Transportation Research Part C 43 (2014) 198–211

Contents lists available at ScienceDirect

Transportation Research Part C

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

Using cell phone data to measure quality of service and passenger flows of Paris transit system

Vincent Aguiléra [a,*], Sylvain Allio [b], Vincent Benezech [a], François Combes [a], Chloé Milion [b]

[a] Université Paris Est, LVMT, UMR T9403 (ENPC IFSTTAR UPEMLV), Marne-la-Vallée, France
[b] Orange Labs, Belfort, France

ARTICLE INFO

Article history:
Received 6 January 2013
Received in revised form 25 June 2013
Accepted 5 November 2013

Keywords:
Quality of service
Transit network
Cellular phone data

ABSTRACT

This paper shows that the particular conditions under which a cellular phone network is operated underground can make it possible to measure passenger flows in an underground transit system. With the help of the mobile network operator Orange, some experiments have been conducted in Paris underground transit system to assess the potential of this new kind of data for transportation studies. The results show that good estimates of dynamic quantities, such as travel times, train occupancy levels and origin–destination flows can be derived from cellular data. The travel times, train occupancy levels and origin–destination flows inferred from cellular data have been compared to direct field observations and Automatic Fare Collection data provided by the STIF (the public transport authority in the Paris metropolitan area). The quantities inferred from cellular data are shown to be consistent with those inferred from the other data sources.

1. Introduction

Quality of Service (QoS) is a central concern for public transport users, transport operators and transport authorities, but they do not necessarily see it in the same way. During a particular trip, users are concerned with, among other factors, their own comfort and delays. Over a longer period of time, they may consider more aggregate indicators, such as the reliability of the system or the availability and accuracy of real time information in order to reduce uncertainty. Operators, especially in the railway sector, are mainly concerned with the schedule adherence of services, which may significantly differ from that of passengers. Transport authorities are increasingly seeking indicators that reflect the distribution of passengers riding experiences. The Transit Capacity and Quality of Service Manual TCQSM TCQSM (2003) cites the criteria of comfort, delays, convenience and reliability.

The comfort, or discomfort, experienced by the users of a transit system depends on the variations of the level of occupancy of the coach they have boarded into, and on the time spent in uncomfortable situations (e.g. standing up in an overcrowded coach). The chances of getting a seat when boarding a crowded coach are very low. Along the trip, as more and more passengers initially in the coach alight, a passenger's chance of getting a sit increase with the time spent in the coach. Hence a realistic understanding of the QoS of a transit system, as perceived by passengers, depends on:

- the level of occupancy, for each train and each interstation, during the day;
- the travel times between any pair of consecutive stations, for any departure time;
- origin–destination passenger flows and route choice, for several time-slices during the morning and evening peaks.

* Corresponding author.
E-mail address: vincent.aguilera@enpc.fr (V. Aguiléra).

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.11.007

Traditionally, QoS indicators are computed using costly and error-prone methods, such as observed timetables and manual passenger counts at stations. The development of Automated Fare Collection (AFC) and Automatic Vehicle Location (AVL) has significantly improved QoS measurements. But, to the authors knowledge, the ability of most AFC/AVL systems to estimate QoS indicators is limited by design. Deducing the origins and destinations of trips (and/or travel times) is far from straightforward, if not impossible. Also, in order to deduce the level of occupancy of trains it must be possible to access reliable AFC and AVL data and combine the two. This may be a complex task, especially if passengers are able to choose different missions and/or routes to make a given trip.

This paper investigates a novel approach: the use of mobile phone *signaling* data to monitor an underground transit system. For reasons explained later on, along a trip underground, a switched-on mobile phone may trigger a few so-called *signaling events* (which we will refer to as *GSM data* in what follows). For a given phone, those data is sparse, so tracking a single trajectory brings very little information. But the aggregation of thousands of such loose tracks provides a way of overcoming the aforementionned limitations of AFC/AVL data.

The main objective of this paper is to investigate how GSM data can be used to measure the QoS of a transit system. With the help of the mobile network operator Orange, we conducted a number of experiments in Paris' underground transit system. The results show that good estimates of the quantities we are interested in can be derived from the records of signaling events. For comparison and calibration purposes, the GSM data was then compared to direct field observations and to AFC data provided by STIF, the public transport authority in the Paris metropolitan area.

The rest of the paper is organized in four sections. The literature is briefly reviewed in Section 2. Section 3 provides some background, including (i) a general presentation of the underground part of the Paris transit system, and (ii) a more detailed description of the RER A central segment, which is used throughout the paper for illustrative purposes. Section 4 deals with the estimation of trains' level of occupancy. In Section 5, travel times and origin destination flows computed from AFC data are compared to those computed from signaling events records. The conclusion follows.

2. State of the art and literature review

Bertini and El-Geneidy (2003) list a number of indicators that can be derived from AFC and AVL, both on the supply side and from the user's point of view. A special report by Furth et al. (2006) serves as a comprehensive guide to the use of these data to measure QoS and to provide a basis for good transit management. AVL is easily available for buses. In this context, some innovative measures of service quality (El-Geneidy et al., 2011) and responsive transit management (Feng and Figliozzi, 2011) have been successfully applied. QoS indicators can also be derived from AFC data. For instance, Reddy et al. (2009) describe methods for the reconstitution of trips from smart card validations, but with many remaining issues: the precision is low, with results varying greatly according to day of the week for unknown reasons and multimodal trips are hard to pin down, with serious consequences on the estimation of trip chains. Nassir et al. (2011) introduce a more precise technique, but loose much data in the process. A specific framework has been developed by Frumin (2010) for the transit network of London, where validation is mandatory on exit.

Mobile phone data has already been identified as another potential data source. However, the vast majority of previous experiments considered road networks. The probes can consist of GPS devices—installed in drivers' cars or mobile phones (Lind and Lindkvist, 2006)—or can be the mobile devices themselves. The data is then obtained from telephone companies. Many experiments have been carried out, introducing different techniques, pursuing different aims and achieving different precisions. Friedrich et al. (2011) constructed O–D trip tables from disaggregated phone data that takes both private vehicle and transit; Wang et al. (2010) infer mode choice from aggregated data, by clustering according to speed. Other authors try to monitor traffic across sections of the road network (Hofleitner et al., 2012). For a more detailed review of the different possible technologies and past experiments, see Valerio (2009).

However, to the best of the authors' knowledge, no work has focused on the use of mobile phone data to study transit systems specifically. This could be due to two technical difficulties. First, except when transit lines and roads are clearly separated, it is very difficult to discriminate between public transport users and car riders. Second, in most urban networks, train lines are at times underground, where GPS signal is not available, and standard GSM methods cannot be applied.

3. Some elements of context

This section provides the necessary background for a better understanding of the rest of the paper. First, it presents the broad characteristics of the underground part of the transit system in the Paris metropolitan area (Section 3.1), along with the AFC system in operation. Second, detailed attention is given to the central segment of the RER A line, which serves as the basis for the subsequent developments. In particular, the impact of congestion on travel times and dwell times on this segment are analyzed (Section 3.2). Third, Section 3.3 provides an overview of the operation of GSM networks and Section 3.4 details the particularities of the underground operation of the Orange GSM network in the subway. Last (Section 3.5), the hourly flow rate of AFC cards is compared to that of mobile phones. Although the two data sources do not provide the same kind of information, the comparison is instructive.

