# Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys

Noelia Caceres[a]*, L.M. Romero[b] and Francisco G. Benitez[b]

[a] *Transportation Engineering Unit, AICIA, Seville, Spain;* [b] *Department of Transportation Engineering, School of Engineering, University of Seville, Seville, Spain*

Camino de los Descubrimientos s/n, 41092 Seville, Spain. Tel.: +34 95 448 8135. E-mail address: ncaceres@us.es *corresponding author

# Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys

Origin–destination (OD) matrices serve as a basis for travel demand modelling. Traditionally, they are derived from travel surveys that collect detailed trip information but with several shortcomings. Mobile phones are regarded as a useful source of information on people's daily mobility. This work explores the use of mobile data in the context of mobility studies by comparing matrices derived from both types of sources over the same region. The results reveal many common features in the trip information. Moreover, although the use of mobile technology may raise questions for short trips, the huge representativeness of this technology captures the mobility in OD connections extensively regardless the area. This is crucial for non-populated areas (e.g. industrial parks or educational campuses), which constitute important mobility hotspots. Based on these findings, an applicable data fusion approach to obtain the optimum accuracy from these heterogeneous sources is presented and applied.

Keywords: mobile phone data; household travel surveys; OD matrices

## 1. Introduction

Origin–destination (OD) or trip matrices are fundamental inputs for most problems regarding the planning and management of transportation systems. They reflect the mobility in an area of study during a particular period of time. Each cell in the matrix indicates the number of trips that depart from each origin zone to each destination zone. Hence travel demand is usually represented using an OD matrix. In the past few decades, OD matrices have been derived from travel surveys in which people are asked to describe their travel behaviour on an average day or to reconstruct their travel pattern on one or more previous days. Many types of surveys, such as home interviews, roadside interviews, or even a combination of them, are used to obtain OD matrices in practice; such surveys vary in complexity regarding the information which can feasibly be collected and in the level of interaction between the survey designer and the respondents (Richardson et al. 1995). Home or household interview surveys are the most widely used and are essentially intended to yield data on the travel behaviour of the residents of the household (e.g. number of trips made, their origin and destination, purpose of trip, mode(s) of travel, departure and arrival times, etc.) and the general characteristics of the household or respondent (e.g. family size, age, sex, income, vehicle ownership, etc.). Survey-based approaches involve costly and laborious processes for collecting, coding, and processing data (Stopher and Greaves 2007; Santos et al. 2011; Ortuzar and Willumsen 2011). As a result, surveys are not conducted frequently (about every five or even ten years) and various methods are used for estimating an OD matrix in practice (as reviewed in Abrahamsson 1998). Among them, the use of traffic counts as measurements of link flows in a network model in order to update an existing (probably outdated) matrix has been widely considered by many researchers (Van Zuylen and Willumsen 1980; Doblas and Benitez 2005; Cascetta et al. 2013). However inference approaches based on just link count data are known to be a challenging problem; normally there are a large number of matrices which reproduce

the observed traffic counts (Abrahamsson 1998). Advances in data collection and computational techniques have incorporated other observed data sources in the estimation procedure, such as automated vehicle detection (Zhou and Mahmassani 2006; Castillo et al. 2013) or routing data (Herrera et al. 2010; Parry and Hazelton 2012). Obviously the accuracy of OD matrices obtained by these methods will be highly dependent on the quality of the information used. Road planners and practitioners demand more accurate and updated information for transport planning and decision making. Survey-based approaches strive to collect very detailed trip information, but they present several deficiencies (Brög and Erl 1999; Bonnel 2003; Ampt and Ortuzar 2004; Cools et al. 2010; Ortuzar et al. 2011), mainly related to sampling biases (e.g. falling response rates, vacant dwellings) and reporting errors (e.g. unreported trips, rounding of arrival and departure times). They no longer seem adequate to provide reliable OD matrices. In recent decades, mobile phone data have also become a promising data source. Mobile phones are pervasively embedded in people's lives – more than one-third of consumers worldwide said they check their phone within five minutes of waking up in the morning, and 20 percent of them check their phone more than 50 times a day (Global Mobile Consumer Trends 2017). The deployment of mobile technology is producing a massive increase in the volume of data regarding where people have been and when they were there, which can not only lead to accurate and updated estimation of mobility matrices, but also to infer other valuable features for planning transport, exhaustive reviews can be found in Milne and Watling (2018) and Gadzinski (2018).

The idea of using mobile phones to acquire transport information is becoming more and more widespread; reviews of current practices can be found in Steenbruggen et al. (2015); Chen et al. (2016); Diao et al., (2016); Rojas et al. (2016); Lee et al. (2016); Lu et al. (2017) Malleson et al. (2018); Wang et al. (2018) among others. In the field of OD matrices and travel behaviour, the use of mobile phone data has been explored by researchers working on

simulated frameworks (Caceres et al. 2007; Sohn and Kim 2008; Zhang et al. 2010; Hofer et al. 2018) as well as in field tests (Mellegard et al. 2011; Calabrese et al. 2013; Widhalm et al, 2015, Demissie et al. 2015; Horn et al. 2017; Wang and Chen 2018; Ni et al. 2018). Their results reveal that mobile data can overcome typical limitations of traditional surveys, with higher sample size, wider coverage, and reduction of the time and cost of data collection and processing, while also providing valuable information on temporary mobility patterns. By contrast, these works also mention the problems associated with using mobile technology as a mobility data source, especially in terms of low spatial resolution and accuracy. To obtain more realistic matrices, recent approaches have combined mobile phone data with other types of data, such as traffic counts (Caceres et al. 2013; Iqbal et al. 2014; Wu et al. 2015; Meng et al. 2017), GPS traces (Gong et al. 2014; Ge and Fukuda 2016; Seo et al. 2017; Nigro et al. 2018), and even a fusion of crowdsourced geospatial information, smartcard transactions, census records, and/or surveys (Toole et al. 2015; Anda et al. 2017; de Regt et al. 2017; Bonnel et al., 2018). Despite the important efforts made in exploring the use of mobile phone data to characterise people's mobility, there are still pending issues to be reviewed in the context of matrix estimation. This study is an attempt to extend the effort in the area to gain a clearer understanding of the potentialities and challenges of this technology compared to traditional survey sources, based on the outcomes derived from a real case study.

