Benefits of Cellular Telecommunication and Smart Card Data for Travel Behaviour Analysis

Discussion Paper

186

Roundtable

Patrick Bonnel
ENTPE, LAET
Lyon

Benefits of Cellular Telecommunication and Smart Card Data for Travel Behaviour Analysis

Discussion Paper

186

Roundtable

Patrick Bonnel

ENTPE, LAET
Lyon

The International Transport Forum

The International Transport Forum is an intergovernmental organisation with 62 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes. The ITF is politically autonomous and administratively integrated with the OECD.

The ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

The ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society. The ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

The Members of the Forum are: Albania, Armenia, Argentina, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China (People's Republic of), Croatia, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Mongolia, Montenegro, Morocco, the Netherlands, New Zealand, North Macedonia, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Tunisia, Turkey, Ukraine, the United Arab Emirates, the United Kingdom, the United States and Uzbekistan.

ITF Discussion Papers

ITF Discussion Papers make economic research, commissioned or carried out in-house at ITF, available to researchers and practitioners. They describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the ITF works. Any findings, interpretations and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the International Transport Forum or the OECD. Neither the OECD, ITF nor the authors guarantee the accuracy of any data or other information contained in this publication and accept no responsibility whatsoever for any consequence of their use. This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Comments on Discussion Papers are welcome.

Acknowledgements

Table of contents

Figures

Tables

Introduction

Research on human mobility traditionally relies on household travel surveys that are often limited to recording trips on a single day per household (Wolf et al., 2003; Stopher and Greaves, 2007). Such surveys offer cross-sectional snapshots of the population's travel behaviour. However, they do not support the fine-grained temporal analysis of the hourly, daily or weekly variability of individual travel behaviour. Understanding the dynamics of human mobility patterns is essential for transport planning and demand management (Gundlegård et al., 2016). Very few surveys have collected data over more than a week and yet such research provided evidence of significant and systematic intrapersonal variability (Huff and Hanson, 1986; Hanson and Huff, 1981, 1988; Schlich and Axhausen, 2003).

Nevertheless, the high cost of travel surveys limits their frequency which prevents the monitoring of mobility dynamics in a large sample. The use of some big data like mobile telecommunication data or public transport smart card data represent a potential to completely change the current techniques to estimate behavioural transport models. The ubiquity of the data as well as the relatively cheap deployment makes it possible to conduct studies on mobility trends. Hence, dynamic origin-destination trip matrices can be generated using methods for assigning trips into target time windows. A number of studies have been conducted to extract dynamic trip metrics using different forms of mobile network or smart card data.

Most studies based on mobile network data have explored customer billing data known as Call Detail Record (CDR) and estimated the temporal distribution of user trips in limited geographical areas. However, these methods perform rather poorly, especially in urban zones (Wang et al., 2013; Zhao et al., 2016). Only a few pieces of research have validated the results against external data sources.

Research presented in this paper feeds from richer mobile network data, not limited to CDR, that includes other mobile network operating data. This paper presents a data processing technique that seeks to enhance the quality of trip matrix estimation. It compares results with what one generally considers as "ground truth data". Lastly, it explores the distribution of trips by time of day.

Smart card data present a number of opportunities and challenges for the analysis of public transport use. Most of the fare collection systems only collect tap-in data, leaving it to analysts to infer trip destination. Fraud and non-validation might also affect data quality. This paper proposes a data processing method using an automated passenger counting system in order to infer data to the whole public transport trips universe. The paper then compares mobility indicators elaborated from smart card data with "ground truth data".

The paper is organised in two sections, starting with the analysis of mobile telecommunication network data, followed by that of smart card data. Each section contains details of the data processing techniques, a comparison with ground truth data, and examples of dynamic analysis. The conclusion summarises empirical findings and offers future directions of research.

Mobility analysis with mobile network data

A big quantity of data is not necessarily representative of travel behaviour in a population. Mobile telecommunication network data contains timestamped geolocation information, which one needs to process and transform into trip information. Whilst the accuracy of this process often correlates with the quantity of observation points per mobile device, this quantity is not evenly spread over the population: this might bias trips imputation. Even if a big quantity of data is available, it does not represent the whole population. Data therefore need to be expanded to the whole population. Lastly data is not error free. This section starts with data processing methods that extract and expand origin-destination trip matrices. It continues with a comparison of the result against a "ground truth" and ends with the analysis of trip volumes by time of day.

Cellular signalling data and its processing methods

Mobile telecommunication operators continuously collect mobile network use for billing and operational purposes. Among mobile network technologies, this paper focusses on the traditional GSM (Global System for Mobile communications) network, which provides 2G services, and the UMTS (Universal Mobile Telecommunication System) network for 3G ones. Both GSM and UMTS networks have different infrastructures, but they still work with the same coverage concept. Each antenna covers a zone called a cell. A group of cells together make up a Location Area (LA). Mobile telecommunication companies use LAs to optimise operations.

This paper explores a dataset that consists of 2G and 3G signalling records corresponding to over two million anonymous mobile phone users in June 2017 using the mobile network services provided by a company called Orange. Due to data protection laws, only one day worth of data is used. It however includes about 300 million events. This dataset covers the entire Rhône-Alpes region in France, an area populated by about 5.6 million inhabitants. Orange operates about 2 230 cell towers in the region, and each cell tower may handle several antennas.

The data consists of cellular signalling traffic transiting through the 2G and 3G networks. It includes all the events that are generated by mobile devices or by the network itself (Smoreda et al., 2013). 3G traffic captures more events than 2G traffic as a result of the extra internet services it is able to monitor. Each record in the dataset corresponds to one event and includes the anonymised user ID, the event type, the coordinates of the antenna serving the mobile phone and the assigned timestamp. Different types of cellular signalling events include:

a. telecommunication events (i.e., calls and SMS);

b. data/Internet connections;

c. itinerancy events: handovers (i.e., cell changes during a call) and Location Area Updates (LAU) that occur when crossing a LA boundary or when in idle mode which generates at least an event every three hours);

d. events called attachment and detachment that typically occur when devices are switched on and off.

The above-mentioned set of event types is larger than that found in CDR (typically limited to items a and b.). This study therefore benefits from having more data points and a higher temporal granularity. This approach better captures the mobility of people who are not using their mobile phone intensively.

Below is a short outline of the data processing method, with more details available in Fekih et al. (2020a) and in Bonnel et al. (2015, 2018). One starts with a set of rules to retain only those devices that are used by human beings and remain switched on over the whole day. In other words, the filtering process removes devices part of the Internet of Things, as well as devices that are turned off or outside of the region. Devices are included in the analysis only if the following three conditions are met.

- **Maximum Inter-event Time (MIT) ≤ 180 minutes**. This condition ensures the continuous presence of the mobile phone in the region, and ensures the device remains switched on. The mobile network system operates in such a way that a mobile phone that remains inactive during three hours (180 minutes) generates a periodic event LAU to maintain its location updated. Each user should have at least eight events (also called observations) if their mobile phone remains switched on during the 24-hour period.

- **Entropy (H) ≤ 0.9**. This condition seeks to exclude Machine-to-Machine communications between objects equipped with SIM cards. Devices with an entropy value higher than 0.9 are excluded as they generate events whose volume and uniformity are likely generated by machines. The threshold of 0.9 is set to match the 99th percentile in the distribution of entropy values across the devices. (Note: The entropy measures the randomness of a system or on the opposite its regularity. High measure of entropy corresponds to very regular signals like those generated by machine-to-machine communications or IOT [Internet of Things]).

- **Number of observations (NO) ≥ 4**. This condition ensures that one can detect at least one trip. A necessary condition to detect a trip is to detect two stationary locations, interpreted as two activities. One needs at least two consecutive observations in the same zone to identify a stationary location. No trip can thus be detected with less than 4 observations.

The combination of these conditions excludes some devices. After the filtering process, a large sample of 985 483 users remains available for the analysis. This represents around 50% of observed users in the original dataset. This filtering process classifies as "soft" in comparison with other research, in order to retain a representative sample of the population.