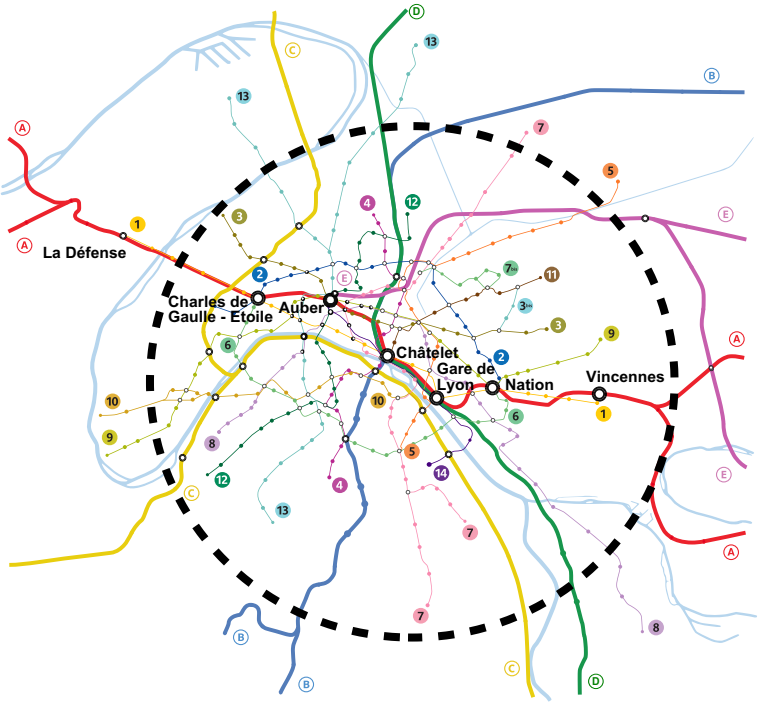

Fig. 1. The Paris subway. Most stations inside the dashed circle are located underground.

3.1. The Paris subway and its AFC system

The metropolitan transit system in Paris comprises in fact two systems. One is the Metro (short for Metropolitan). It runs in downtown Paris, and has 14 lines, numbered 1 through 14. The other is the RER (*Réseau Express Regional*, the Regional Express Network) which runs through both the city of Paris and the rest of the urban area, and consists of 5 lines, lettered A through E. In what follows, the underground part of this transit system will be referred to as the subway. The approximate boundary between the subway and the aboveground part of the transit system is shown on the map of the transit system (Fig. 1).

All stations in the subway are equipped with contactless AFC gates. According to the figures published by the transport authority, during an average working day, the vast majority (more than 90%) of the trips use AFC. With respect to AFC, the RER is a closed system, but the Metro is not. Metro users have to badge on entry, but not on exit. A Metro user transferring to the RER has to badge three times: first when entering the Metro, then when entering the RER, and finally when exiting the RER. RER users have to badge at their entrance and exit station. Users do not need to badge at the transfer station when they transfer between RER lines, for instance between RER A and RER B at Châtelet. As RER A and RER B are the two most trafficked lines, a significant proportion of trips have their AFC entry event at an RER B (or A) gate, and their AFC exit event occurring at a RER A (or B) gate. Those details about AFC operation in the subway will be of importance in Section 5, where travel times and origin–destination flows estimated from GSM data will be compared to AFC data.

3.2. Congestion in the RER A central segment

RER A is the most trafficked line of the Paris transit network, with more than one million trips during an average working day. The eastern part of the line has two branches and the western part has three. The arc between Châtelet and Auber is the most heavily loaded, carrying around 50,000 passengers per hour in each direction during the morning peak of an average working day. In this paper, the *RER A central segment* is the part of the RER A line situated between Vincennes and Charles de Gaulle-Etoile (Fig. 1). The targeted frequency during peak hours on the central segment is 30 trains per hour. But recurring delays due to congestion make this impossible to achieve. The flow rate of trains during the morning and evening peaks hardly reaches 26 trains per hour, so that the capacity actually provided is 10% lower than the nominal level.

RER A also suffers from a very high level of irregularity. Its quality of service is generally perceived as very low, with severe delays and discomfort. This section provides a quantitative insight into the phenomena at work, through the statistical analysis of two quantities: the travel time and the dwell time.

During all the working days of October and November 2011, we constructed the observed train timetables using data sourced from the RER A operator's website.[1] This website provides the same real time information that is displayed on the

[1] http://www.ratp.fr.

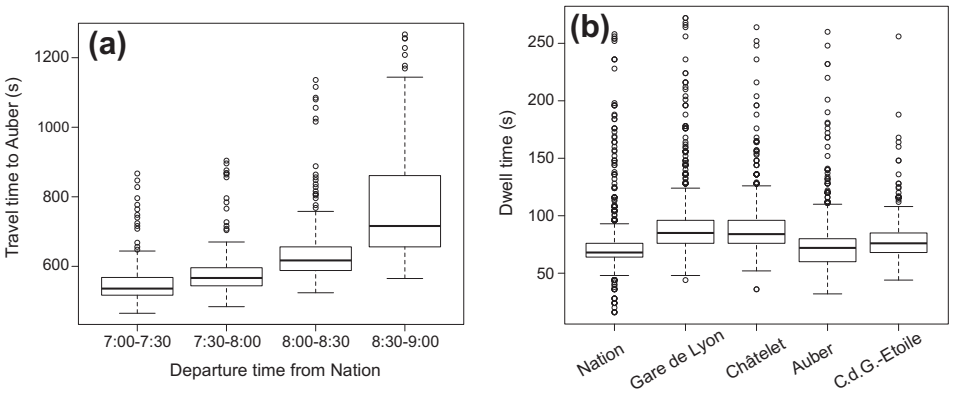

Fig. 2. Boxplots of the distribution of (a) travel times from Nation to Auber during the morning peak, and (b) dwell times for 5 stations in the central segment of RER A. Drawn from the real time information system of the operator, during working days in October and November 2011. By design, the travel time from Nation to Auber should be constant and equal to 540 s. The dwell time should remain below 60 s.

platforms, including the estimated time of arrival of the next train, and whether or not a train is standing at the platform. These data were recorded every 4 s during the study period. A train's arrival time at the station was estimated on the basis of the first instant the *train standing at platform* message appeared in the records. A train's departure time from station was estimated on the basis of the first instant the *train standing at platform* message disappeared from the records. This enabled us to construct a complete observed timetable for the line for the two months of October and November 2011.

Box-plots of the distribution of travel times between Nation and Auber for 30 min intervals within the morning peak are shown in Fig. 2a. Both the mean value and the dispersion around it increase with the time of departure from Nation. The situation during the morning peak is not homogeneous. Delays accumulate from 7:30 a.m. to 9:00 a.m., during the morning peak. The degradation of travel times gradually increases during the peak. It is much worse between 8:30 and 9:00 a.m. than it is between 7:00 and 7:30 a.m. According to the planned timetables, all trains should take 540 s to get from Nation to Auber. During the two observed monthes, between 8:30 a.m. and 9:00 a.m., all trains took longer than 540 s, and half of them more than 700 s. The situation is similar observations for all stations in the central segment of RER A.

Fig. 2b shows the distribution of dwell times during the morning peak, for 5 stations in the central segment of RER A. In the absence of dedicated devices (such as automated gates on platforms) dwell times are mostly influenced by (i) the flow of boarding passengers, (ii) the flow of alighting passengers, and (iii) the interactions between those two. The mean value and variability of the dwell time reach their maximum at Gare de Lyon and Châtelet (Fig. 2b). Both are major hubs in the transit system. At Châtelet, RER A is connected to RER B, the second busiest line in the transit network. Gare de Lyon is connected to the intercity high speed train network. Due to these connections, the flows of boarding and alighting passengers are substantially larger at Châtelet and Gare de Lyon than at the other stations.