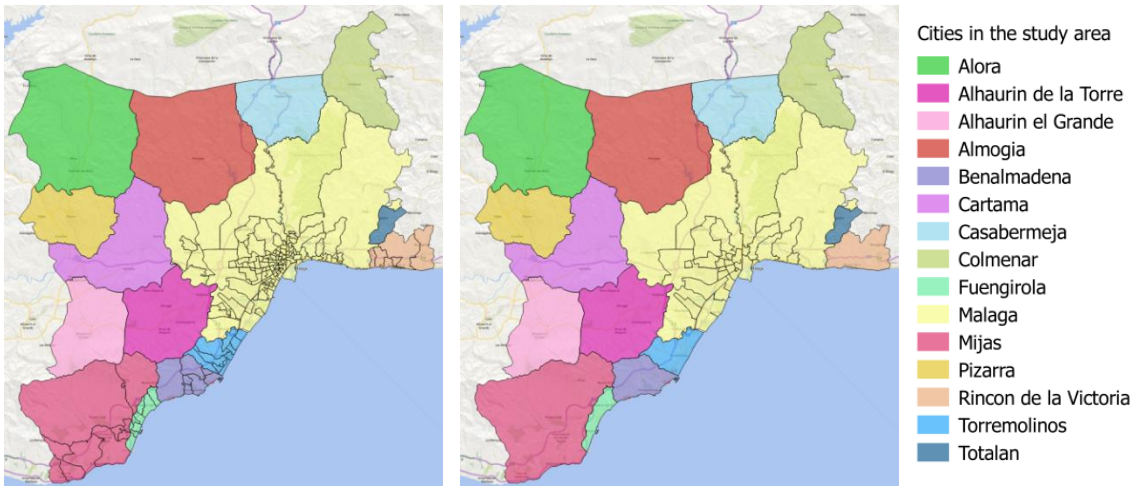
This paper is organised as follows: Section 2 introduces an overall view of the case study, presenting the study area and the data sources. Section 3 exploits the transport data of the pilot case, derived from a household travel survey and a mobile phone telecommunication provider; a comparative analysis is carried out, highlighting the differences and similarities from the qualitative and quantitative points of view. Section 4 provides a discussion on the main advantages and drawbacks revealed by the analysis. Finally, the potentialities and

challenges of mobile phone data in complementing traditional survey methods are illustrated in Section 5 along with some suggested directions for further practical applications.

## 2. Data and methods

### *2.1. Study area*

The study area corresponds to the urban agglomeration of Malaga, located in the south of Spain on the Mediterranean coast. It consists of the city of Malaga, which is the sixth most populated city in Spain, and 14 surrounding municipalities in both the coastal zone and the interior. The study area has a population of around one million inhabitants and covers approximately 1400 km$^2$, divided into 178 transport zones (TZ) defined from census data (Fig. 1a). The zoning is also further aggregated into 46 macro-zones (MZ), more general areas based on adjacent (similar socioeconomic) transport zones (Fig. 1b).




(a) (b)

**Fig. 1** Zoning of study area: (a) transport zones and (b) macro-zones. Cities are represented by different colours.

### *2.2. Trip matrices*

This study is based on available OD matrices derived from two sources: household travel surveys (HTS) and mobile phone data (MPD). The last household travel survey, led by the regional transport system planning administration, was conducted in October 2014 by the

Malaga Area Metropolitan Transport Consortium (a Spanish public transport body) with the aim of characterising a detailed picture of workday mobility and travel choices made by residents of the study area. The survey reported all trips made by each resident of the sampled households. In particular, approximately 30,000 persons were interviewed (around 3% of the population in the region). The results were then statistically expanded and validated based on other socioeconomic datasets obtained by governmental agencies, deriving the corresponding OD matrices.

Mobile phone data come from network operators which collect, store, and process massive datasets on subscribers so they can route calls and offer services. The data used for this study are based on aggregated and anonymised phone events collected and processed by a telecommunication operator with a large market share (around 40%) in the studied area. These events consist of active interactions related to phone calls and text messages, as well as passive interactions, which occur in the background (or idle status) without the user's active participation. These passive interactions are associated with signalling, such as losing/regaining mobile signal or periodic records created when phones are on but have not created any other events for a sustained period of time (typically of the order of a few hours). Besides, there is also a relevant passive event associated with movement from one specific group of cells (called a location area, LA) to another. In mobile systems, the service coverage area is classified into cells and these are grouped into LAs. For mobile operating purposes, whenever a phone enters a new LA, it notifies the mobile network of its new position. This type of event substantially increases the number of the MPD events. Each one is characterised by an encrypted user ID (following strict anonymised protocols to guarantee privacy), a timestamp when the event occurs, and an event location estimated using algorithms based on triangulation of mobile phone mast signals. These events provide 'footprints' regarding where people have been and when they were there. In this study, a trip is regarded as a one-way

movement from a zone of origin to a zone of destination at a particular starting time. Users are more likely to engage in an activity (by means of a trip) after a 'stay' at a particular location (that represents the origin or destination of a trip). The remaining footprints provided by the events occur during the displacement that takes place. Therefore, the first step is to identify which footprints are 'stays'. For this purpose, several authors have developed different algorithms (Widhalm et al. 2015; Alexander et al. 2015; Toole et al. 2015; Jiang et al. 2017). In this work, the identification is based on a time threshold in the subsequence of events. This threshold for the time between consecutive events ($t_{bce}$) has been taken as a simply rule-of-thumb for identifying if an event belongs to a possible new trip ($t_{bce} \geq 30$ min) or to the same trip after a brief stop ($t_{bce} < 30$ min). An event defines the end of a trip when the time difference to the next event is more than 30 minutes; this end defines the beginning of the 'stay' but also the origin of the next subsequence of events. By processing all existing events created by the sample of users, trips are inferred. These trips are derived by considering the proximity of events not only in terms of time but also regarding the space. For instance, it is necessary to analyse whether events reveal an actual movement or they are generated by static users due to the ping-pong records in the neighbouring towers. The characteristics of the transport network topology are also analysed. For instance, the associated distance between two consecutive events as well as the difference between their timestamps has to be checked to ensure that they are compatible with the travel distance and travel time, respectively, for the possible routes. Based on this, all events generated over the study area during two consecutive weeks in February 2015 (ten working days and four weekend days) were processed to infer trips made by the considered sample (approximately 200,000 persons, aged 18 years and above) based on subscribers of the particular operator. The results were expanded to represent the full population of the studied area based on census data taking into account the area where user's home is located, similarly to expansion made in

travel surveys, besides additional mobile phone data features and privacy-assured specificities have been considered which affected the expansion: i) only data corresponding to users over/equal 18-years old are processed (it is forbidden by national regulations to provide mobile data concerning children or teenagers), ii) OD pairs data with less than 5 detected trips are eliminated (national data protection laws do not allow to exploit mobile records with ~~that lead to infer~~ 1-4 trips because it may be possible to identify the person(s) making the trip(s), which is absolutely forbidden to be inferred from mobile technology), iii) data expansion follow the market share of the phone operators of the region of study, because the mobile sample data used in this case only represent 40%. These steps were supplementarily applied to the trip inference taking into account the implicit (and controllable) restrictions related to privacy. For the expansion procedure, ~~T~~the home was determined from events generated late at night on weekdays (when people usually stay at home). In this sense, it is necessary to highlight that, although there are several approaches for expanding data, the approach applied to mobile data was based on census data in order to be similar to the one used in the surveys provided for this study (this is a task customarily conducted by the Regional transport system planning administration and in many cases executed by specialised~~st~~ survey companies, following standard procedures). So that both data sources (HTS and MPD) were expanded in ~~a~~ similar manner~~way~~ with the final purpose of not introducing additional bias in comparison associated to the expansion stage. As it is previously commented, p~~P~~rivacy is a serious concern in using data from this kind of technology; hence, to increase privacy, the data are translated to be referred at transport zones shown in Fig. 1. For this purpose, the estimated location is directly linked to the corresponding traffic zone in the zoning system. Trips are also hourly aggregate; that is, a trip is assigned to each hour period based on its starting time. Therefore, data are anonymised, aggregated, and expanded, thus it is not possible to associate the data with individual users.