Table 1. Average mobile phone activity indicator values for filtered devices

Indicator	Average value
Maximum Inter-event Time [7am-10pm] (MIT) in minutes	143
Entropy (H)	0.67
Number of observations (NO)	85
Average Inter-event Time (AIT) in minutes	54

Source: Orange data from the Rhône-Alpes region extracted in 2017.

In order to expand the data to the whole population, one has to infer the home location of the person carrying each device. The provision of a limited 24-hour dataset for privacy reason makes this process particularly challenging. Home detection consists in four consecutive steps:

1. Select events occurring at night (from 3 a.m. to 7 a.m. and from 10 p.m. to 3 a.m.);

2. Discard events that only can occur when a device is in motion, such as handover from one cell to another;

3. For each device in the resulting selection, extract the identifiers of all antennas to which the device was connected;

4. For each device, consider the most frequently observed antenna, assign it to the corresponding sector and consider it as the home location of the person using the device.

Home detection is then used to expand data to the whole population of the area using census data for each sector. Expansion factors range from 3.5 to 10 with a median value of 5.39.

One defines a trip as a movement between two activities. This requires making some assumptions on what can be interpreted as an activity. This paper proposes the use of a time threshold: if an individual is present for a duration at least equal to this time threshold in a sector, one will assume that this individual performed a stationary activity there. Of course, the number of stationary activities is sensitive to the choice of the time threshold. Fekih et al. (2020a) tested different time thresholds: 30-minutes appears to be the most relevant choice when looking at trips within the Rhône-Alpes region.

Validation against "ground truth" data

The Rhône-Alpes region authorities conducted a household travel survey (HTS) covering the entire region for the first time between 2012 and 2015. The survey consists in the telephone interviews of 37 450 individuals aged 11 and over, who represent 0.7% of the population in that age group. Individuals in the sample reported 143 000 trips. The Rhône-Alpes region covers an area of 43 700 km^2 with a population of 5.2 million inhabitants aged 11 and over.

A geographical stratified random sampling technique was used. The geographical stratification corresponds to a system of 77 zones (denoted as HTS-sectors) for the whole region. Figure 1b shows the 77 HTS-sectors and, with different colours, their aggregation in 14 macro-zones. Each HTS-sector involves at least 450 surveyed individuals.

The survey collects socio-demographic characteristics of the individuals and of the household, as well as information about all the trips that they made in a 24-hours period. This period runs from 3:00 a.m. on the day before the interview to 3:00 a.m. on the day of the interview. The most important attributes characterising a trip are transport mode, start and end time of the trip at minute-level granularity, activity at the origin and activity at the destination, location of the origin and location of the destination. Data has been collected through three waves in 2012-13; 2013-14 and 2014-15 from late autumn to early spring gathering only working day trips (Monday-Friday, excluding school and public holidays). Survey methodology is similar to other travel surveys conducted in urban areas in France (CERTU, 2008).

Figure 1. Maps of the Rhône-Alpes region with (a) 3G cells and coloured 3G location areas and (b) household travel survey sectors and 14 coloured macro-zones
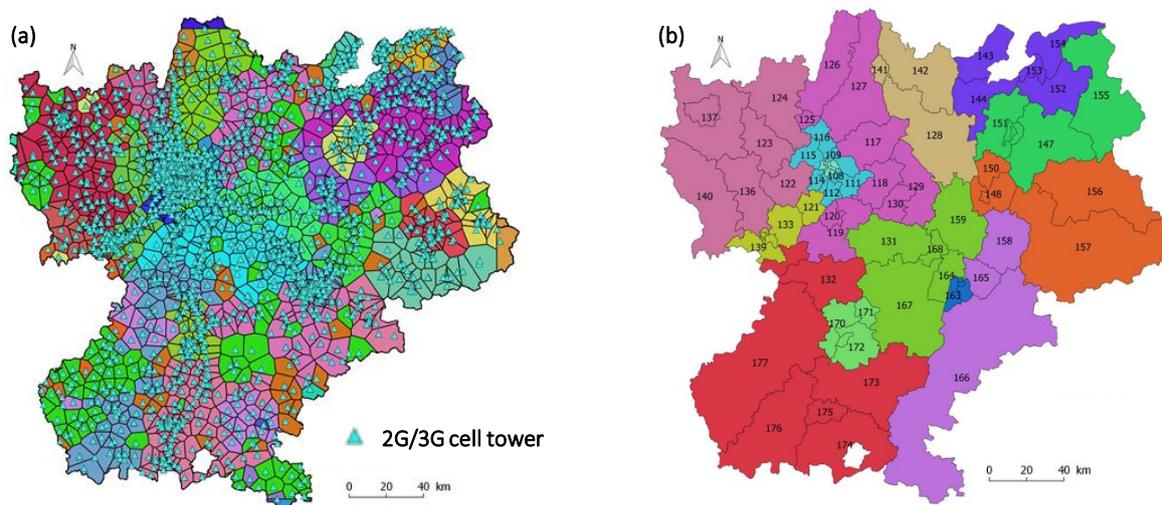

Source: Fekih et al. (2020a).

Due to the number of observed trips in the HTS, origin-destination matrices have been aggregated into a 14-zone system (Figure 1b). Regression analysis is performed to compare the amount of trips for each origin-destination (O-D) pair. That helps validate the scaling step. For this comparison, the threshold of 30minutes applies for both signalling and EDR data. In other words, the author identified activities shorter than 30 minutes in the HTS dataset and removed trips to and from such activities.

In addition to the total amount of trips, the coefficient of determination R^2 with value *0.96* between macro-zone trips gives a high-level indication that the distributions of O-D flows are similar to the following regression equation:

$y_{ij} = 0.85 * x_{ij} + 877$, R^2 **=** 0.95 (Figure 2a)**,**

where y_{ij} is the number of trips from cellular signalling data for the O-D pair *ij* and x_{ij} is the number of trips from HTS. This result is observed if some trips are excluded from the regression analysis. These are called the Lyon O-D pairs and correspond to the trips between the city-proper macro-zone and the surrounding suburban macro-zone (Figure 1b). Figure 2b shows a slope coefficient closer to 0.70 if the two Lyon O-D pairs are included in the regression analysis.

Figure 2. Number of trips by origin-destination pair, estimated from cellular signalling data and from the regional household travel survey
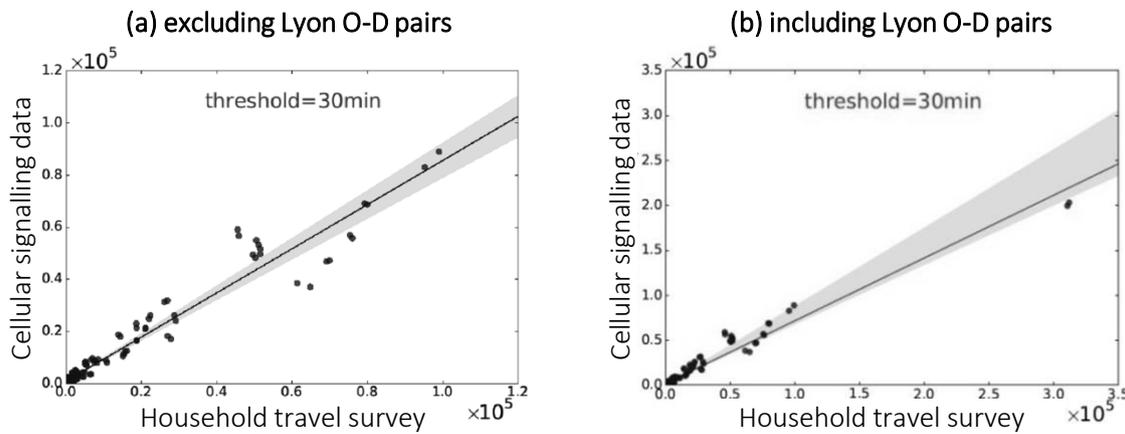
Note: The 30-minute threshold applies to the detection of activities in cellular signalling data. The same 30-minute threshold applies to select activities longer than 30 minutes in the household travel survey data.