The two plots in Fig. 2 show that the operation of the line is very poor, and that congestion effects mean the situation becomes worse and worse during the peaks, when demand exceeds supply. Indeed, trains and platforms are overcrowded, dwelltimes reach unacceptably high values, the operator cannot adhere to the transfer time target, the rate of train traffic decreases finally causing the trains and platforms to be even more overcrowded. This vicious circle is initially due to the differential between supply and demand, and it causes the quality of service to drop progressively during peak periods.

3.3. Operation of a GSM network

In order to properly assess the possibilities offered by GSM data, it is necessary to provide a basic understanding of GSM networks.

At the heart of a wireless GSM communication network is a set of Base Transceiver Stations (BTS). Each BTS is equipped with one or more antennas to communicate with mobile devices. The area covered by the antennas of a BTS is called a cell—hence the name cellular network. A mobile phone connects to the network by searching for cells in its immediate vicinity. For outdoor aerial cells, the mean horizontal radius of a cell depends on numerous parameters, including the antennas' height and local propagation conditions. It varies from a couple of hundred meters to tens of kilometers. A mobile phone frequently exchanges information with the closest antennas to keep connected to a BTS. When the phone moves, the BTS it is connected to may change.

BTS are grouped together and controlled by a Base Station Controller (BSC). The BSC are themselves grouped together and controlled by a Mobile Switching Center (MSC). The area covered by an MSC is designated as a Location Area (LA). An LA typically contains hundreds of cells. To operate the network efficiently, each MSC maintains a database of so called signaling events. The most common of these are:

HANDOVER: When cell change happens during a communication.
SMS: When a short message is sent or received.
IMSI ATTACH/DETACH: The phone is switched on or off.
LAU: i.e. a Location Area Update.

LAU events are related to the LA Update procedure that makes the network informed of the locations of phones. Phones are responsible for detecting LA codes. When a phone finds that the LA code differs from its last update, it sends a LAU-N event. Each phone also regularly reports its location by periodically sending LAU-P events. Like other signaling events, LAU events are logged in the MSC database. For each signaling event, the MSC database records a certain amount of data, including:

- The instant at which the event started (or ended), in milliseconds. The typical duration between the start and the end of an event is around 1 s. This is the time needed for the LAU request initiated by the phone to be processed by the network and finally acknowledged by the phone.
- The BSC the phone was connected to when the event started.
- The event status, either rejected or accepted. An LAU request may be rejected for bandwidth management reasons, partly because LAU traffic has a lower priority than voice and data. When a burst of LAU requests are emitted within a narrow time frame—as occurs when a crowded train enters a tunnel—some LAU requests can be rejected. When this happens, the rejected phones will emit a new LAU request a few tens of seconds later.
- The LA and Cell ID from which the event was sent.
- A Temporary Mobile Subscriber Identity (TMSI). This is a unique, random, and temporary number assigned to the phone. It changes a few times a day.

3.4. Operation of the Orange GSM network in the Paris subway

The principles presented in the previous section are valid for aboveground and underground networks alike. However, underground networks have particularities thanks to which GSM data offers unique possibilities. Those are explained here.

First, surface antennas cannot communicate with mobiles located underground. Also, the indoor propagation of electromagnetic waves differs significantly from outdoor. As a consequence, the underground operation of a GSM network requires dedicated antennas. Classical indoor antennas are used in underground stations for covering the platforms, whereas dedicated devices, such as radiating cables or waveguides, are used in tunnels. Second, it so happens that the Orange GSM network in the Paris subway consists of a unique dedicated LA, to which all underground antennas (i.e. indoor antennas in stations and tunnel antennas) belong. Thereafter, we will refer to this LA as the *the underground LA*.

During a typical trip in the subway, a switched on mobile phone may emit several signaling events, including LAU events. First, a few seconds after the phone has entered the underground LA, a LAU-N event is likely to be recorded in the MSC database of the underground LA. The phone informs the network of the change it has detected in its location area, from a aboveground LA to the underground LA. Then, during the trip underground, some signaling events may be emitted. If this is the case, they are recorded in the MSC database of the underground LA. Last, when the phone detects it has exited the underground LA, a LAU-N event is likely to be recorded in the MSC database of the surface antenna closest to the exit. Depending on local propagation conditions, that antenna may or may that which is geometrically the closest.

In the remaining of the paper, the term *GSM data* is used and should be understood as a shortcut for *a set of signaling events*.

3.5. Comparison between GSM data and AFC data

As already mentioned, GSM data and AFC data do not offer the same possibilities. AFC data only measures entries, and sometimes exits and transfers, whereas GSM data measures the actual space–time trajectories of passengers, but with a significant probability not to be detected at all in some stations (this particular issue is discussed in detail in Section 4.5). In addition, every passenger does not carry a phone operated by Orange. These two data sources are therefore quite different. The objective of this section is to assess to what extent they can be compared.

To this end, the variations in the hourly flow rate of mobile phones and of AFC cards, averaged over 15 min intervals during a single day (13 October 2011), have been plotted Fig. 3, for the 6 stations in the central segment of RER A. More precisely, for a given station, the plot is the flow rate of distinct mobile phones (or distinct AFC cards) detected underground during consecutive 15 min interval. The two datasets clearly do not provide the same information. On the one hand, in a given station, the flow rate of mobile phones is the number of distinct mobile phones that triggered at least one LAU event in the underground LA within a 15 min interval. The phones may be aboard trains, on the platforms, or in pedestrian tunnels leading to the platforms. On the other hand, the flow rate of AFC cards is the number of cards that enter (or exit) the station, from (or to) the outside or the Metro. It should be remembered that connecting passengers from one RER line do not badge when connecting with another RER line.

Despite the differences between the two datasets, the plots in Fig. 3 exhibit interesting features. The morning peak and the evening peak clearly emerge. Up to a certain multiplicative constant—which depends on the station and on the peak period—the two flows vary simultaneously. At Nation (Fig. 3b) and Châtelet (Fig. 3d), the fact that they are almost equal is a pure coïnci-

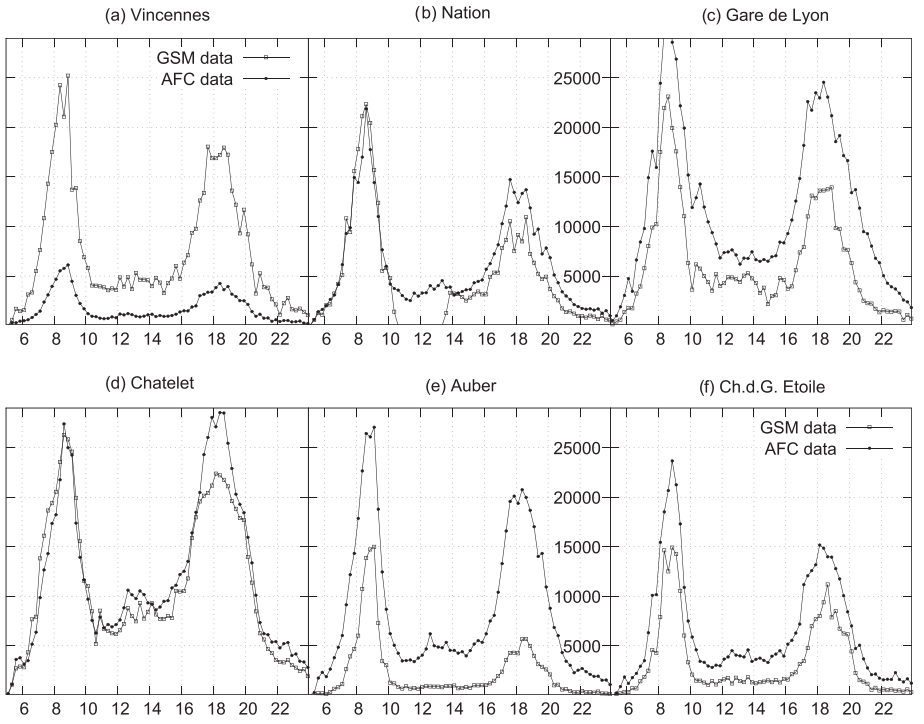

Fig. 3. Variations of the hourly flow rate of mobile phones and AFC cards, averaged over 15 min intervals during the day. Data for 13 October 2011.

dence: among other factors, the observed flows depend on the market share of both systems in the population of passengers. The gap around noon at Nation is due to a malfunction of the GSM network that lasted a couple of hours. At Vincennes (Fig. 3a), the flow of phones by far exceeds the flow of AFC cards. This is because the station at Vincennes is located in a short tunnel. To the west of Vincennes for approximately 1 km between Vincennes and Nation, the RER is aboveground. Hence, almost all phones aboard trains traveling in both directions are likely to trigger a LAU event there. At the same time, Vincennes is not a major connecting hub. The number of passengers boarding or alighting at Vincennes is small in comparison to the number remaining on the trains. The converse is true at Auber (Fig. 3e), where the flow of AFC cards by far exceeds the flow of phones.