Then, from both kinds of sources, the number of trips departing from one transport zone to another in a particular one-hour period (hourly matrices) and in the 24-hour period (daily matrices) were derived and compared. The two data sources were not extracted in the same time period, there was an interval of several months between HTS and MPD. However, they are sufficiently close in time to assume that the mobility behaviour did not suffer significant changes; the metropolitan area of Malaga presents a very homogenous mobility features during these months of the year.

## 3. Comparative analysis

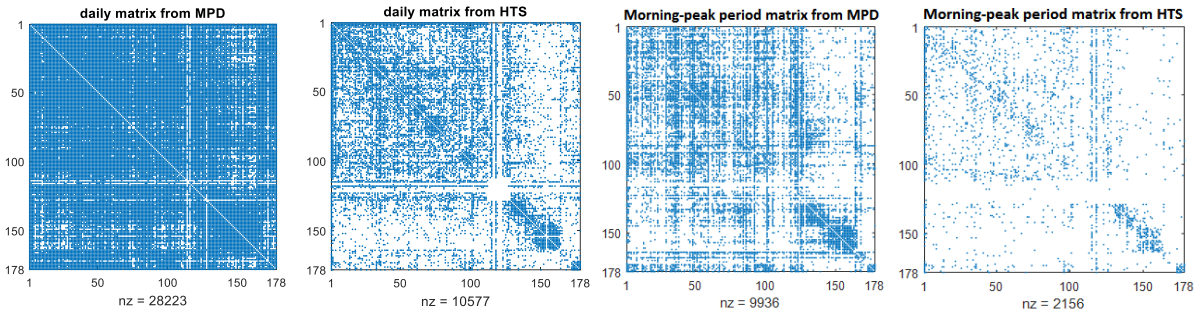
This section compares and analyses the trip data contained in the mobility matrices derived from the two types of sources (mobile data and surveys), providing the main qualitative and quantitative findings. For comparative purposes, data trips are pre-processed in order to conduct the comparison under the same context. Due to legal issues, only data from persons aged 18 years and above are processed., W~~w~~hile survey data collects information from all residents of the sampled households, including infants and children, MPD source does not provide such information affected by privacy regulations. To avoid this age effect ~~and~~regarding ~~-~~to compar~~inge~~ the same mobility phenomenon (trips made by people aged 18 years and above), the matrices derived from survey data were processed to contain only trips made by adults.

### *3.1. Quantitative findings*

#### *3.1.1. Sparsity of matrices*

Before making any comparison, the first step to be taken is to analyse the information contained in the matrices. A lack of trips between a pair of regions indicates that these regions are not generating or attracting trips between each other for some reason. A straightforward pairwise cross-checking between HTS and MPD matrices identifies cases where single pairs

of regions present, with regard to the number of trips: i) similar values (either null or non-null ones), ii) very different non-null values, or iii) a null value in one of the matrices versus a non-null value in the other. In order to explore this issue, the sparsity of the matrices is analysed and compared. The sparsity of the matrix is defined as the number of zero pairs (i.e. matrix cell or element) divided by the total number of pairs. Fig. 2 plots this concept for the matrices derived from both sources by marking the locations of the nonzero pairs with blue circles; the number of nonzero pairs (nz) is also reported.




(a) (b) (c) (d)

**Fig. 2** Sparsity of matrices based on the 178 transport zones. At daily level: (a) derived from MPD, (b) derived from HTS. For the morning-peak period: (c) derived from MPD, (d) derived from HTS. ('morning-peak' refers to the time interval 08:00–08:59).

The visual analysis of the plots in Fig. 2 reveals that the sparsity in matrices from the two sources is somehow similar for a particular subset of OD-pairs. There are pairs in the two sources that do not have trips; this is more acute during late night hours (as expected) or between certain regions. However, the comparison also shows that there are numerous cells with zeros in HTS-based matrices that do not correspond to zeros in the MPD-based ones, giving the impression that mobile data capture mobility in a higher percentage of all possible OD connections. This can be better appreciated by looking at the number of nonzero pairs in the two sources. With a zoning-system granularity of 178 transport zones, the matrix consists of 31,506 pairs of inter-zonal trips. Then, at daily level, the number of nonzero pairs in the MPD-based matrix is around 89% of the total, while in the HTS-based matrix it is around

34% (Fig. 2 a and b, respectively). In Fig. 2 c and d, for the morning-peak period (trips starting between 8:00 and 8:59), the percentage of nonzero pairs in the MPD-based matrix is 32% versus less than 10% in the HTS-based matrix. This huge difference in the number of nonzero pairs between sources also appears at other hour-periods (Fig. 3); for instance, in periods in which workday traffic is usually concentrated (between 7:00 and 20:00 hours) the percentage of nonzero pairs remains around 30% in the MPD-based matrix while it is around 5% in the HTS-based one. This can be explained by the presumed wider representativeness of the mobile phone sample, based on the fact that surveys cannot observe all possible OD pairs. In general, observed trip matrices (e.g. derived from surveys) have a large number of empty cells and probably a set of cells with large values. The reason for this is that in a particular time interval some OD pairs are more likely to contain trips than others, thus leaving numerous cells with a very low number of expected trips. Thus the probability of making no observations of a particular OD pair is large. This sparsity property of the HTS-based matrices can be considered as a weakness of this methodology, since this kind of matrices have to distribute the total number of trips T among a substantially fewer number of elements, the nonzero OD pairs. Therefore the expansion process assigns more trips to the OD pairs collected in the survey than the real matrix, with trips between all pairs. Consequently, the number of trips in these cells is overestimated; meanwhile trips from non-collected OD pairs are neglected. By contrast, the wide representativeness of the mobile sample produces dense matrices in which most of the cells are filled, although a share of them contains small values.

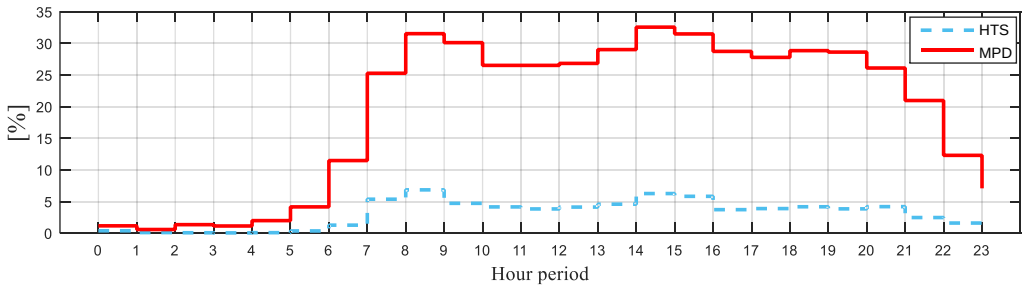



**Fig. 3** Number of nonzero elements in the hourly matrix derived from both HTS and MPD.

### *3.1.2. Similarities in trip information: OD pairs*

Once the differences in sparseness have been assessed, the level of similarity between matrices from the two sources (HTS and MPD) is reviewed and compared. First of all, the analysis focuses on trip flow between transport zones (each cell in the matrix) in the two sources. Taking into account that the study area is further divided into macro-zones, the trip information at this zoning level has also been considered. One of the measurements evaluated is the linear correlation coefficient, Pearson's coefficient ($R_p$), to determine the level of similarity from the quantitative point of view. This coefficient is the most widely used measure of the relationship between two variables (i.e. when a change in one variable is associated with a proportional change in the other variable). At transport zone level and on a daily basis, the Pearson's coefficient reveals a weak positive linear relationship between mobile-based and survey-based trips, with $R_p$ = 0.44; by contrast, the coefficient at macro-zone level yields a strong linear relationship between the numerical values (trips) in OD pairs ($R_p$ = 0.81). On an hourly basis, similar tendencies are obtained (Fig. 4), and the two sources get stronger correlation when the zoning system is rolled up to a lower level of granularity (i.e. broader zones, less refined discretisation). This remains in time periods in which workday traffic is usually concentrated (between 7:00 and 20:00 hours). They coincide with hours during which mobile events are less dispersed according to the coefficient of variation, meaning there is a higher reliability for analytical purposes.