Source: Fekih et al. (2020a).

What follows is the investigation of differences in trip numbers estimated by the two data sources for each O-D pair. Some very significant discrepancies occur, typically where cellular signalling data yields much lower trip volume estimates than what the HTS would suggest. In most cases, this occurs on non-adjacent zones and affects for both directions.

The comparison of trip numbers between HTS and cellular signalling data reveals some very consistent estimates. This occurs on some long-distance O-D pairs. For instance, Figure 3 show percentage differences in trip numbers between the two data sources, with a focus on trips to and from macro-zone 2 that correspond to the suburban ring around the city of Lyon. It is clear that flows between zone 2 and distant zones such as 7, 8, 9, and 10 are significantly underreported (blue flow lines) in the household travel survey. This could be due to weaknesses in the sampling of households.

However, cases where cellular signalling data under-estimates trip volumes occur mainly in high population density zones, such as between the Lyon urban and suburban macro-zones. The omission of 4G cellular signalling events might cause this phenomenon, as could imagine that the 4G coverage is particularly high in densely populated areas.

Some differences appear between the two matrices, but it is difficult to identify if these differences result from biases in the cellular signalling data or from the limited HTS sample size (probably both). For example, the HTS-based trip matrix clearly suffers from the problem of empty cells: at HTS sector level (Figure 1b) only about 40% of O-D pairs capture trips. In comparison, cellular signalling data captures trips on 95% of O-D pairs. Nevertheless, one concludes that a time threshold of 30 minutes yields trip matrices that are sufficiently consistent.

Figure 3. Trip volumes to and from the Lyon suburban area (macro-zone 2) estimated from cellular signalling data compared with household travel survey data

(a) Trips to macro-zone 2   (b) Trips from macro-zone 2
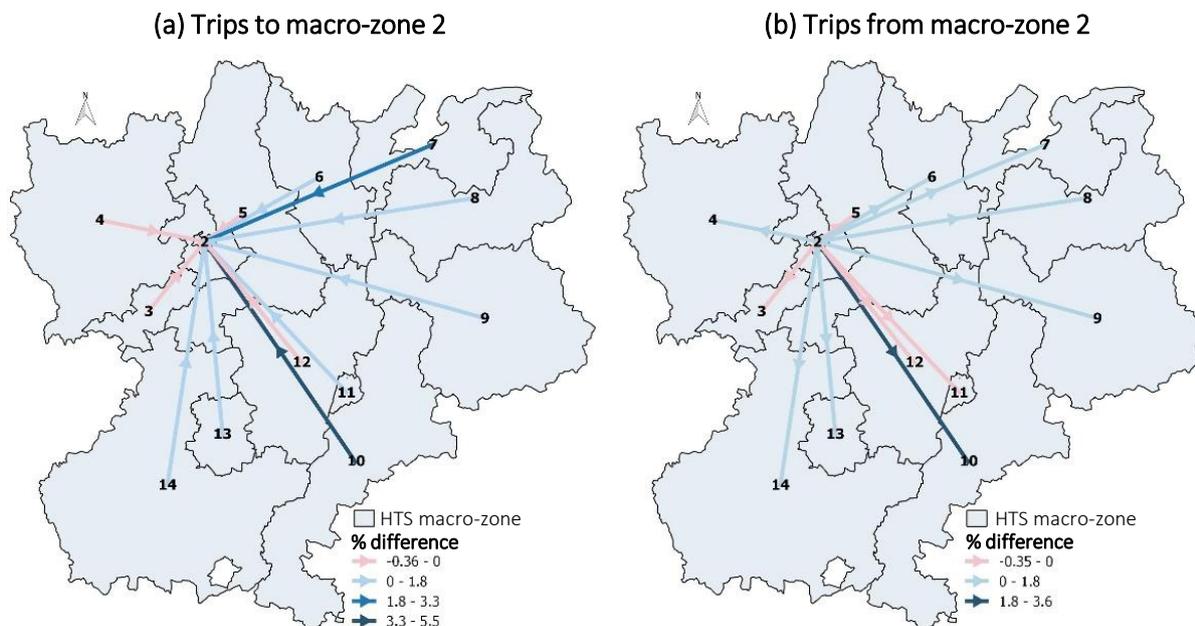
Note: Household travel survey (HTS).

Source: Fekih et al. (2020a).

Analysis of travel demand by time of day

The previous section proposes to split cellular signalling data across a spatial dimension, resulting in O-D trip matrices. This section adds a temporal dimension to create dynamic origin-destination matrices. As a first step, trip starting time is imputed to build hourly trip profile which is compared to the one of HTS survey. In a second step, the paper proposes a spatial clustering process combined with land-use data.

In order to study the dynamic trip-making patterns, one needs to associate to each trip a time window. For each user, a trip is imputed between every two consecutive stationary activities (i, i+1) happening within the 24-hour period. The trip occurs at a specific time spanned by the interval between the timestamp of the last mobile network event e associated to activity i and the timestamp of the first mobile network event e associated to activity i+1, noted as e_i, and e_{i+1}, respectively. Hence the start time is considered to take place in $[e_i, e_{i+1}]$. The real time instants associated to the end and the beginning of every consecutive pair of activities are unknown and can only be inferred from cellular signalling observations. Then the start time of the trip is estimated with the assumption that the start time is uniformly distributed in the time window $[e_i, e_{i+1}]$, and thus it is estimated to be in the middle of that interval. A first temporal profile is built using this hypothesis (Figure 4)

The signalling-based demand profile exhibits less sharp morning and afternoon peaks compared with the survey. The relatively crude estimation of trip start times from signalling data could explain this result. Another explanation could be the existence of a large fraction of users in the cellular signalling dataset who could be described as "static people". For those individuals, it is possible to detect the home sector but no trip can be observed, as the only stationary activity produced is performed at the home sector.

The proportion of such static people in cellular signalling dataset amounts to 46%. Even though a certain portion of static users could be actually stationary (e.g. elderly people), it appears highly probably that another large portion of them could be mobile, but due to their low mobile phone use or their reliance on 4G signal, no associated trips have been identified.

Figure 4. Temporal demand profiles from cellular signalling data
and household travel survey data
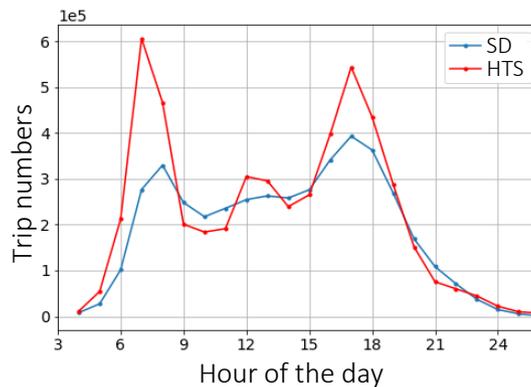
Note: Signalling data (SD), Household travel survey (HTS).

Source: Fekih et al. (2020b).

To address this bias, the author applied a uniform correction factor on all mobile phone trips with a start time estimated during the morning peak period (5-8 a.m.). This factor has been calculated as the ratio of the afternoon and morning peaks in the LAU profile which is less dependent of mobile phone use, allowing to consider the non-observed cell phone transactions and extract the hidden information from them. The applied temporal correction factor results equal to a value of 1.3 (Fekih et al., 2020b). It is important to note that the temporal correction is a debiasing procedure totally independent from the survey data (used for the comparative analysis), thus being easily reproducible by solely relying on information collected by the mobile network operator (i.e. the LAU events distribution). It highlights again the importance to use signalling data instead of CDR data.

