Combined use of Smartcard and Wifi detections to estimate real-time operational information of a public transport system

Pablo Guajardo · Juan C. Herrera

Keywords: WiFi scanners · smartcard data · ITS · OD matrix

1 Introduction

Several public transport systems (PTS) around the world use smartcards (SC) for the payment of fares when entering the system. Each time a system user validates the card, data on the time and place of validation and the card’s unique identifier are usually stored. Different authors have developed new methods-to-process and uses-of SC data for public transport. The reader is referred to a survey of this literature reported by Pelletier et al. (2011) to check studies on the use of SC data in public transport systems. Munizaga and Palma (2012) constructed an OD matrix using data generated by smartcard users in Santiago, Chile. They estimated the destination for each user validation from the place where the immediately following validation was made, based on certain assumptions regarding the time of day of the trip, the time elapsed between the two validations and other data. Munizaga et al. (2014) showed that this methodology correctly estimated the origin bus stop in 98.9% of cases and the destination bus stop in 84.2% of cases. More recently, Ma et al (2017) inferred several pieces of information, such as residence location, workplace and departure time of commuters in Beijing, using SC data.

On the other hand, Wi-Fi scanners (WS) can capture smartphone signals sent by users as they enter a detection zone around the scanner. If enough scanners are deployed on the network, the movement of device on the network can be observed. Danalet et al. (2012, 2014) trace the movement of pedestrians on a university campus while Abedi et al. (2014) analyze the human traffic flows within an office, detecting how certain people do not move about as independent entities but rather as
a group. Similar phenomena could be studied in the context of a PTS. In Abedi et al. (2015), the authors identify whether a given device is being used by a pedestrian, a jogger or a cyclist on the basis of its trip time. Other studies in the literature have used Bluetooth scanners instead of WS to determine vehicle speeds and travel times (Abbott-Jard et al., 2013) and OD matrices (Barceló et al., 2012, Michau et al., 2013), among other phenomena.

SC data and WS data can be combined, thus multiplying their potential for use by transport analysts and planners in developing and evaluating a series of useful indicators on system operation and user behaviour. The present study proposes a methodology to capture this potential and reports on a pilot experiment to test it in the real-world context.

2 Problem description

We consider a closed-type PTS such as a Metro or a BRT in which all users validate their entry into the network by tapping in with a smartcard at a card reader that stores the time and place of the validation. We assume a linear (corridor-like) network, while the methodology could be easily extended to consider grid networks as well. Each stop or station has its own WS whose detection area covers the station in its entirety. The PTS is assumed to be isolated from other transport modes and blocks the signals of devices that are external to it; if this is not in fact the case, the data would have to be filtered before our methodology could be applied.

We propose a methodology to combine both SC data and WS data to estimate in real time the OD matrix and arc load (in terms of number of passengers both), and other key PTS indicators. Additionally, we discuss on the variations in the experimentally observed smartphone penetration among the public by time of day and over geographic space, and the consequences of these variations for our methodology.

3 Methodology

3.1 Estimation of OD matrix and arc load

The OD matrix and arc load are estimated in two main stages. The first stage processes WS data relating to the starting, updating and ending of trips by devices while the second stage expands the trip sample represented by this information into real users using the SC data.

A device \( n \) first detected at station \( i \) is considered to have initiated its trip at station \( i \) and instant \( t \) only if it is detected at station \( j \) (\( j \neq i \)) at a later time \( t' > t \). If no new detections of device \( n \) are obtained at other stations after a while, device \( n \) is discarded as a user.
New detections at stations $j$ of a device that was first detected at station $i$ ($j \neq i$) are used to update the location of the device and to increase the load of all the arcs connecting stations $i$ and $j$.

If a device $n$ has not been detected by any WS for more than a certain lapse of time $L$, it is assumed that the device exited the network at the station and time instant of its last detection. By ending the trip of device $n$, the OD matrix is update accordingly.

Parameter $L$ must be big enough so that a device still travelling on the network will not be treated as having ended its trip before later detections updating its position in the system can be received, yet small enough that a device which has left the network but returns to it later will be recognized as having ended one trip and started another.

3.2 Estimation of weighting factors

In this stage, the WS data are combined with the SC data in order to expand the sample obtained with WS. The value of weighting factor $W_{pi}$ is given by Equation (1):

$$W_{pi} = \frac{ASC_{pi}}{AWS_{pi}} \quad \forall \ p, i$$

$ASC_{pi}$ and $AWS_{pi}$ are the number of users entering the network at station $i$ in period $p$ according to SC data and to WS data, respectively. Any device starting a trip at station $i$ during time interval $p$ is multiplied by $W_{pi}$ when computing the OD matrix and arcs load.

3.3 Other measures

The at-station waiting time is derived by first calculating the time difference between the first and last detections of each device at its origin station and then averaging the results for each station-period. Excluded are outliers and devices whose first and last detections occur at the same moment. Assuming that the first detection occurs when the device is at the station’s waiting area, the results will be an underestimate given that by construction it will always represent a quantity contained entirely within the real waiting time.

