.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Z. Zhao, H. N. Koutsopoulos, and J. Zhao. Individual mobility prediction using transit smart card data. *Transportation research part C: emerging technologies*, 89:19–34, 2018.