As shown before (Figure 3) the discrepancy in trip emission between HTS and cellular signalling data is abnormally highly-correlated to the urban density of the sectors. Specifically, the difference is much higher for denser urban areas compared to rural one. This aspect can be interpreted as an underestimation of the urban area travel demand in the case of the cellular signalling data, due to the difficulty of such data in capturing short-distance trips (more likely to happen between adjacent/smaller urban areas). In order to address this bias, a spatial correction factor per zone is applied. This factor is estimated through a regression between the total demand difference per zone (estimated through HTS data and cellular signalling data) and the urban density (as computed from land use data). By applying this correction factor Fekih et al. (2020b) were able to decorrelate this difference with respect to the urban density and obtain more similar temporal profile (Figure 5).

Figure 5. Temporal demand profile from cellular signalling data and household travel survey data after spatiotemporal debiasing procedures
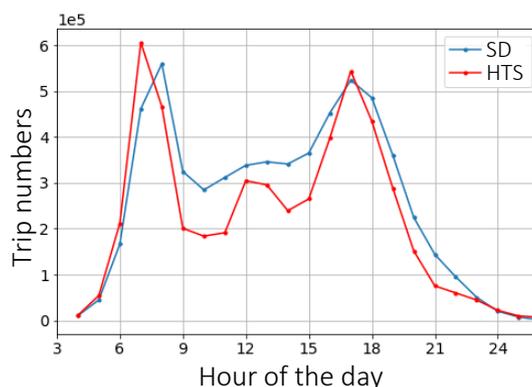
Note: Signalling data (SD), Household travel survey (HTS).

Source: Fekih et al. (2020b).

The author later conducted a spatial clustering analysis. This facilitates the comparison of temporal demand profiles in various land-use configurations. In urban areas, it reveals that signalling-based estimations are slightly lower than those estimated from a travel survey in the afternoon peak, but they nevertheless match the temporal demand profile relatively closely. In other areas, rural or mixed, signalling-based estimations match those from survey with slight differences in the morning peak. These results confirm that cellular signalling data can provide very useful information and solve the sampling rate problem of surveys in large mixed and rural areas, if properly debiased.

Cellular signalling data clearly present a high potential for the detailed spatiotemporal analysis of mobility. Nevertheless, debiasing procedures are fundamental to develop these analyses. Some of these procedures require no additional data sources, but others require input from travel surveys or other big data sources.

Dynamic mobility analysis with smart card data

As for mobile network data, this section briefly presents data processing methods in order to extract and expand trip origin-destination matrices from public transport smart card transactions. It then compares the result to "ground truth". A dynamic analysis of the matrices follows.

Smart card data and its processing methods

Data used in this section come from Keolis Lyon, the company operating the public transport network in Lyon under the supervision of SYTRAL, the public transport authority of the Lyon metropolitan area. The network consists of 4 lines of metro, 2 lines of funicular, 5 lines of tramway and more than 100 regular lines of bus. Before going into detail on data process terminology used in this paper is clarified. A trip refers to the movement

of a passenger from an origin to a destination. On bus and tramways, a trip-leg is the movement of a passenger on a single vehicle between a boarding and an alighting points. In the subway networks (metro and funicular) transfer don't require to validate a fare. To ensure comparability between the datasets, in this environment a trip-leg is defined as the movement of a passenger between an entry station and an exit station. A trip is made of one or more trip-legs. Transfer between trip-legs can include waiting time and walking time but shall not include activities. An itinerary is a set of lines or entry stations (in the case of subway) that together make up a trip. A day (or service day) is defined from 4.30 a.m. to 4.30 a.m. of the next day.

The public transport fare system in Lyon is an entry-only transaction system. Passengers can use smart cards or magnetic paper ticket, with the latter representing a little more than 20% to total ticket validations. Whatever the fare support, passengers are normally required to validate every time they board a moving vehicle (tramway and buses) or enter the subway network (metro and funicular station). For this study, all fare transactions from 13 to 17 March 2017 were extracted. This period corresponds to 5 standard weekdays of a month deemed to be representative. In the fare transaction database only smart card can be uniquely identified (all paper tickets share the same id).

To enrich these data sources the following steps were implemented (Egu and Bonnel, 2020a):

- **Data correction**. In a moving vehicle, the automatic fare transaction system is normally integrated with the automatic vehicle location (AVL) making it possible to know the boarding stop, line, direction and route. In cases where this information was missing, it was imputed using the AVL dataset;

- **Transfer identification**. To identify the beginning of a new trip and isolate transfer the following rules was implemented: 1) The first transaction of a day is always the beginning of a new trips; 2) two boarding transactions that occur within 60 minutes and that are not made on the same bus route or on subway station are considered part of the same trip (Munizaga and Palma, 2012; Spurr et al., 2015; Devillaine et al., 2012; Seaborn et al., 2009). For paper tickets, it is assumed that all paper tickets transaction identified in the database as first stamp-in correspond to a trip departure;

- **Destination inference.** Inference of alighting stop or exit station was done using the trip chaining method minimising the distance between successive boarding transactions (Trepanier et al., 2007; Nunes et al., 2016; Zhao et al., 2007; Li et al., 2018). The maximum walking distance was set at 600 metres. For the last transaction of the day, a return to the closest stop of the first transaction of the day was considered.

This method applied to smartcard data yields a percentage of alighting point estimated of 81% which is quite comparable with various results obtained on entry only system (Munizaga and Palma, 2012; Trepanier et al., 2007; Li et al., 2018; Nunes et al., 2016). For paper tickets trip destination and transfer information remain undocumented and could not be inferred. Lastly, some trips simply do not appear in the database because of fare non-interaction and fare evasion – about 21% according to the last fraud survey (SYTRAL, 2017). Altogether, the estimated origin-destination trip matrix contains about 48% of all public transport trips and therefore needs to be expanded.

Without any additional information, one would make the hypothesis that the spatial structure of the matrix is correct and one would elaborate a uniform scaling factor. It is clear however that paper tickets and fraud are not uniformly spatially distributed. Automatic passenger counting system are available on all public transport vehicles in the Lyon area and can be used for scaling purpose. The method proposed by Gordon et al. (2018) is thus implemented, that computes a scaling factor for each trip so as to ultimately match a set of known control flows. In this paper, flow data is collected over a 5-day period (13 to 17 March 2017).

Validation against "ground truth" data

In order to validate the matrix built with smart card and APC data, two kinds of surveys are available in Lyon conurbation: household travel survey and origin-destination survey.

In the Lyon metropolitan region, the last household travel survey was performed between October 2014 and April 2015. This survey was done according to the French standard methodology defined by the CEREMA (CERTU, 2008). Within the perimeter of Lyon transit services, the survey is based on face-to-face interviews. The survey focuses only on weekdays and is performed from Tuesday to Saturday to collect trips made on the day before, from Monday to Friday. Survey sample is based on a geographically stratified random selection of household with a response rate of about 50%. To build a public transit O-D trip matrix with this data source, all trips that were declared as made on the transit network were extracted (without taking into consideration the household zone of residence). The unweighted sample consists of 10 570 trips carried out by 4 708 distinct individuals. Origin and destination of trips are known at the level of the finest zoning system which range from communal level to small zones of less than 100m squares in areas with high population density. They are therefore always much smaller than the zoning system of Figure 1. The expansion of the matrix to the full population of residents is done using the individual calibration weight estimated by CEREMA (CERTU, 2008).

In Lyon transit network area, all lines are manually surveyed every five years in a continuous process. During the survey, each passenger is asked to describe their current trip. The following data is recorded:

- boarding and alighting stops for the trip-leg on the surveyed line;
- if any, connecting line before the surveyed line (up to 3);
- if any, connecting line after the surveyed line (up to 3);
- origin and destination of the full trip coded at the stop level;
- socio-demographic of the passenger, trip purpose and public transit access modality.