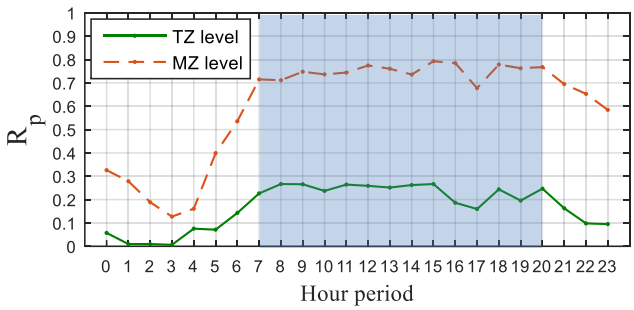
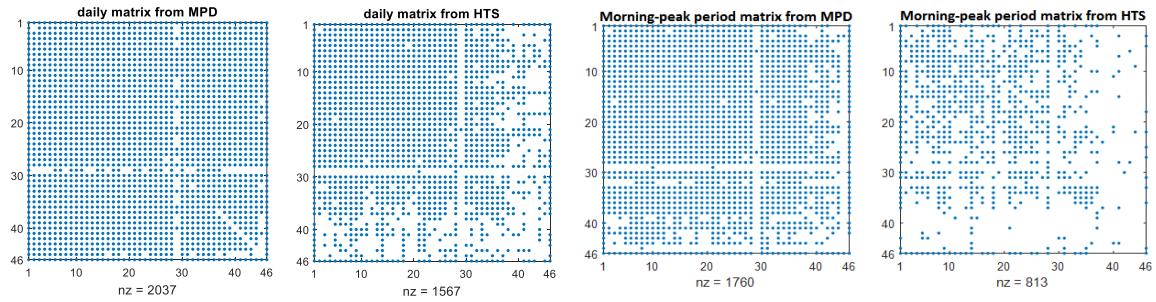



**Fig. 4** Pearson's coefficient ($R_p$) by one-hour periods of the day, correlating trips in OD pairs derived from the MPD and HTS sources, at both transport zone and macro-zone level.

However, it is necessary to remark that this kind of traditional correlation metrics may not properly deal in terms of comparing matrices with differences in sparsity. As previously discussed, matrices derived from HTS have a larger number of empty cells that those derived from MPD. This difference in the number of nonzero pairs (compared in Fig. 2) reveals that the comparison may lose reliability at transport zone level. Of course, the empty cells in the HTS-based matrices can be attributed to sample size limitations, as discussed above, but also to the geographical specification of the traffic zones. Then, the results can be more easily understood when the zoning system is rolled up to a lower granularity. In such a case, although the difference persists, the sparsity of the two matrices based on the 46 macro-zones becomes somewhat more comparable at both daily level (Fig. 5 a and b) and hourly level (Fig. 5 c and d). Then, the comparative analysis of sources is probably more consistent at macro-zone level, for which the Pearson coefficient gives a reasonably high relationship between the trip information derived from the two sources.




(a) (b) (c) (d)

**Fig. 5** Sparsity of matrices based on the 46 macro-zones. At daily level: (a) derived from MPD, (b) derived from HTS. At morning-peak period level: (c) derived from MPD; (d) derived from HTS.

In the previous comparisons, OD matrices are treated as a set of unconnected numbers, ignoring spatial and temporal facets of the mobility that are also synthesised in a matrix-based form. Therefore, it is advisable to also evaluate the similarity between matrices in this context, but traditional statistical measures (e.g., r-squared, mean square error, etc.) are not able to find this kind of structural correlation in data. Recent works in the literature

propose the use of the Mean Structural SIMilarity (*MSSIM*) index as an OD matrix comparator (Djukic et al. 2013; Pollard et al. 2013; Day-Pollard and Van Vuren 2015). The most important feature of this metric, developed for comparing images at pixel level (Wang et al., 2004), is the use of additional information in the evaluation process on the basis of the structural patterns in data. In general, images are highly structured, with pixels that are close to each other appearing to have strong dependencies. Based on this, the *MSSIM* computes statistics on groups of pixels (those within a window) and then takes the average (mean), rather than computing statistics based on all the pixels in the image together. If an image is equated to an OD matrix, the cells in the matrix (or OD pairs) can be seen as pixels that exhibit strong dependencies as well. If the two matrices describe similar transport patterns then, as with pixels, it is reasonable to expect similarities on windows of OD patterns (Pollard et al. 2013). The *MSSIM* index is designed to capture this property in a [–1, 1] scaled range. The largest positive value means the highest similarity (full match), while the smallest negative value entails the highest dissimilarity (significant differences). Although the *MSSIM* approach requires further refinement for use with OD matrices, the *MSSIM* index can identify structural differences better than traditional measures such as r-squared; conversely, *MSSIM* is successfully used as a measure of matrix similarity in the works referred to above. The *MSSIM* is calculated by summing and averaging Structural SIMilarity (*SSIM*) values across a whole matrix using an iterative procedure in which the *SSIM* is calculated over a part of the matrix (local window), generally a few cells wide by a few cells high. The expression used for calculating the *MSSIM* between two matrices, *A* and *B*, is:

$$MSSIM(A,B) = \frac{1}{M} \cdot \sum_{j=1}^{M} SSIM(a_j, b_j) \tag{1}$$

where $a_j$ and $b_j$ are the OD matrix contents at the $j$-th local window and $M$ is the number of local windows in the matrix. The local window, which may be a square box of $N \times N$ elements, moves cell-by-cell over the entire matrix. At each step, the local statistics and the