4. Estimation of trains' level of occupancy

The occupancy level is defined as the ratio between the number of passengers aboard the train and its capacity. It is a major determinant of transit QoS, and not easy to estimate. The purpose of this section is to assess whether or not GSM data is able to provide accurate estimates of the number of passengers on a train. To this end, we have used two datasets. The first contains field observations made on 7 April 2011, in the central segment of RER A (Section 4.1). The second contains the records of LAU events in the underground LA, on the same day (Section 4.2). GSM data does not allow the complete trajectory of each phone to be tracked. As suggested in Section 3.4, along the same underground route, the number and location of signaling events emitted by different phones may vary significantly. Section 4.3 examines this in greater depth, and compares train trajectories (from field observations) to mobile phones trajectories (from GSM data). We have shown that, between two consecutives stations, the GSM data allows us to estimate the number of phones per train, up to a multiplicative constant (Section 4.4). Under some assumptions, this constant, that differs for each pair of consecutive stations, can be inferred (Section 4.5). Then, using the train capacities obtained from field observations, it is possible to estimate the interstation occupancy levels of the trains. These estimates are then compared to field observations (Section 4.6).

4.1. Field observations

Three different types of trains are operated on the RER A line. The number of seats varies in $\{432; 600; 1056\}$. The total capacity varies in $\{1760; 1900; 2580\}$. In each case, the nominal passenger density is 4 passengers/sqm. Between 7:15 a.m. and 9:03 a.m., for every train traveling east–west, and for each station between Vincennes and C.d.G-Etoile, 6 observers located on the platform recorded:

- the train type.
- an estimation of the occupancy level in the coach that stops in front of the observer, on the following 4-degree qualitative scale:

low: seats are available, no standing passengers.
medium: no seat is available, some standing passengers.
high: a large number of standing passengers.
huge: standing up passengers are pressed up against the doors.

- the instants at which several events occurred: the complete stop of the train, the doors opening, the doors closing, and the departure of the train.

4.2. GSM data

The whole set of LAU events recorded on 7 April 2011 was extracted from the MSC database of the underground LA. In what follows, we will consider only the events recorded at the underground BTS at Vincennes, Nation, Gare de Lyon, Châtelet, Auber and C.d.G.-Etoile. Essentially, this dataset—denoted by L—contains (m, h, s) triples, where m is the (temporary) unique identifier of a phone, h is an instant (the number of milliseconds elapsed from midnight), and s is a station on the central segment of RER A. L contains about 1 million records, and around 293,000 distinct phone identifiers. Hence, each phone was detected 3.65 times on average. 32% of the phones were detected only once; 63% of the phones were detected three times or less, and 4% were detected 10 times or more.

4.3. Train and phone trajectories

The train trajectories observed during the field survey are plotted in the time–space diagram of Fig. 4a, for departure times from Vincennes in the range [7:15 a.m.; 7:55 a.m.], along with the few 5-trajectories contained in L. The term 5-trajectory is used here to denote a subset T of L such that all triples in T share the same value for m, with 5 consecutive stations in the central segment appearing at least once, in order. As one can expect from the figures given in Section 4.2, 2-trajectories are far numerous than 5-trajectories. Fig. 4b shows the 2-trajectories observed in the GSM data between Vincennes and Nation, and between Nation and Gare de Lyon, for departure times from Vincennes in the range [7:30 a.m.; 7:44 a.m.]. Each train is clearly correlated with a burst of 2-trajectories in GSM data. However, the exact location at which LAU events occurred are unknown. We only know the ID of the antenna, and the station where this antenna is located. Between Vincennes and Nation the burst of 2-trajectories seems to precede the train because a lot of phones trigger a LAU event when entering the tunnel upstream of Vincennes, before the train stops at the station. Similarly, the LAU events observed before Nation are triggered in the tunnel between Vincennes and Nation, before the train stops at Nation. Also, it should be noted that phone trajectories do not always match trains trajectories. Some 2-trajectories look as though phones were jumping from one train to another. This is indeed the case. RER A trains do not circulate on all the possible routes between the 3 western branches and the 2 eastern branches. For instance passengers traveling from the northeast to the northwest have to alight in one of the stations of the central segment and wait for a connection.

4.4. Number of 2-trajectories per train

Fig. 5 is an alternative representation of the trajectories plotted Fig. 4b. A cell-phone 2-trajectory between Vincennes and Nation is represented by a point in a two-dimensional plane. The abscissa is the instant at which a phone was detected at Vincennes. The ordinate is the instant, if one exists, at which the phone was detected at Nation after having being detected at Vincennes. In other words, a point appears in this plane if there exists $l_1 = (m_1, h_1, s_1) \in L$ and $l_2 = (m_2, h_2, s_2) \in L$ such that

$$(m_1 = m_2) \wedge (h_1 < h_2) \wedge (s_1 = \mathit{Vincennes}) \wedge (s_2 = \mathit{Nation})$$

For a given phone m, if more than one (l_1, l_2) pair satisfy the above conditions, a single point is arbitrarily selected: one such that $h_2 - h_1$ is minimal. A shaded square dot appears in Fig. 5 if at least one point is contained in a 10 s square cell of the plane. The shade of the dot varies with the flow rate of phones: a light-gray dot indicates less than 1 phone per second, a dark-gray dot indicates less than 2 phones per second and a black dot indicates more than 2 phones per second.

The train trajectories from field observations are represented by line segments. The circled dot on the diagonal indicates the instant of departure from Vincennes, as recorded by the observer on the platform. The cross at the other extremity of the segment indicates the departure time from Nation, as recorded by the observer at Nation.

Standard segmentation techniques (gradient based directed edge detection and region growing) are applied to partition the density map of cell-phones 2-trajectories into distinguished, rectangle shaped, sets of 2-trajectories, one per train, and hence to associate a number of 2-trajectories to each train. This number clearly depends on the probability of a phone having been detected at two consecutive stations. For reasons suggested in Section 3.5, this probability varies with the pair of stations under consideration, and with the period of the day. Now, if we:

(i) let $P(s)$ denote the probability of a phone of being spotted at station s;
(ii) let s^- (resp. s^+) denote the station downstream (resp. upstream) s;
(iii) assume $P(s^-), P(s)$ and $P(s^+)$ are independent.