Each line is surveyed for three types of days: weekdays, Saturdays and Sundays. On bus lines, this survey is done on board and is supposed to be exhaustive (all daily runs of each day type are surveyed). For tramway and subway lines, this survey is done at stations. A random sample of approximately 35% of passengers for tramway lines and 25% of passengers for subway lines are surveyed. The results are then scaled according to half-hour count by station derived from automatic passenger counting system. To obtain data for the complete networks, all the OD surveys collected between 2013 and 2017 on weekdays were compiled into one single database. The goal of this survey is to obtain accurate information about transit usage on each line. It is used for tactical and operational level planning (route itinerary modification, network redesign, route level transfer analysis etc.). Due to its high sampling, it is the most used dataset for public transport planning in Lyon.

Table 2 reveals substantial differences between data sources. The total number of trips legs and trips are similar for O-D survey and smart card data weighted with APC data, but significantly different from household travel survey. Results are closer in terms of mode shares.

Table 2. Comparison between smart card, origin-destination survey and household travel survey data

	Smart card data	Public transport origin-destination survey	Household travel survey (HTS)
Trip legs (million)	1.56	1.51	1.11
Trips (million)	1.10	1.16	0.80
Bus trip legs (%)	41	39	43
Tramway trip legs (%)	23	22	21
Subway trip legs (%)	37	39	36

Source: Egu and Bonnel (2020a).

Figure 6 shows major differences between smart card data and household travel survey data in the number of trips per day and per person. Smart card data reveals a much lower share of people (50%) doing two trips per day, and a much higher share of people (18%) doing only one trip per day.

Figure 6. Distribution of the number of trips per day and per person
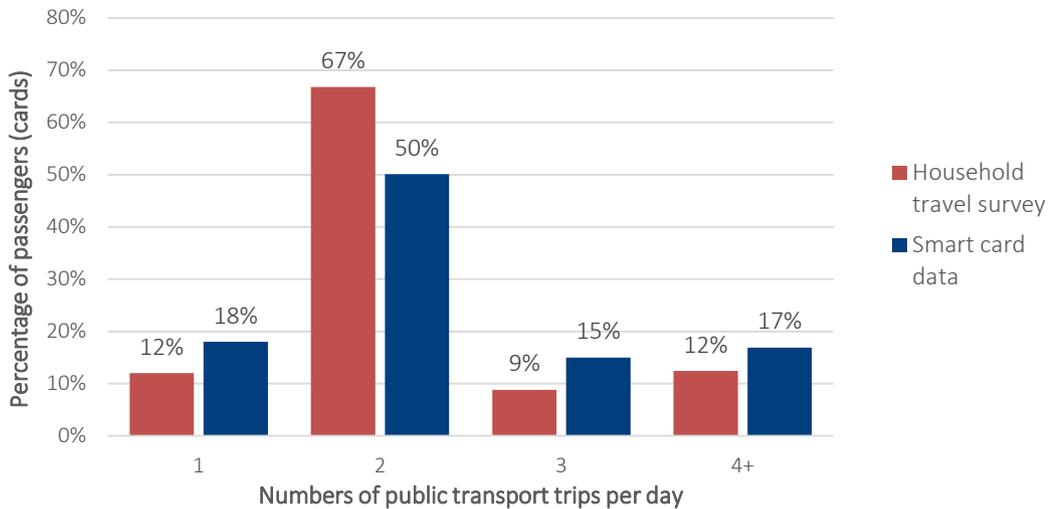
Source: Egu and Bonnel (2020a).

Interpreting such differences remain challenging due to the incomplete nature of smart card data, which excludes paper tickets and fraud. Figure 6 however suggests that household travel surveys under-estimate the number of people who do not use the same mode of transport on the way out and back, as already observed in Spurr et al. (2015) and Seaborn et al. (2009). This observation could be relevant to transport planning authorities and public transport operators as they would otherwise under-estimate the share of people doing a single public transport trip per day and completing their tour using other modes.

Figure 7. Regression analysis between matrices
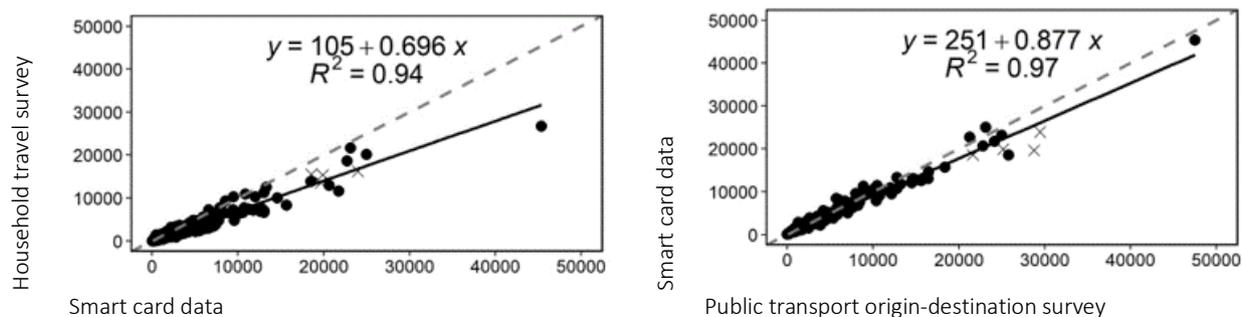
Source: Egu and Bonnel (2020a).

The differences were further inspected at the level of the origin-destination matrices using a zoning system with 18 zones in order to have enough trips for each origin-destination of the HTS survey. The regression plots (Figure 7) indicate that all the matrices structures are globally consistent but with some differences. As indicated by the regression equations, the slopes are varying from 0.7 to 0.88 as a result of different total trips volume. Matrices developed from O-D survey and smart card data (weighted by APC) are most consistent with a slope close to 1 as expected if matrices were identical. With a slope close to 0.7, the comparison of matrices from HTS and smart card data confirms the results of Table 2.

One generally considers public transport O-D surveys as a robust solution to measure patronage. There exist some differences between smart card data and O-D survey but they might occur because an O-D survey consolidates data sampled over a 5-year period. Nevertheless, these differences are minor compared to differences observed with HTS data. When combined with external sources like automated counting system and after imputing any missing information, smart card data can therefore offer a relevant tool to understand and monitor public transport demand. It also enables the analysis of demand dynamics, which is not possible with cross-sectional surveys.

Analysis of travel demand over a 6-month period

This section explores the day-to-day variability of public transport use at the level of the individual, a concept also called *intrapersonal* variability. Traditional cross-sectional survey data is often analysed with the implicit hypothesis that individuals are performing the same trips every working day. Previous analyses have given evidence that it is not true. It is simply not possible to document the intrapersonal variability with cross-sectional data.

Smart card data can reveal intrapersonal travel variability over a given period on the condition that card identifiers (IDs) are not reset over this period. Keolis Lyon gave access to disaggregate smart card data with the same IDs during a 6-month period (Egu and Bonnel, 2020b). A committee for ethics and privacy protection internal to Keolis authorised the supply of this data, based on the national and European rules for privacy protection. It is the author's understanding that smart card data reveals much less personal information than mobile network data (tracking the user 24/7 across all modes) or smartphone app data (tracking accurate location details, to the exact address point).

In this research, data from 1 January 2017 to 30 June 2017 were extracted from the fare collection database. This study period consists of 181 days with at least 25 days of observations for each day of the week. It includes two school holiday periods: winter break (from 18 February 2017 to 5 March 2017) and

spring holidays (from 15 April 2017 to 1 May 2017) and six bank holidays (1 January 2017, 17 April 2017, 1 May 2017, 8 May 2017, 25 May 2017, 5 June 2017). Those events can affect individual usage pattern and may be of interest in terms of 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 intrapersonal vs interpersonal variability. 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.

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,

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

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 metric, a dissimilarity matrix M is calculated in which each cell Mkl takes the value of D(Xk;Xl). This matrix is an input to the clustering algorithm. Hierarchical clustering is a common approach that does not require to commit a particular number of clusters (Friedman et al., 2001). It produces a dendrogram that illustrates how the objects are joined together.