*SSIM* index are calculated within the local window according to the following expression:

$$SSIM(a,b) = \frac{2 \cdot \mu_a \cdot \mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \cdot \frac{2 \cdot \sigma_{a\,b} + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \quad (2)$$

where $\mu_a$ and $\mu_b$ are the mean values within this part of the matrices *a* and *b*, respectively; $\sigma_a$ and $\sigma_b$ are the variance of each dataset; $\sigma_{ab}$ is the covariance of the two matrices; and $C_1$ and $C_2$ are stabilisation coefficients. The *SSIM* is calculated according to the methodology described by Wang et al. (2004) and Pollard et al. (2013), with their values for $C_1$ and $C_2$. To define the window, it is assumed that the pixels of images that are closer to each other usually have a stronger relationship than the more distant ones. The contribution of the proximity in the *SSIM* calculation is then modulated by a Gaussian weighting function, $\mathbf{w} = \{w_i \mid i = 1, 2, \ldots, N\}$ for the local window of size *N*. The estimates of local statistics for the mean, variance, and covariance in (2) are then modified to incorporate such weights in the *SSIM* calculations. But in this sense, it is necessary to remark that, unlike in images, the proximity of cells in OD matrices does not guarantee that they are also close in space (e.g., zones 1 and 2 may be rather distant) and the window has to be established to consider this proximity. This issue was already considered in other works using this index for comparing OD matrices; Day-Pollard and Van Vuren (2015) proposed the Euclidean distance between two cells *i* and *j* defined in (3), using the xy-coordinates of the origin and destination centroid for OD pair *i* ($x_{Oi}$, $y_{Oi}$, $x_{Di}$, $y_{Di}$) and OD pair *k* ($x_{Ok}$, $y_{Ok}$, $x_{Dk}$, $y_{Dk}$), as a way of considering this fact in the *SSIM* calculations:

$$d(ODpair_i, ODpair_k) = \sqrt{(x_{Oi} - x_{Ok})^2 + (y_{Oi} - y_{Ok})^2 + (x_{Di} - x_{Dk})^2 + (y_{Di} - y_{Dk})^2} \quad (3)$$

This Euclidean distance *d* is then used in the Gaussian function $w = exp\left(d^2 / \sigma\right)$ to determine the contribution of OD pairs, with cells that are further apart contributing less. According to this, the *MSSIM* values between daily and hourly matrices from the two sources

are calculated. At daily level, although slightly better results are obtained at macro-zone level than at transport zone level, the structural correlation reaches high values in both cases ($MSSIM_{TZ}$ = 0.78; $MSSIM_{M\,Z}$ = 0.83). Fig. 6 shows the structural correlation between matrices on an hourly basis, with similar values to the daily case in time periods in which workday traffic is usually concentrated (between 7:00 and 20:00 hours). This means that the structural patterns in the daily matrix derived from MPD are very close to those captured in the matrix derived from HTS, for example in terms of the distribution of trips over destinations. Fig. 7 a and b illustrate the flow distribution at the macro-zone level (only pairs with more than 1000 daily trips) for matrices derived from the HTS and MPD sources, respectively. A visual analysis reveals that both sources capture similar patterns, with the majority of flows directed to and from home and work areas in Malaga as well as a few flows to and from the second most populous city of the study area (Alhaurin de la Torre). In both cases, the highest flow occurs between the two cities located in the southwestern part of the area of study (Mijas and Fuengirola). Another interesting strand revealed by the analysis, focusing on the time periods between 7:00 and 20:00 hours, is that the *MSSIM* index remains around the same value (approx. 0.77) at both macro-zone and transport zone level. Unlike the linear correlation coefficient, which is more influenced by the sparsity of matrices, the *MSSIM* index does not appear to reflect significant changes working at different levels of zoning-system granularity. This is explained by the fact that the *MSSIM* index measures the structural aspects of how cells (OD pairs) relate to one another and not the individual cell values themselves. Therefore the *MSSIM* index, in addition to identifying structural differences better than traditional similarity/correlation indexes, makes it possible to simultaneously compare sparse and dense matrices under similar quantitative criteria.

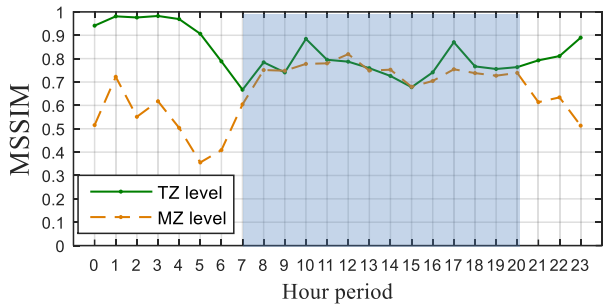
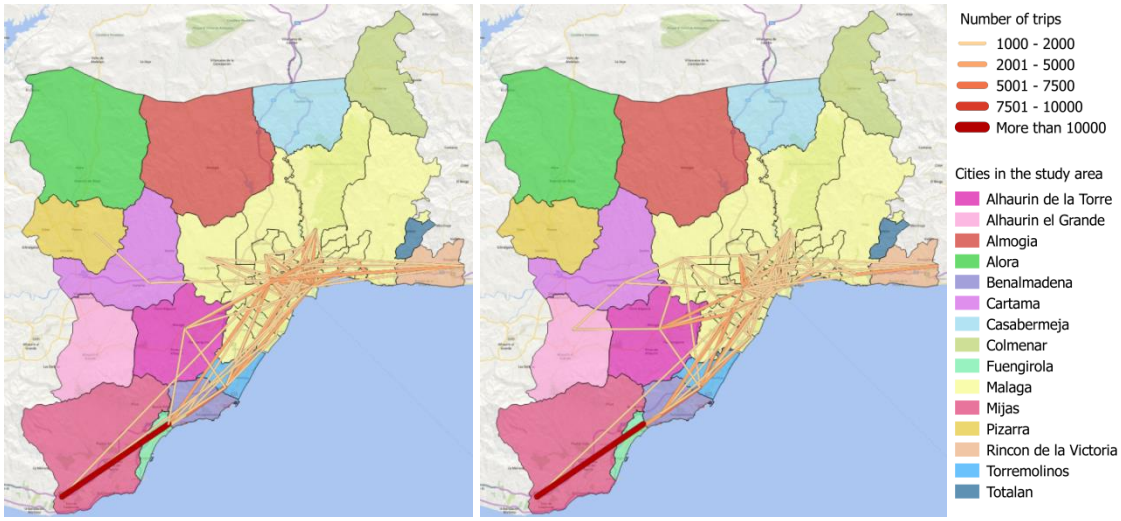



**Fig. 6** *MSSIM* index at both transport zone and macro-zone level for the relationship in the OD-pair data derived from the two sources (MPD and HTS) by one-hour period.




(a) (b)

**Fig. 7** Major OD flows between macro-zones: (a) daily matrix derived from the HTS and (b) daily matrix derived from the MPD.

### *3.1.3. Similarities in trip information: origins and destinations*

Additionally, the level of similarity by origin (aggregating columns in the matrix) and by destination (aggregating rows in the matrix) of trips has also been explored using the correlation coefficient. Table 1 presents the coefficients reached in these terms, with similar values to the values obtained at OD-pair level. The two sources are more correlated when the zoning system is rolled up to a lower granularity. Fig. 8 displays these coefficients for all one-hour periods of a day.