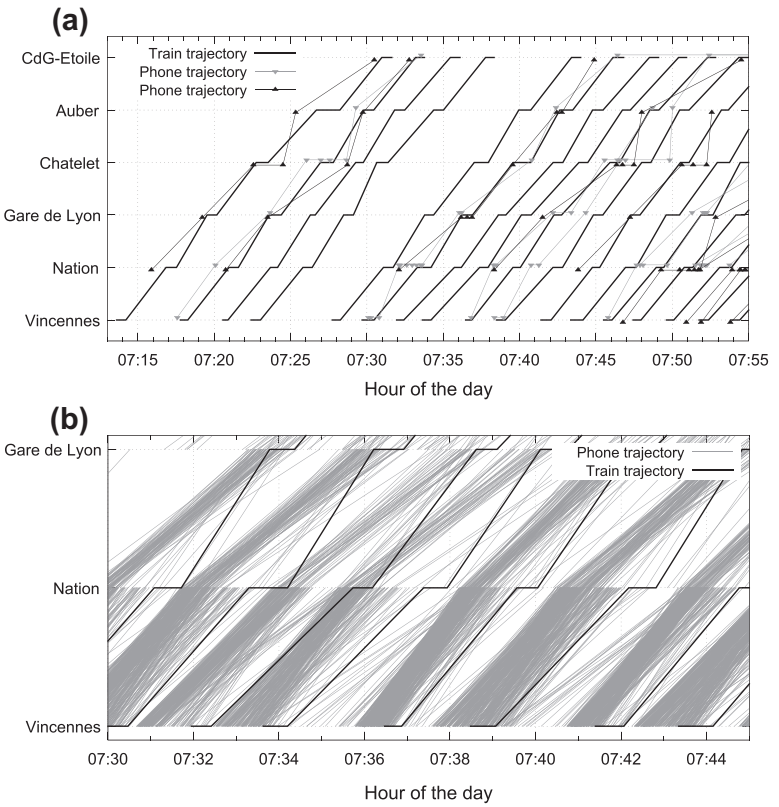

Fig. 4. Time space diagrams of underground LAU events and trains in the central segment of the RER A, on 7 April 2011. (a) Trajectories of some of the few cell-phones that triggered at least one LAU at 5 successive stations. (b) Trajectories of the cell-phones that triggered at least one LAU at 2 successive stations, between Vincennes and Gare de Lyon.

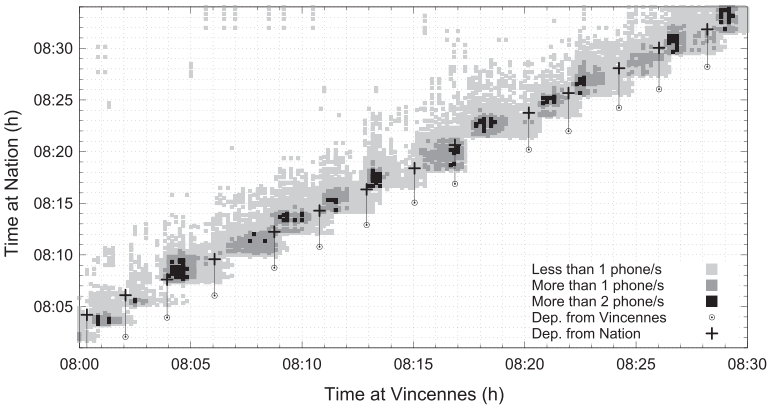

Fig. 5. Density map of the 2-trajectories, in a time-time diagram, between Vincennes and Nation, between 8:00 a.m. and 8:30 a.m., 7 April 2011.

Then the number of phones $N_{t,s}$ aboard a train t with $n_{t,s}$ 2-trajectories is $\frac{n_{t,s}}{P(s)P(s^-)}$. The coming section presents an estimator for $P(s)$.

4.5. Probability of a phone being detected during an underground trip

With the notations and assumptions introduced in the previous section, especially hypothesis (iii), the following holds:

$$P(s) = P(s \mid s^+ \cap s^-) = \frac{P(s^+ \cap s \cap s^-)}{P(s^+ \cap s^-)}$$

In other words, the probability $P(s)$ that a phone traveling from s^+ to s^- will be observed at station s, conditionally to the fact that it has been observed at s^+ and s^-, is the ratio between (i) the probability this phone has been observed at the three stations s^+, s and s^-, and (ii) the probability it has been observed at s^+ and at s^-.

This provides us with a simple estimator of $P(s)$. If, for example, we wish to estimate the probability that a phone traveling from Nation to Châtelet will be observed at Gare de Lyon, we simply need to count the phones observed at those three stations, then count the number of phones observed at Nation and at Châtelet.

Generally speaking, if $n_{s^+\cap s\cap s^-}$ denotes the number of phones observed at all three stations s^+, s and s^-, and if $n_{s^+\cap s^-}$ denotes the number of phones observed at both stations s^+ and s^-, then

$$f = \frac{n_{s^+\cap s\cap s^-}}{n_{s^+\cap s^-}}$$

is an estimator of the probability $P(s)$ of a phone being observed at station s. Additionally, the 95% confidence interval around f is given by:

$$P_{95}(s) = f \pm 1.96\sqrt{\frac{f(1-f)}{n_{s^+\cap s^-}}} \tag{1}$$

The P_{95} estimator depends to a great degree on s and on the direction, as well as the time of the day. Variations of P_{95} are plotted Fig. 6a for all 30 min intervals in the range $[7a.m. : 12]$ for the 4 stations Nation, Gare de Lyon, Châtelet and Auber, in the east–west direction.

At Nation, P_{95} remains almost equal to 1 throughout the morning, with a narrow confidence interval. All phones traveling from Vincennes to Gare de Lyon in the morning are likely to be observed at Nation. P_{95} for Châtelet increases slightly, from 0.4 and 0.5, during the morning. For Gare de Lyon, P_{95} is in the range $[0.4 : 0.6]$ during the morning peak, and in the range $[0.2 : 0.4]$ outside the morning peak. Auber behaves almost like Gare de Lyon, with a noticeable plateau during the morning peak.

Things are different during the evening in the west–east direction, for the same 4 stations (Fig. 6b). Nation remains constant, at around 0.4. Auber is almost constant, at around 0.2. Châtelet seems to decrease, from 0.7 to 0.4 during the evening peak as does Gare de Lyon, from 0.4 to 0.2.

4.6. Train occupancy levels

As seen in Section 4.4, between any pair of consecutive stations, each train can be associated with a set of 2-trajectories. Let denote n_t the number of 2-trajectories assigned to the train t. Using the P_{95} estimator defined in Section 4.5, the level of occupancy of the train t, between the pair of stations s and s^-, denoted $O_{s,t}$, is estimated by:

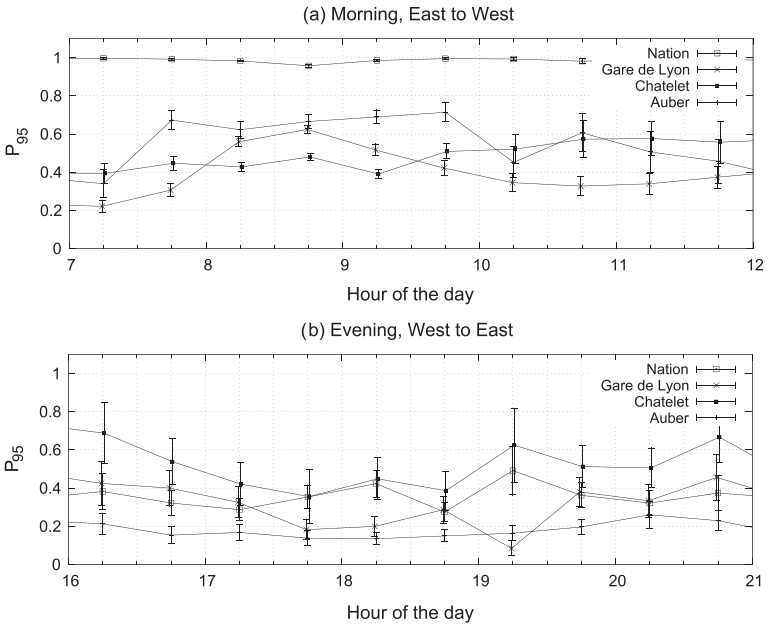

Fig. 6. Estimated probability for a phone to be observed at station s, using the P_{95} estimator (Eq. 1). The y-error bars indicate the boundaries of the 95% confidence interval. (a) In the morning, from east to west. (b) In the evening, from west to east.