A random sample of 40 000 cards among the 591 124 cards observed travelling at least once between January and June 2017, was drawn to apply the clustering method. The results are represented on a heatmap in Figure 8. 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 8 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 periods in weeks 8-9 and weeks 16-17 are also visible and can lead to episodic break of usage. Lastly, some users exhibit clear changes in usage intensity over the six months.

Figure 8. Dendogram resulting from the application of the clustering method
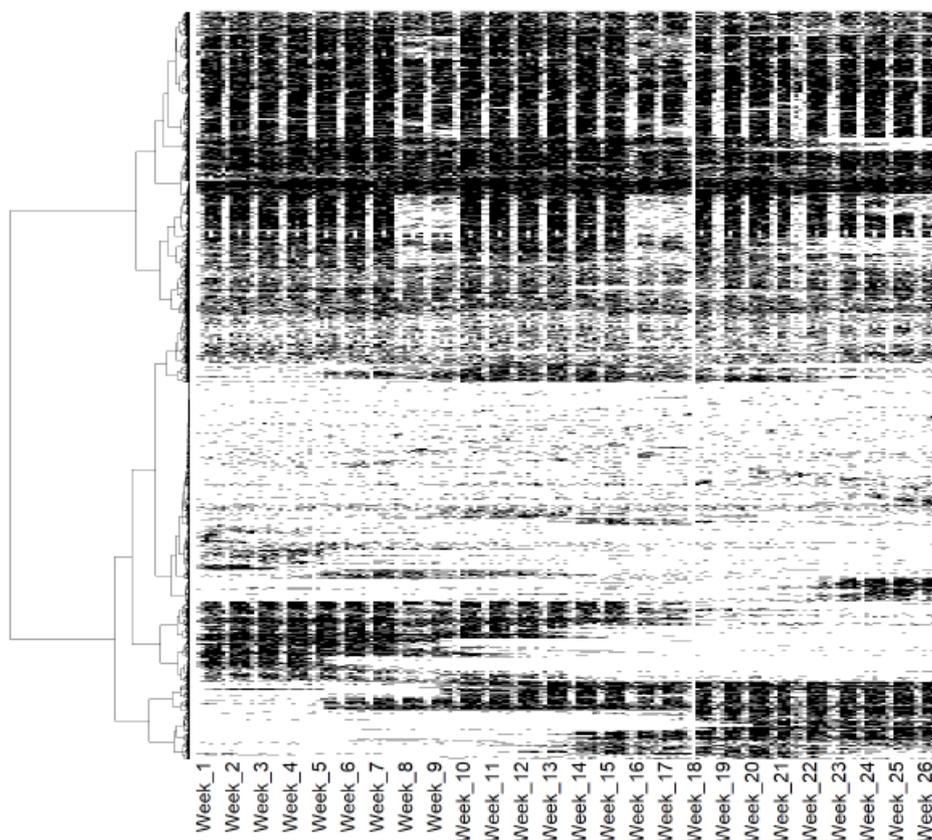
Source: Egu and Bonnel (2020b).

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

- **The low-frequency users (LF)**, mainly located in the middle of the dendrogram. They almost never use the system on a multi-month scale. 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.

- **The consistent transit users (CT)**, mainly located at the top of the dendrogram. They used the transit system consistently over the 6-month 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. Transit usage is part of their daily routine from the beginning to the end of the study period.

- **The intermittent transit users (IT)**, mainly located at the bottom of the dendrogram. They present characteristics of low-frequency users but also characteristics of consistent transit users. Transit usage is part of their routine but at one point of the study period, they exhibited a marked change in day-to-day usage intensity.

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; 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; 3) otherwise users are classified as CT.

The authors applied those rules to the 40 000 cards of the initial sample. Table 3 presents the number of users in each group. The smallest group corresponds to the LF users with a total of 5 456 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.

Table 3. Distribution of smart card users by public transport usage regularity

	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

Source: Egu and Bonnel (2020b)

The rest of the paper focuses on the behaviour of the 18 186 consistent transit users to examine day-to-day regularity and intrapersonal variability at an aggregate level. It allows to maximise the observation period and focus on a group that share common characteristics in terms of multi-month usage routine. On this subpopulation, at an aggregate level each working day appears quite similar with a similar number of transit trips, a similar spatial distribution of trip ends and a similar distribution of starting time profile. But the similarity at an aggregate level that has been put in evidence by other authors (Schlich and Axhausen, 2003; Hanson and Huff, 1988; Morency et al., 2007) does not mean that individuals are doing the same transit trips every day.

Huff and Hanson (1986) have proposed a trip-based similarity measure between two days **i** and **j** that can measure conjointly the trip and spatiotemporal characteristics of trip and can be expressed as follows:

$$S_{ij} = [1 - \frac{1}{2}\sum_k |P_{ic} - P_{jc}|] \frac{n_i}{n_j} \quad , \quad n_j \geq n_i$$

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.

Equivalent class are defined based on trip characteristics. With smart card data, transaction time and boarding stop are available. Both features have high cardinality as there are many stops in the network (more than 4 000) and the timestamp is known with second precision. To reduce the dimension of those features, two grids are used that make sense both from a practical and behavioural point of view:

- **Temporal grid**. Trip starting times are grouped into the following time slots: before 7 a.m., 7 a.m. to 10 a.m., 10 a.m. to 12 p.m., 12 p.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.

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.

A first aggregation of Sij could be to calculate for all users the mean similarity between any two days of the week (Table 4). This analysis results in the following observations:

- Weekend days are less similar than other days of the week, which confirms the finding of Schlich and Axhausen (2003). 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 with other weekdays decrease to 0.08.

- 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 comparing Monday with Friday.

- As found by Schlich and Axhausen (2003), Friday is the weekday that exhibits less similarity with the rest of the weekdays.

- 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.

Table 4. Similarity between days of the week

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	0.33	0.32	0.28	0.29	0.27	0.10	0.08
Tuesday		0.36	0.29	0.31	0.28	0.11	0.08
Wednesday			0.33	0.27	0.25	0.11	0.08
Thursday				0.34	0.28	0.11	0.08
Friday					0.31	0.11	0.08
Saturday						0.18	0.10
Sunday							0.18

Source: Egu and Bonnel (2020b).

Those results support the idea of Hanson and Huff (1988) that there is a high degree of regularity in public transport travel pattern, but there is also systematic day-to-day variability. Thus, to characterise travel patterns, collecting data on a single day is insufficient. In order to better characterise intrapersonal variability among the 18 186 consistent users, the clustering technique described above is used. With the help of the dendogram, six clusters are defined. To visualise the clusters, Figure 9 plots 100 cards randomly selected from each cluster.