**Table 1.** Correlation between trip information from the two sources (MPD and HTS).

| Trip information | Zoning system | $R_p$ |
|---|---|---|
| *By origins* | Transport zone level | 0.38 |
| | Macro-zone level | 0.80 |
| *By destination* | Transport zone level | 0.38 |
| | Macro-zone level | 0.79 |

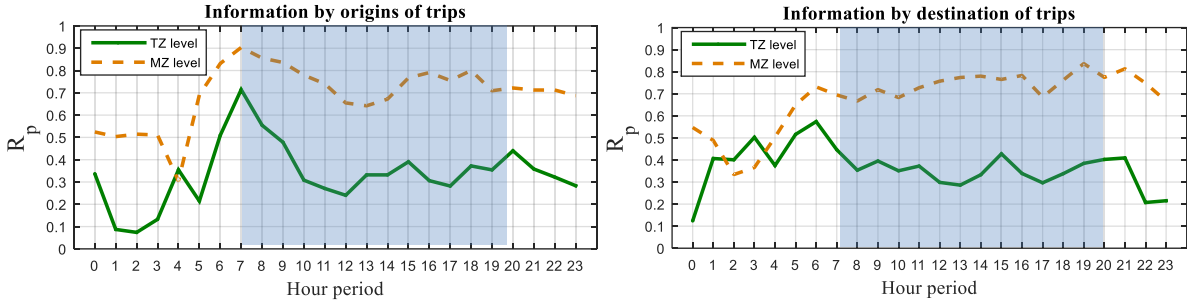



Fig. 8 Pearson's coefficient ($R_p$) by one-hour periods of the day, correlating MPD and HTS sources: (a) by origins and (b) by destinations of trips.

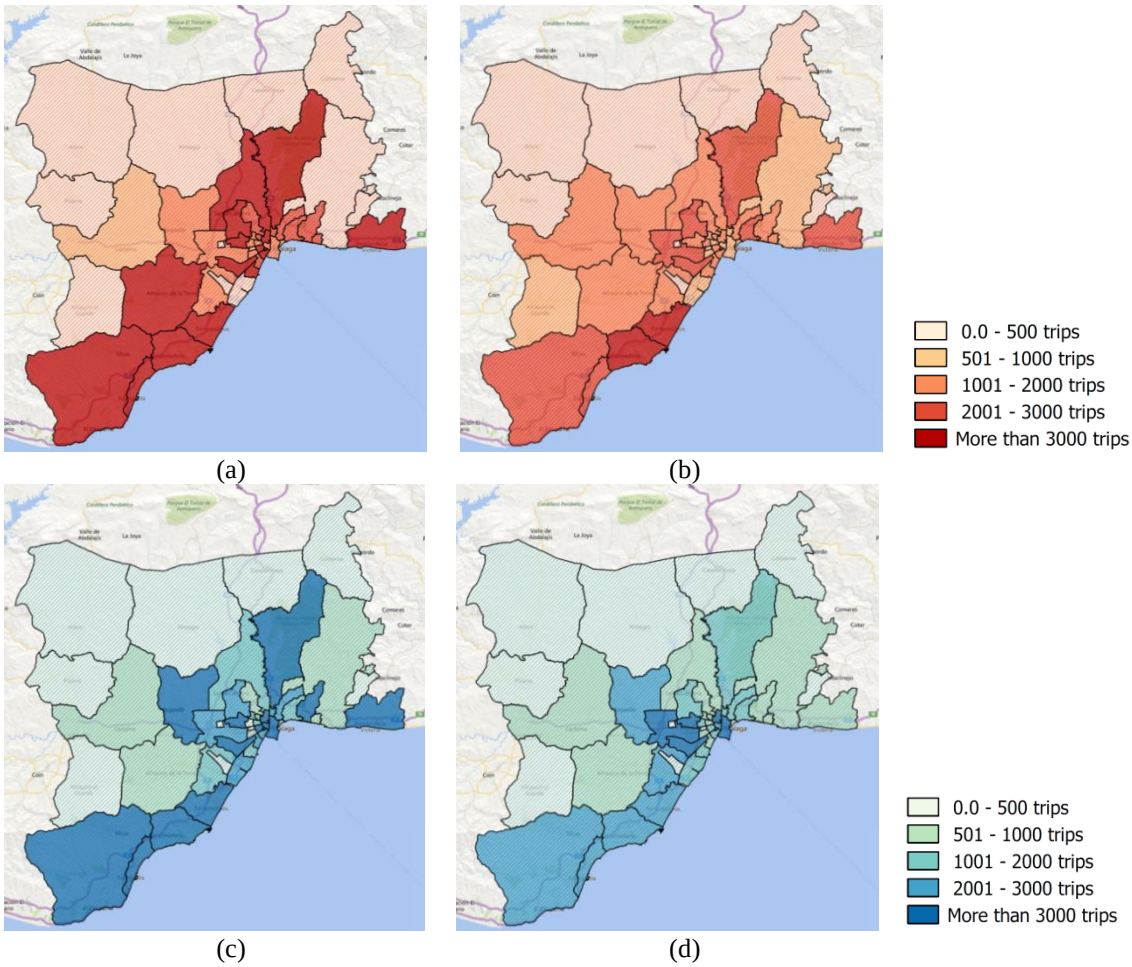



**Fig. 9** Number of trips during the morning-peak period (08:00–8:59): (a) trips originated in each macro-zone derived from HTS; (b) trips originated in each macro-zone derived from MPD; (c) trips terminated in each macro-zone derived from HTS; (d) trips terminated in each macro-zone derived from MPD.

Fig. 9 represents the study area coloured by the number of trips originated or terminated in each macro-zone from the HTS and MPD, all of them referred to the morning-peak period. In general, the two sources generate similar results, except for major cities (with an important component of short-distance trips) and macro-zones including non-populated areas. Regarding this last point, an interesting observation can be made by analysing the macro-zones in which the major divergences are concentrated. For this purpose, Fig. 10 displays the proportional distribution of the numerical differences (errors) between the two data sources. A positive sign means that the number of HTS-based trips is larger than the

number of MPD-based trips for each macro-zone, whereas a negative sign means the opposite. As can be seen, the major differences between sources occur in the macro-zones numbered 1, 14, 21, 26, and 27. In particular, for macro-zone number 1, containing the city centre, the analysis reveals that the number of trips originated using HTS is greater than the number from MPD. This macro-zone is characterised by an important component of shopping, services, and other leisure activities which imply mobility by walking and short trips, for which mobile technology may be less precise as a mobility probe (as discussed in the next sections). In contrast, the number of trips originated in macro-zones 14, 21, 26, and 27 derived from MPD is larger than the number derived from HTS. These macro-zones include university campuses and major industrial parks, indicating a significant divergence regarding the data-collection capability of HTS in these sectors.

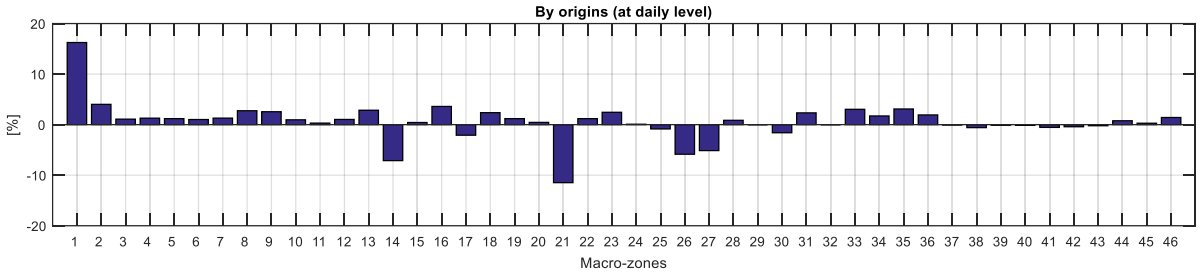



**Fig. 10** Total error distribution in the number of daily trips originated in each macro-zone between the two data sources (HTS and MPD-based matrices).