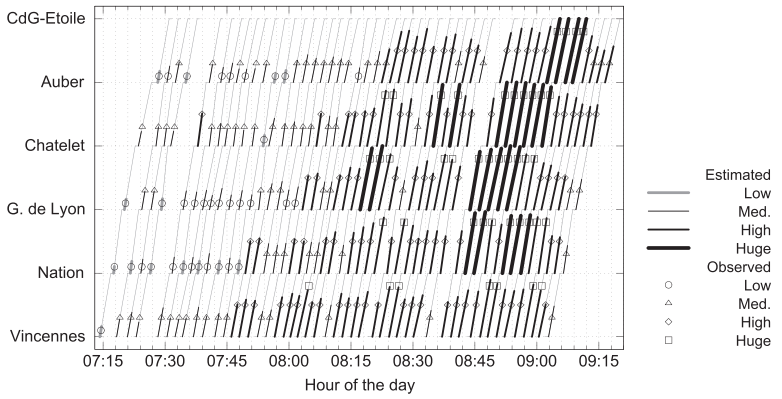

Fig. 7. Train occupancy levels estimated from GSM data (as defined by Eq. 2), compared to field observations. Estimations from GSM data are plotted with line segments of varying length, thickness and shade. Field observations are plotted with symbols. See Section 4.6 for detailed comments.

$$O_{s,t} = \frac{n_t}{\alpha K_t P_{95}(s) P_{95}(s^-)} \tag{2}$$

where K_t is the capacity of train t, and α the network market share[2] of the mobile phone operator Orange.

The estimated train occupancy levels (Eq. 2) between Vincennes and C.d.G.-Etoile is plotted in Fig. 7. The length of the line segment for each train–station pair is proportional to the estimated level of occupancy. The shade and thickness indicates the level of comfort. A thick gray segment indicates that the estimated number of passengers is below the number of seats in the train. A thick black segment indicates that the estimated number of passengers exceeds 95% of the train's total capacity. Ratios above 100% are cut off for readability. Field observations are plotted also in Fig. 7 using symbols. A circle (resp. triangle, diamond, square) indicates a low (resp. medium, high, huge) occupancy level. The train occupancy levels estimated from GSM data are in good accordance with field observations (Section 4.1). More than 80% of the level of occupancy observed during the field survey agree with the estimations. When they do not, the discrepancy is small and may be well explained by threshold effects and uncertainties in observers' ability to estimate the occupancy level of a whole train. For instance, at Gare de Lyon, between 7:30 a.m and 08:03 a.m., 10 field observations are ranked *low* while GSM data indicates *medium*.

5. Comparison with AFC data

The experiments and results described in Section 4 relate to our first attempt at using GSM data, on 7 April 2011. The results are encouraging, and make it possible to estimate interstation travel times (using 2-trajectories) and train occupancy levels. We intentionally restricted ourselves to signaling events recorded only in the underground LA, for two reasons. First because the objectives of this preliminary experiment did not require additional data, and second because GSM data are huge datasets, and manipulating gigabytes is error-prone and time-consuming. Starting with data from a single LA allowed us to gradually acquire the necessary skills to manage all the steps of data processing.

However, this restriction to the underground LA prevented us from capturing exits from the underground, because, as we have already seen, when a phone exits the underground, it is likely to trigger an LAU event that is recorded by the above-ground LA the phone is entering. In order to explore the possibilities offered by this additional information, a larger scale experiment was conducted on 13 October 2011. Its purpose was to estimate travel times and origin–destination (OD) flows, and compare the results to estimates from AFC data.

This section presents the datasets used during this second experiment. The AFC dataset is analyzed in Section 5.1 and the GSM dataset is analyzed in Section 5.2. The travel times computed from AFC data and GSM data are then compared (Section 5.3), and so are the OD flows (Section 5.4).

5.1. The AFC dataset

We used AFC records for the whole of the RER A line, for 13 October 2011. This dataset contained 1.9 M rows. Each row contained a anonymized card identifier, a timestamp, and the identifier of the station where the event (either an entry or an exit) occurred. The AFC dataset contained around 684,000 distinct card identifiers. On average, there were 2.71 events per card. The distribution of cards over the number of events during the day is plotted Fig. 8. Most AFC cards were detected either twice during the day (commuters connecting to/from another RER line) or 4 times during the day (commuters using RER A in

[2] The network market share includes the traffic induced by Mobile Virtual Network Operators. It was equal to 46.6% in 2011, according to http://goo.gl/1HSI7.

the morning and in the evening). Surprisingly, the figures for 1 and 3 events during the day were quite high. In theory, this could be explained by the fact that the operator sometimes leaves the AFC gates open, but this was not the case on the day of the survey. Another possible explanation is that some passengers take the RER in one direction and the metro in the other direction. For example, Metro 1, which is parallel to RER A, is a viable alternative for traveling to La Défense in the morning from the east of Paris: the travel time is higher but passengers have a good chance to find a seat, whereas the RER is already very loaded when it enters Paris from the east. In the evening, there is no point in using the metro: the large number of passengers boarding at La Défense means the metro is not much comfortable than the RER. This may explain the structural asymmetry in the number of validations. Of course, this attempt at an explanation should be confirmed by statistical analysis, but this is unfortunately not an easy task with the data at hand and without an additional survey, and falls outside the scope of this paper.

5.2. The GSM dataset

The GSM dataset contains the signaling events recorded on 13 October 2011, for 49 LAs. One of them is the underground LA and the 48 others are aboveground LAs that cover a large band around the central segment of the RER A. The aboveground LAs comprise more than 8,000 aerial cells, covering all the subway entrances in the broad vicinity of the central segment of the RER A. The GSM dataset—denoted by E—contains nearly 29 M rows, and a total of 2.9 M distinct phone identifiers, hence an average of 10 events per phone. Each element e of E is a 5-tuple (m, c, h, l, l'), where m is the mobile identifier, c the identifier of the cell where the event was triggered, h the event's timestamp, l the identifier of the LA the cell c belongs to, and l' the identifier of the LA the phone was previously located in. As explained in the introduction of this section, using data from the aboveground LAs allows to detect the phones at the entrance and at the exit of the underground LA.

For any 5-tuple $e = (m, c, h, l, l') \in E$, if l is the identifier of the underground LA and l' the identifier of an aboveground LA, e is referred to as an *in* event. Conversely, if l is the identifier of an aboveground LA and l' the identifier of the underground LA, e is referred to as an *out* event. The distribution of mobile phones identifiers over the number of events during the day is plotted in Fig. 8, together with the distribution of *in* and *out* events. The distributions of *in* and *out* events, for more than 1 event during the day, are almost equal. The dataset efficiently captures the exits of mobile phones that enter the underground LA more than once during the day. The figures differ for the population of phones that did a single *in* (resp. *out*) event during the day. The number of entries exceeds the number of exits. This can be explained by the fact that the dataset captures all the entries into the underground LA, while some exits, those out of the coverage of the 48 aboveground LAs, are not detected.