Figure 9. Visualisation of the day-to-day usage pattern of 100 random users selected from each cluster
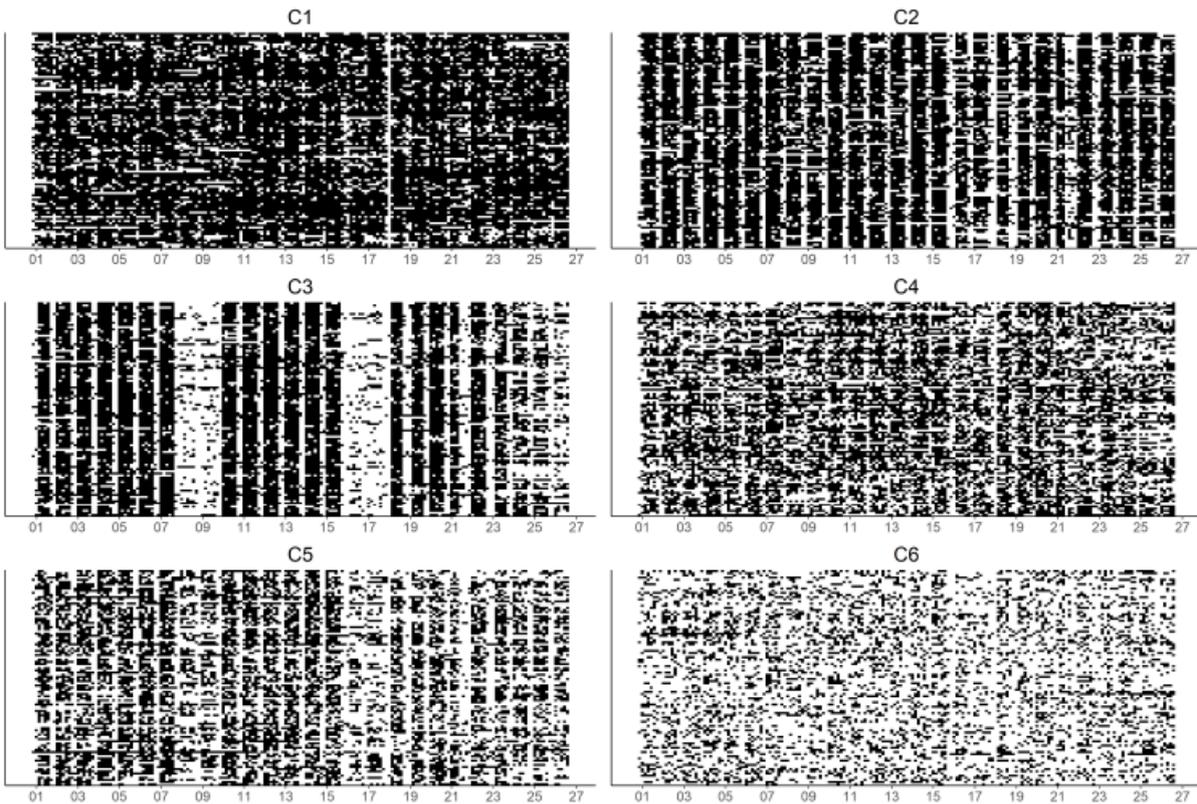
Source: Estimations from smart card data (Keolis Lyon), Egu and Bonnel (2020b).

Cluster 1, with 3 211 cards, is characterised by the highest percentage of days of usage (86% on weekdays) and almost no calendar structure as the usage rate remains 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 4 592 cards (25% of the CT users). It is characterised by a very high usage rate on workings days but 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%).

Cluster 3 exhibits a concentration of trips in the two peak-periods, similarly to cluster 2, but their day-to-day usage pattern differs. Users in cluster 3 almost do not use the transit system during the holiday 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%.

Cluster 6 is formed by vectors that are very sparse i.e. with lots of zero. In this cluster, individuals consistently use the system over the 6 months but their usage rate is less than 30%. Cluster 6 is also characterised by a lower trip rate and a very high percentage of one trip day (39%).

Conclusion

Big data represent an opportunity to analyse travel behaviour. This paper illustrates that the potential for temporal dynamic analysis is very important at different levels of temporality. Due to the size of the data it is possible to analyse temporal dynamic within a day at a fine spatial granularity even for origin-destination matrices.

Origin-destination matrices built from travel surveys often present a high percentage of cells without any trips. In the context of the Rhône-Alpes region, considering the 77 zones used in the stratified sampling for the household travel survey, trips were sampled in only 40% of O-D pairs. On the other hand, this figure reaches 95% using cellular signalling data from a single telecommunication company. Of course, if one adds a temporal dimension to the analysis the problem becomes even worse. Big data offer the possibility to develop new spatiotemporal analysis that was not possible with traditional data due to sample size problems.

Contrary to cross-sectional surveys, big data generates a continuous feed of information. One can develop dynamic analysis like that using smart card data and presented in this paper. It provides the opportunity to analyse the variability of mobility over a long period. Different temporal dynamics can be identified like the weekly rhythm which distinguish each day of the week. The difference between working days and the weekend is well known, but even between working days the number of trips and characteristics of trips are not identical. The concept of average day is questionable and analysts should account for its variance.

A better understanding of travel demand dynamics would allow a better adaptation of the supply or of the regulation of the demand. An example of such analysis is presented in Egu (2020) who proposes a forecast by half-hour interval for each bus route and for each day of the year for the Lyon transit network using 4 years of smart card data through machine learning methods. Calendar effects and supply measurements are combined with yearly trends to propose a forecast at a highly disaggregated level. This kind of method represents a tool to accompany tactical planning (van de Velde, 1999) in between strategic planning (5 years and more horizon) where 4-step models are mostly used (Ortuzar and Willumsen, 2011; Bonnel, 2002) and real-time (or very short-term) planning where operating measures are taken. Such analysis does not require to keep the same identifier over time, as it does not require to follow the same individuals but only to have information for each day of the year (or the period studied).

When the same used identifier is kept over a longer period than one day, richer analysis can be performed. It is possible to follow the same individual to analyse the variability of their behaviour. An example of analysis has been proposed in this paper. It identified different kinds of public transport users. Transit authorities or operators can adapt marketing or pricing strategies in order to increase public transport use in each cluster as for example those identified in Figure 9. Other analyses are proposed in the abundant literature on big data.

Many pieces of research have demonstrated the usefulness of big data, including cellular signalling data and smart card data. It is however important to identify also the limitations of these new sources of information.

For durations longer than one day, big data might present some limitations due to privacy protection laws. The response is always context-dependant. It depends on national laws and regulations, the content of the data, choices made by the organisation that manages the data, etc. For example, in the French context, the user identifier changes every day in the cellular signalling data from Orange, whilst the same identifier remains unchanged for a year in smart card data from Keolis Lyon.

The privacy threat is obviously most acute with most precise user positioning techniques and with door-to-door tracking. For this reason, GPS traces collected by smartphone apps are probably much more intrusive than the data used in this article. Clearly this question requires more research and the development of ethical rules in order to present research results at a suitable level of aggregation. In spite of the strict anonymisation rules like the ones developed by Orange, the potential of big data remains important.

Big data is not necessarily representative of a population. This paper provided many examples of users excluded from or misrepresented in the sample: those using another mobile network operator, making exclusive use of 4G, making too little use of the cellular network, travelling using paper tickets, etc. In the case of cellular data, not only can people be left out, but machines part of the Internet of Things can also be mistaken for humans.

The validation of the data is therefore always required. This validation cannot be performed internally by some kinds of norm quality. It requires the confrontation of the data to other sources, to "ground truth". Furthermore, the technology which generates big data continuously evolves which implies the continuous update of data treatment and validation protocols.

The question of representativeness is also dependent on data processing. In this paper, the number of trips is dependent on some thresholds like the activity duration for cellular signalling data or on some rules to impute destination for smart card data. Most often the quality of the result is related to the number of observations for each individual. One has to seek a trade-off between the omission of some individuals whose observations are too limited and the bias that their omission generates. For example, individuals who make limited use of their mobile phone are probably not travelling like the average population. Individuals who use paper tickets are probably not travelling like smart card users.

Weighting processes are therefore required. Generally, their quality is improved by the use of external sources. In case of mobile phone data, imputation of home location allows combining the data with census data or land use data to generate non-uniform weighting parameters. More research is required to improve data fusion methods in order to improve these processes.

Big quantities of data rarely come with rich data semantic. Big data often comes with very few attributes. Trip characteristics are often missing which require imputation rules to complement the data. For privacy reason socio-demographics are mostly not available. Again, some imputation rules are developed. Most of these rules are based on data fusion with external sources. Traditional travel surveys are often good candidates for these analyses. In this sense, the development of big data solutions for transport planning depends on the availability of more traditional surveys.

References

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

Bonnel, P., M. Fekih and Z. Smoreda (2018), "Origin-Destination Estimation Using Mobile Network Probe Data", *Transportation Research Procedia*, Vol. 32 (January), pp. 69–81, https://doi.org/10.1016/j.trpro.2018.10.013.

Bonnel, P. et al. (2015), "Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations", *Transportation Research Procedia*, Vol. 11, pp. 381-398, https://doi.org/10.1016/j.trpro.2015.12.032.