The number of trips made by each device on a daily or weekly basis is readily calculated using the MAC code. The resulting frequencies can then be employed to classify users into those regularly making the same trips and those who travel more sporadically. The same analysis can also be conducted with SC data. The
disadvantage with WS data is that there may be users travelling with more than one 
device or no device at all. On the other hand, SC data for this indicator are biased in 
that some users have more than one card while some cards are used by more than 
one person.

3 Experiments and results

3.1 The data

The PTS tested considers a 3.7 km long segment of the metro system of Santiago, 
Chile that includes stations Los Dominicos (E1), Hernando de Magallanes (E2), 
Manquehue (E3) and Escuela Militar (E4). The four stations are underground and 
ext to one of Santiago’s main streets. E1 is the terminal station and E4 connects 
this subnetwork to the other 104 stations of the metro network.

Since only four (out of 108) stations are considered, numerous users travelling to the 
stations in our subnetwork must obviously have entered the network at one of the 
many other stations in the PTS that were not in our experiment and so did not have a 
WS installed. Since these users would have had to pass through E4, our 
methodology assumes that they entered the system there. This implies that for E4 
there would be little point in using the weighting factors or calculating any results. 
Had we installed scanners in every station, of course, this limitation would not have 
arisen.

SC user arrivals data were obtained for the four stations in 15-minute periods over a 
six-week stretch in 2016. Each observation included the day, the time period, the 
name of the station and the number of users, separating out senior citizens and 
students (school and post-secondary).

A single WS was installed in each of four consecutive stations. The device detection 
data collected by the WS at each station covered the same six weeks as the SC data, 
and to ensure compatibility the detections were also grouped into 15-minute periods. 
Parameter $L$ is set to 10 minutes.

3.2 Results

Since ground truth OD matrix is unknown, we compare the structure of the OD 
matrix obtained by using the proposed methodology ($ODM_{WS}$) and the one 
estimated from the SC ($ODM_{SC}$). The $ODM_{WS}$ was obtained by summing the 
expanded WS detections over the 15-minute periods between 5 pm and 8 pm on the 
Tuesday in our experiment. The $ODM_{SC}$ was estimated based on Munizaga and 
Palma (2012) and for the same day and time of day, but 17 months earlier. The PTS 
network did not undergo any structural changes during that period that would 
materially affect the compatibility of the two versions.
Figure 1 shows both the ODM\textsubscript{SC} and ODM\textsubscript{WS} in terms of the number of trips. The structure of both matrices is similar, as confirmed by the MSSIM\textsuperscript{1}. As expected, E4 generates and attracts most of the trips in both cases.

Two versions of arc load were calculated for comparison purposes as well: LOAD\textsubscript{WS} and LOAD\textsubscript{SC}. The results for both directions of travel are shown in Figure 2.

The arc-load estimates are more stable than those for the OD matrices given that they can be based on detections received from any stage in a trip and thus do not depend strictly on signals picked up at the origin or destination stations. LOAD\textsubscript{SC} and LOAD\textsubscript{WS} for direction 1 (from E1 to E4) display greater similarity than those for direction 2. This may be due to the way in which the weighting factors were applied, since for users travelling in direction 1 the weights were calculated in terms of their respective origin stations. This is not the case for direction 2 users.

\textsuperscript{1}Mean Structural Similarity Index. This index has been proposed to compare the structure of OD matrices (see Pollard et al., 2013).
The individual at-station wait times for stations E1, E2 and E3 on the Monday in our experiment are graphed in Figure 3. The wait times grow during periods of peak arrivals, although the values are highly variable, especially during the afternoon peak.

![Fig. 3 Wait times for stations E1, E2 and E3 estimated using WS data (Monday).](image)

The increase during busy times is likely due to two factors observed in practice: (i) an arriving train may be full or almost full, forcing users to wait for a later one because they cannot board or they decide not to board crowded vehicles and prefer to wait for one less crowded; and (ii) there may be a queue just to get onto the station platforms.

Since the encryption of the MAC code changed every day, we were only able to calculate the frequency of trips by each device on a daily basis. For the week in our experiment, we found that the proportions taking different numbers of trips per day was very stable: 78% travelled once a day while 21% travelled twice a day. Regardless of the reasons behind these figures, our results reveal that one-way trips are the great majority, suggesting there is much variability in the daily activities and commuting habits of those who use the PTS in our experiment.

We finally analyze the behavior of the weighting factors. In summary, weighting factors varies with time and space. Higher penetrations are observed during the morning peak than during the rest of the days. Station E2 also exhibits a higher penetration than the other stations. The full paper further discusses on these issues, mentioning potential reasons to explain this behavior and their impact in the methodology.

Acknowledgements: This research was supported by the Centro de Desarrollo Urbano Sustentable (CEDEUS) (FONDAP 15110020 from CONICYT). We would also like to thank Metro de Santiago for providing the smartcard data used in this study.

References (of the full version)