5.3. Travel times

Both datasets contain enough data to estimate the distributions of travel times between any pair of stations in the central segment. The travel times measured with AFC data are gate-to-gate, so they include the walking time from the gate to the platform, the waiting time on the platform, and the walking time from the platform to the gate. The travel times measured from GSM data are those of 2-trajectories, as defined in Section 4.4. They approximate train travel times. Fig. 9 illustrates the density of the distribution of travel times, in the morning, from Vincennes to C.d.G. Etoile (Fig. 9a), and from Nation to Auber (Fig. 9b).

In both cases, the density are very similar, up to a shift of GSM data travel times by a 3 min constant. The difference is explained by walking times and waiting times captured by AFC data. In both figures, the dotted line represents the expected travel time, according to the operator timetables. More than 80% of the users experience a travel time that is higher than the target. More than 40% experience a travel time that exceeds the target by 20%. Although not plotted here, the change in the

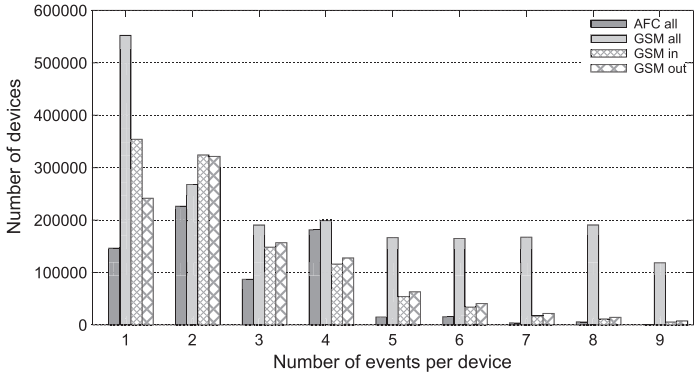

Fig. 8. Distribution of devices (either AFC cards or mobile phones) over the number of events per device, during the day. For AFC data (AFC all): an event is either an entry or an exit. For GSM data: GSM all is related to the total number of signaling events per mobile phone during the day; GSM in corresponds to entries in the underground LA; GSM out corresponds to exits from the underground LA.

distribution of travel times over 30 min time slices during the morning peak follows a very similar pattern to that observed for a two-month observation period (see Fig. 2a).

5.4. Origin–destination flows

The purpose of this section is to estimate morning OD flows from stations in the central segment to La Défense. La Défense is the largest business district in the Paris metropolitan area. The RER A station at La Défense is the next station to the west after C.d.G. Etoile. The OD flows estimated from the AFC data correspond to trips with a first spot in the morning in the central segment, and a second spot in the morning at La Défense. Therefore, trips that used the RER to connect with RER A were not captured, but trips that used the Metro to connect with RER A were. The OD flows estimated from AFC data are plotted in Fig. 10. The peak period extends from 8:00 a.m. to 10:00 a.m., with a hyperpeak between 8:30 a.m. and 9:30 a.m. Most trips are originated at Auber, with nearly 40% of the trips during the hyperpeak.

To compare OD flows drawn from GSM data to OD flows from AFC data, some additional data is needed. First, for any mobile phone that has triggered both an *in* event and an *out* event in the morning, we need to know if the destination is La Défense, i.e. if the *out* event was triggered in an aboveground cell in the vicinity of La Défense.

Second, for each such mobile whose destination has been assigned to La Défense, we need to deduce if the trip should be assigned to RER A, and if so which station in the central segment is likely to be the RER origin of the trip. This is because the *in* event may have been triggered not only in the central segment, but also in one of the Metro station where connection with RER A is possible. To this end, we used the GSM dataset to define a naïve assignment model, based on the travel time likelihood that a trip has connected with RER in the central segment.

5.4.1. Th aboveground vicinity of La Défense

From the AFC dataset, more than 95% of the phones that triggered an *out* event during the morning after having been detected underground at La Défense did so within a 3 min time frame. We shall now assume that the set of aboveground cells where such an *out* event occurred defines the aboveground vicinity of La Défense. Hence, every phone that triggered an *out* event during the morning in the aboveground vicinity of La Défense, whether or not it was detected underground before, has its destination assigned to La Défense.

5.4.2. Travel time between underground stations

If (u, v) is a pair of underground stations, let

$$T_{u,v} = \{(o,d) \in E^2, o.c \in u, d.c \in v, o.m = d.m, o.h < d.h\}$$

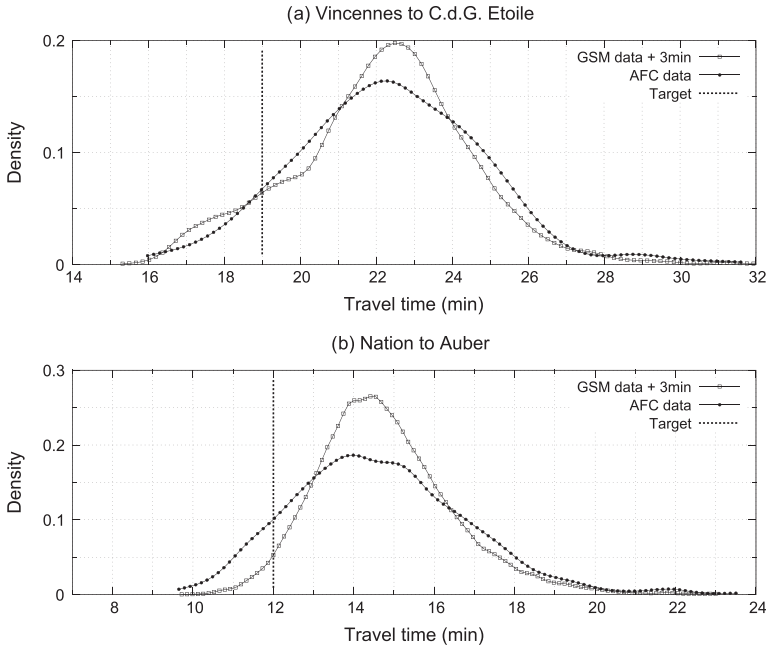

Fig. 9. Density of the distribution of travel times, in the morning. (a) From Vincennes to C.d.G. Etoile. (b) From Nation to Auber. Thick line: density of the distribution of travel times computed from AFC data. Thin line: density of the distribution of travel times computed from GSM data, shifted by 3 min. Data for 13 October 2011.

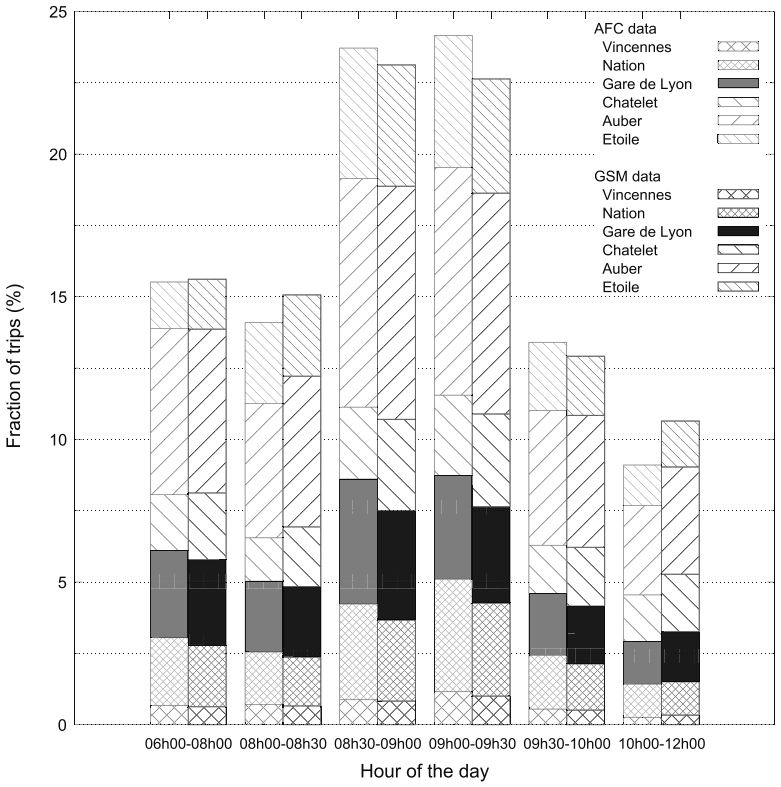

Fig. 10. Distributions of OD trips to La Défense, for six origins between Vincennes and C.d.G. Etoile, and six time slices in the morning. For each time slice, the distribution inferred from AFC (resp. GSM) data is plotted on the left (resp. right) hand side.