CERTU (2008), *L'enquête ménages déplacements standard CERTU*, Éditions du CERTU.

Davies, D. L. and D. W. Bouldin (1979), "A cluster separation measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, No. 2, pp. 224-227, https://doi.org/10.1109/tpami.1979.4766909.

Egu, O. (2020), *Apports des données passives à la compréhension des comportements de mobilité ? Enjeux pour la planfication et l'organisation des transports en commun*, thèse de doctorat de l'Université de Lyon, LAET, Lyon, 152p.

Egu, O. and P. Bonnel (2020a), "How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination survey and household travel survey? An empirical investigation in Lyon", *Transportation Research Part A Policy and Practice,* 138, pp. 267-282,.

Egu, O. and P. Bonnel (2020b), "Investigating day-to-day variability of transit usage on a multimonth scale with smart card data: A case study in Lyon", *Travel Behaviour and Society*, Vol. 19, pp. 112-123. https://doi.org/10.1016/j.tbs.2019.12.003.

Devillaine, F., M. Munizaga and M. Trepanier (2012), "Detection of activities of public transport users by analyzing smart card data", *Transportation Research Record: Journal of the Transportation Research Board*, 2276 (1), pp. 48-55, 2012, https://doi.org/10.3141%2F2276-06.

Fekih, M. et al. (2020a), "A data-driven approach for origin-destination matrix construction from cellular network signalling data: A case study of Lyon region (France)", *Transportation* (2020). https://doi.org/10.1007/s11116-020-10108-w.

Fekih, M. et al. (2020b), "Assessing The Potential Of Cellular Signalling Data To Estimate Dynamic Travel Demand Patterns: A Comparative Study With Travel Survey Data", in 99th Transportation Research Board Annual Meeting (TRB), Washington, United States.

Friedman, J., T. Hastie and R. Tibshirani (2001), "The elements of statistical learning", Vol. 1, *Springer Series in Statistics*, Springer, New York.

Gordon, J. B., H. N. Koutsopoulos and N. H. Wilson (2018), "Estimation of population origin-interchange-destination flows on multimodal transit networks", *Transportation Research Part C: Emerging Technologies*, Vol. 90, pp. 350-36, https://doi.org/10.1016/j.trc.2018.03.007.

Gundlegård, D., C. Rydergren, N. Breyer and B. Rajna (2016), "Travel Demand Estimation and Network Assignment Based on Cellular Network Data", *Computer Communications*, Vol. 95, pp. 29–42, https://doi.org/10.1016/j.comcom.2016.04.015.

Hanson, S. and J. O. Huff (1988), "Systematic variability in repetitive travel", *Transportation*, Vol. 15(1-2), pp. 111-135, https://doi.org/10.1007/BF00167983.

Hanson, S. and J. O. Huff (1981), "Assessing day-to-day variability in complex travel patterns", *Transportation Research Record*, Vol. 891, pp. 18-24.

Huff, J. O. and S. Hanson (1986), "Repetition and variability in urban travel", *Geographical Analysis*, Vol. 18 (2), pp. 97-114.

Li, T., D. Sun, P. Jing and K. Yang (2018), "Smart card data mining of public transport destination: A literature review", *Information*, Vol. 9(1), https://doi.org/10.3390/info9010018.

Morency, C., M. Trepanier and B. Agard (2007), "Measuring transit use variability with smart-card data", *Transport Policy*, Vol. 14(3), pp. 193-203.

Munizaga, M. and C. Palma (2012), "Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile", *Transportation Research Part C: Emerging Technologies*, Vol. 24, pp. 9-18, https://doi.org/10.1016/j.trc.2012.01.007.

Nunes, A. A., T. G. Dias and J. F. e Cunha (2016), "Passenger journey destination estimation from automated fare collection system data using spatial validation", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17(1), pp. 133-142, doi: 10.1109/tits.2015.2464335.

Ortuzar, J. d. D. and L. Willumsen (2011), *Modelling transport,* 4th Edition, John Wiley & Sons, Chichester.

Rousseeuw, P. J. (1987), "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

Schlich, R. and K. W. Axhausen (2003), "Habitual travel behaviour: Evidence from a six-week travel diary", *Transportation*, Vol. 30(1), pp. 13-36, https://doi.org/10.1023/A:1021230507071.

Seaborn, C., J. Attanucci and N. H. M. Wilson (2009), "Analyzing multimodal public transport journeys in London with smart card fare payment data", *Transportation Research Record*, Vol. 2121(1), pp. 55-62, https://doi.org/10.3141%2F2121-06.

Smoreda, Z., A.-M. Olteanu-Raimond and T. Couronné (2013), "Spatiotemporal data from mobile phones for personal mobility assessment", in J. Zmud, M. Lee-Gosselin, J.A. Carrasco, M. Munizaga (eds.) *Transport Survey Methods: Best Practice for Decision Making*, Emerald Group Publishing Limited, pp. 745-768.

Spurr, T., A. Chu, R. Chapleau and D. Piche (2015), "A smart card transaction "travel diary" to assess the accuracy of the montréal household travel survey", in C. Morency, M. Munizaga, S. Greaves and T. Raimond (eds.), *Transport Survey Methods: Embracing Behavioural and Technological Changes*, *Transportation Research Procedia*, Vol. 11, pp. 350–364.

Spurr, T., A. Leroux and R. Chapleau (2018), "Comparative structural evaluation of transit travel demand using travel survey and smart card data for metropolitan transit financing", *Transportation Research Record*, Vol. 2672(8), pp. 807-816, https://doi.org/10.1177%2F0361198118773897.

Stopher, P. R. and S. P. Greaves (2007), "Household travel surveys: Where are we going?" *Transportation Research Part A*, Vol. 41(5), pp. 367– 81, https://doi.org/10.1016/j.tra.2006.09.005.

SYTRAL (2017), *Rapport Enquête Fraude de Mars 2017*, Sytral, Lyon.

Trepanier, M., N. Tranchant and R. Chapleau (2007), "Individual trip destination estimation in a transit smart card automated fare collection system", *Journal of Intelligent Transportation Systems*, Vol. 11(1), pp. 1-14, https://doi.org/10.1080/15472450601122256.

van de Velde, D. M. (1999), "Organisational forms and entrepreneurship in public transport: Classifying organisational forms". *Transport Policy*, Vol. 6(3), pp. 147-157.

Wang, M.-H., S. D. Schrock, N. Vander Broek and T. Mulinazzi (2013), "Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data", *International Journal of Intelligent Transportation Systems Research,* Vol. 11 (2), pp. 76–86, https://doi.org/10.1007/s13177-013-0058-8.

Wolf, J., M. Oliveira and M. Thompson (2003), "Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System- Enhanced Household Travel Survey", *Transportation Research Record,* Vol. 1854, pp. 189–98, https://doi.org/10.3141/1854-21.

Zhao, J., A. Rahbee and N. H. Wilson (2007), "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 22(5), pp. 376-387.

Zhao, Z., S.-L. Shaw, Y. Xu, F. Lu, J. Chen and L. Yin (2016), "Understanding the Bias of Call Detail Records in Human Mobility Research", *International Journal of Geographical Information Science*, Vol. 30(9), pp. 1738–62, https://doi.org/10.1080/13658816.2015.1137298.

International Transport Forum

Benefits of Cellular Telecommunication and Smart Card Data for Travel Behaviour Analysis

This paper proposes the estimation of trip origin-destination matrices using big data through two case studies. In the first, trip matrices are estimated from mobile network data and compared with household travel survey results. In the second, public transport trip matrices are derived from smart card data and compared with passenger survey data. The paper concludes that sample size and longitudinal data collection are big data's main strengths, yet are limited by privacy protection constraints and by the need to control for biases in the sample.

All resources from the Roundtable on Use of Big Data in Transport Models are available at:

www.itf-oecd.org/big-data-transport-models-roundtable

OECD