A closer look at the information by origins (MPD vs. HTS) now focusing on trips originated in each transport zone reveals an important finding when colour and size are added to include information on the population (Fig. 11). For trips generated during the whole day (Fig. 11a) and during the morning-peak period (Fig. 11b), the HTS source tends to underestimate those trips originated in non-populated zones (identified by big red dots). These transport zones are associated with mass transportation facilities (like the railway station, central bus terminal, and airport) as well as university campuses, industrial parks, and hospitals located in the study area. This suggests that these types of zones may not be

properly represented in survey campaigns, despite being important centres where trips originate and terminate. In most cases, this is the fault of designing samples in survey-based campaigns based on population size and in terms of the number of households in the study area. Non-populated zones are excluded from the sampling frame (there are no households to be interviewed); other cases, such as non-resident populated zones, are also not identified by surveys based on census data (e.g. university residences). Hence, the sample might not be representative of all people travelling from/to these zones. There are other types of surveys (e.g. intercept surveys) that take place at non-residential sites while the respondent is in the course of carrying out an activity of some type (e.g. shopping). Although they can better synthesise the mobility over these zones, these types of surveys using uncontrolled quota sampling have controversial drawbacks regarding reliability. It is necessary to use a combination of survey methods to take advantage of their pros and cons, but even a combination of intercept and home interview surveys may fail to produce matrices where all cells have been sampled (Ortuzar and Willumsen 2011). However, the representativeness of the MPD sample, which is homogeneously distributed across the territory, captures trips with independence of the socioeconomic characteristics of the zone (origin or destination) of trips. Other concerns regarding sample size are addressed in the next section.

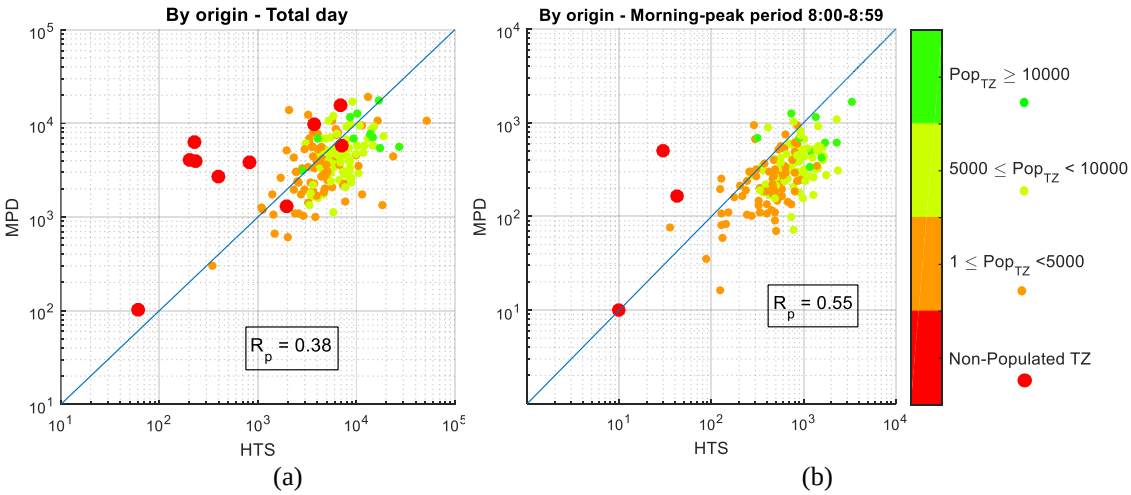
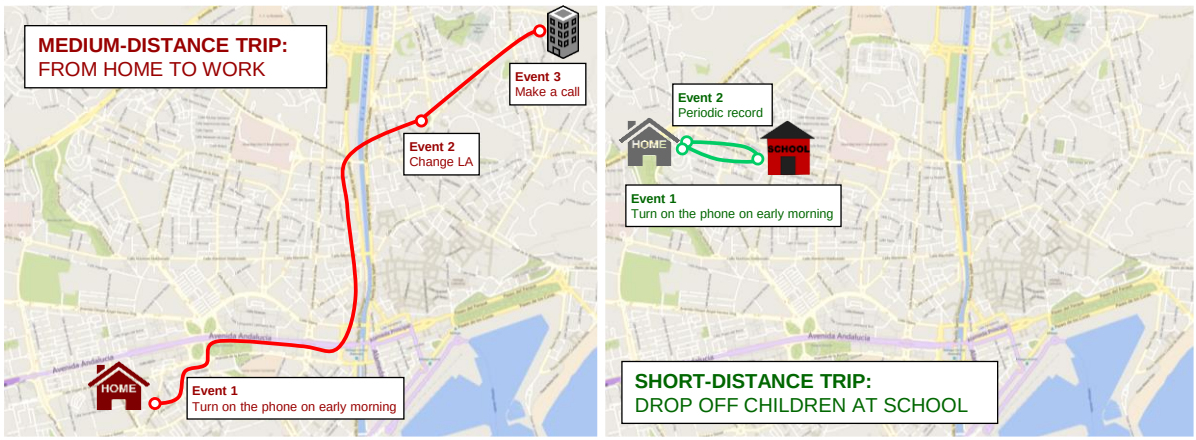



**Fig. 11** Number of trips originated in each transport zone (MPD vs. HTS) coloured according to the population: (a) during a day, and (b) during the morning peak period (08:00–8:59).

### *3.1.4. Trip distribution by distance*

From the mobile-data perspective, the detection of movements is strongly subject to the number of events generated by phones as they communicate with the network. The more events are generated, the more footprints are available from which to infer trips. In this regard, a longer trip duration increases the possibility of a call, message event, or even the abovementioned passive events. In a similar way, a longer trip distance also offers more opportunities to generate events (e.g. due to movement events created when a user changes from one group of cells to another). An example of this is depicted in Fig. 12a. In contrast, when trips are made in less time (because they imply shorter distances or are made at faster speeds), mobile phones leave fewer footprints of their 'approximate' locations during their movement (Fig. 12b). The consequences of these issues cannot be ignored when using MPD, since these types of short trips tend to be undercounted or completely overlooked when exploiting mobile data.