denote the set of 2-trajectories between u and v. $T_{u,v}$ contains pairs of events in E that are triggered by the same phone spotted first in a cell located in station u, and later on in a cell located in station v. It $T_{u,v}$ is not empty, then we define the travel time from u to v, denoted $t_{u,v,}(h)$, for any departure time h from u, as

$$t_{u,v}(h) = d.h - o.h,\ (o,d) = \begin{cases} \arg\min_{t \in T^{h^+}_{u,v}} t.d.h & \text{if } T^{h^+}_{u,v} \neq \varnothing \\ \arg\max_{t \in T_{u,v}} t.o.h & \text{otherwise} \end{cases}$$

where

$$T^{h^+}_{u,v} = \{(o,d) \in T_{u,v}, h < o.h\}.$$

5.4.3. A naïve assignment model

Let m be a mobile phone that has triggered (i) an *out* event $d \in E$ during the morning in the vicinity of La Défense, and (ii) an *in* event before d. Let $o \in E$ denote its last *in* event before d. The travel time from o to d is $t_m = d.h - o.h$. With the hypothesis that the trip included a connection in one of the stations of the central segment, there is a station s in the central segment that minimizes the difference between t_m and the travel time from o to d when connecting at s, that is

$$s = \arg\min_s \left\{ \left(t_m - [t_{o,s}(o.h) + t_{s,d}(o.h + t_{o,s}(o.h))]\right)^2 \right\}$$

Hence, if the relative difference between t_m and the travel time of the route passing through s exceeds a arbitrarily chosen threshold (set to 10%), the hypothesis is rejected. Otherwise, the trip is assigned to the OD pair (s,d). To soften this all-or-nothing assignment, in its place we used a logit discrete choice model among viable alternatives.

The resulting assignment is plotted in Fig. 10. For each time slice in the morning, the total number of phones assigned to La Défense from one of the stations in the central segment is slightly over 50% of the number of AFC cards. On the whole, during the morning, the ratio is around 52%, slightly above the network market share of the telecom operator. Several reasons may explain the difference, including: (i) the market share of Orange in the population of commuters to La Défense may differ from the country-wide market share; (ii) GSM data capture a fraction of trips that are not captured by AFC data (e.g. occasional business trips to La Défense). For each time slice in the morning, the proportion of phones assigned to each origin is remarkably close to that of AFC cards, although slight differences exist: the total volume of phones assigned during the first

half of the hyperpeak exceeds the volume assigned during the second half; the converse is true for AFC data. Also, the fraction of trips with an origin assigned to Châtelet using GSM data is quite high compared to the AFC data.

6. Conclusion

The methods described in this paper show the potential of mobile phone data to estimate QoS in transit. Some points are worth noticing.

First, the mobile phone data we used consists of signaling events that are recorded by the telecom operator for the normal operation of its network. The signaling procedures and events are GSM standards. They are not proprietary or specific to Orange, nor do they require, from a technical point of view, dedicated data acquisition resources.

Second, various analyses can be performed which do not require any data from the transit operator's side. It has been shown throughout the paper that train occupancy levels, travel times, and origin–destination flows can be estimated at a very fine-grain level, and with a promising level of accuracy. In some cases where the existing QoS indicators are a bone of contention between the operator and the users, transport authorities could benefit from having at their disposal alternative means to estimate the QoS.

Third, mobile phone data allows passengers to be tracked throughout their trip, whereas AFC only contains entries and, in some cases, exits. Mobile phone data allows both the supply and the demand side can be studied simultaneously: this is much more complicated with AFC data. This opens up the possibility of investigating poorly understood transient phenomena (e.g. users behavior in the event of occasional perturbations), or subtle recurring effects in transit systems, such as the influence of occupancy levels on route choices. Significant progress could be achieved in the design and calibration of a new generation of more dynamic transit assignment models.

Fourth, the market penetration of mobile phones among the population is known. However, it probably depends on variables that are correlated with the place of work and residence. Moreover, the market share of one phone company (in our case Orange) may also depend on some other, unobserved factors. These factors introduce biases which may be difficult to measure and correct.

Fifth, the comparison of mobile phone data to AFC data and field surveys shows a promising level of consistency: due to the biases discussed above, converting mobile phone data into passenger flows is not straightforward; but the distributions of travel times, levels of occupancy or OD flow inferred from mobile phone data are consistent with other data sources.

In any case, the preliminary results presented in this paper are very encouraging. Operators and transport authorities could benefit from new means of monitoring the system they are in charge of. Users could benefit from richer, real-time information.

Acknowledgments

The authors would like to thank Orange and STIF for supporting this research. In particular, this research is funded by the Chaire Ecole des Ponts ParisTech – STIF on public transport modelling and economics. The authors would also like sincerely to thank anonymous reviewers, whose fruitful comments considerably helped in improving the paper. The usual disclaimers apply.

References

Bertini, R., El-Geneidy, A., 2003. Generating transit performance measures with archived data. Transportation Research Record: Journal of the Transportation Research Board 1841, 109–119.

El-Geneidy, A., Horning, J., Krizek, K., 2011. Analyzing transit service reliability using detailed data from automatic vehicular locator systems. Journal of Advanced Transportation 45 (1), 66–79.

Feng, W., Figliozzi, M., 2011. Using archived avl/apc bus data to identify spatial–temporal causes of bus bunching. In: Proc. of the 90th Transportation Research Board Annual Meeting.

Friedrich, M., Immisch, K., Jehlicka, P., Otterstätter, T., Schlaich, J., 2011. Generating origin-destination matrices from mobile phone trajectories. Transportation Research Record: Journal of the Transportation Research Board 2196, 93–101.

Frumin, M., 2010. Automatic data for applied railway management: passenger demand, service quality measurement, and tactical planning on the London overground network. Ph.D. thesis, Massachusetts Institute of Technology.

Furth, P.G., Hemily, B., Muller, T.H.J., Strathman, J.G., 2006. Tcrp report 113 – using archived avl-apc data to improve transit performance and management. Tech. rep., Transit Cooperative Research Program.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. TTransportation Research Part B: Methodological 46 (9), 1097–1122.

Lind, G., Lindkvist, A., 2006. Optis – optimised traffic in Sweden. Tech. rep., Movea Traffic Consultancy Ltd.

Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M., 2011. Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transportation Research Record: Journal of the Transportation Research Board 2263, 140–150.

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., Rudenko, S., 2009. Application of entry-only automated fare collection (afc) system data to infer ridership, rider destinations, unlinked trips, and passenger miles. Transportation Research Record: Journal of the Transportation Research Board 2110, 128–136.

TCQSM, 2003. Transit Capacity and Quality of Service Manual, second ed. Tech. rep., Transit Cooperative Research Program.

Valerio, D., 2009. Road traffic information from cellular network signaling. Tech. rep., Forschungszentrum Telekommunikation Wien.

Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., September 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: Proc. of the 13th International IEEE Conference on Intelligent Transportation Systems, pp. 318 –323.