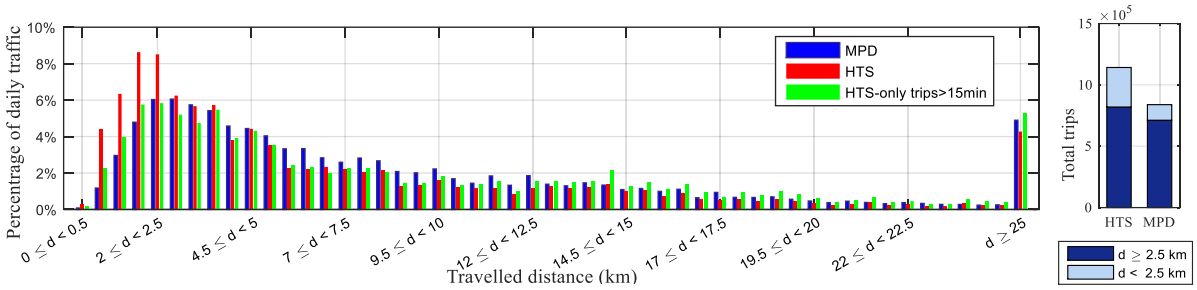



(a) (b)

**Fig. 12** Examples of mobile-event generation during (a) a non-short trip and (b) a short trip.

Therefore, focusing on distance travelled (based on the network shortest distance between the origin and destination centroids), Fig. 13a reveals that for medium and long distances, which are customarily made by motorised modes, travel rates are very similar for the two sources (MPD in blue and HTS in red). But for distances shorter than 2.5 km, a

significant reduction of the MPD rates is appreciated. This suggests that mobile data tend to under-report short-distance trips, as shown in Fig. 13b. In this respect, it is worth underlining that transport modes related to short-distance trips are usually non-motorised (Ryley 2008) and are primarily ‘walking’, although ‘cycling’ also occurs. Research in the literature shows that the average ‘walking’ speed is around 4 km per hour (km/h), while that of ‘cycling’ is around 10–12 km/h (Jensen et al. 2010), depending on factors such as the user's age, gender, or even surface condition. By cross-checking short-distance trips with low travel speeds, a dynamic mobility pattern on the scale of the neighbourhood in cities can be obtained; this pattern is difficult to detect with the spatial resolution offered by mobile technology (it is strongly dependent on the granularity of the mobile network), requiring additional methodologies like GPS (Ge and Fukuda 2016). Moreover, in terms of travel time, many of those short trips take less than 15–30 minutes, a quite reduced time window for generating mobile events. However, this comparison does not seem to be the most appropriate for contrast purposes, since survey-based approaches also present resolution problems in such time windows. For example, trips tend to be rounded to the nearest 10-minute or even 15-minute interval by survey respondents (Stopher and Greaves 2007) and daily travel times per person are considerably overestimated for the first 15-minute time interval (Gerike et al. 2015). Thus, to overcome these issues, Fig. 13a also displays travel rates derived from HTS considering only trips with durations (reported by respondents) greater than 15 minutes. These travel rates (bars in green) are certainly close to MPD rates (bars in blue), especially in the context of short distances. But the difference in magnitude of total trips between the two sources (Fig. 13b) reveals a concern that should be addressed when using MPD sources.




(a) (b)

**Fig. 13** Trip distribution from the two sources (MPD vs. HTS) by ranges of the distance travelled (in kilometres, km): (a) percentage of daily trips (b) total daily trips.

## *3.2. Qualitative findings*

### *3.2.1. Cost and time consumption*

The resources needed to conduct any travel survey can be defined in terms of budget and time. Besides the effort expended on sample design and data acquisition, a considerable amount of these resources are needed to code and process data. As a result, the time between surveys often ranges from 5 to 10 years. With this frequency, the data soon become outdated and, during intermediate periods, the transport demand must be updated using other available information (e.g. traffic counts) to derive estimates. Nowadays, administrators demand more updated information for analysing transportation policies and strategies. In this study, matrices derived from MPD were available in a few weeks, rather than the extremely long and costly traditional survey-based approaches. In this sense, mobile phone data offer substantial opportunities to improve the cost effectiveness and frequency of data collection and processing while also providing valuable information on temporary mobility patterns.

### *3.2.2. Sample design*

Representativeness is of paramount importance in describing mobility over a region. In this sense, the sample size plays an important role in determining the generalizability of findings. As mentioned above, the size of the sample in survey-based approaches is usually determined

on the basis of the residential population of the study area. Since survey budgets generally tend to be tight, the sample size forces a reasonable balance between quality and cost to generate good estimates in practice. In this study, approximately 30,000 persons were interviewed to construct the matrices derived from HTS (a population of around 3% in the region). A main strength of surveys is that they typically use statistical sampling in order to make inferences about a general population, within pre-specified margins of error; this allows findings to be generalised.

In the case of mobile data, the sample is customarily limited to subscribers of a given mobile phone operator; in this study the matrices derived from MPD were estimated on the basis of a sample of approximately 200,000 persons (aged 18 years and above). Therefore, although the selection of individuals is not subject to statistical sampling, the sample size is larger than in most household travel surveys. Certainly, MPD are provided by a single operator who has a finite market share, but one can consider that the sample follows a homogeneous distribution in time and space based on the pervasive use of mobile technology, with a mobile subscription penetration of 109% in Spain (CNMC 2015), assuming a random sample extracted from the whole population. This assumption may be not adequate when mobile operators have different penetration rates in the area of study, which may occur when broader areas are analysed (not in this case study). Due to the inherent dependency on the users' profile, several works have suggested sampling problems when using datasets from mobile phone users (Frias-Martinez and Virseda 2012; Ranjan et al. 2012). For instance, high-frequency callers may not always be representative of an entire population or certain population segments may not be properly represented in the sample (e.g., the usage of mobile phones seems somewhat less widespread among people aged 65 and over). However, recent works have demonstrated that mobile phone data are robust to biases in terms of phone ownership (Wesolowski et al. 2013) and that even phone users with different phone usage

patterns do not have systematic differences in travel behaviour (Jiang et al. 2017). This verification confirms the validity of using a sample of phone users for expansion to the whole population. Moreover, while the HTS sample only consists of residents of the study area supplemented by specific passer-by surveys at particular zones (with the consequent questions that may arise in non-populated zones), the MPD sample also includes visitors within the region. This reduces the need for additional resources (e.g. intercept surveys) to draw an overall comprehensive picture of travel patterns within an OD matrix.

### *3.2.3. Feasibility and timeliness*

A main problem faced in conducting high-quality travel surveys is tied together with non-response and non-reported information. In general, it is difficult to find respondents who are willing and able to be interviewed. Even if they agree to participate, it is possible that the responses provided differ from real facts or are left incomplete. The reasons for respondents failing to report trips actually made are varied: unwillingness to devote time and effort to reporting certain activities, belief that specific trips are too insignificant to be reported, or, more frequently, some respondents simply forget trips they have made or forget to record them. These measurement errors may also be attributable to other causes such as the interviewer's training or the questionnaire itself. In addition to non-reporting of the correct number of trips, it is well known that respondents generally do not provide accurate details about other key components of travel, such as travel times, distances, and costs (Stopher and Greaves 2007); for instance, respondents typically report their travel time in quantum leaps of five- or ten-minute intervals. Mobile phone data play a key role in the context of all these errors. This source has the advantage of being collected passively, avoiding many of the abovementioned problems and even the difficulty of finding respondents who are willing and able to be interviewed, a key problem in surveys. Furthermore, another strength of mobile phone data is the timeliness of data collection, without the bias of survey techniques. In travel

survey data, a non-negligible rate of non-mobility users, that is, households and persons who report making no trips on the day of the survey, is frequently assumed. This last issue does not represent a dramatic problem because mobile phones provide real-time mobility data collected over several days from the repeated sample.

### *3.2.4. Level of detail*

It is clear that the determination of trips between origin and destination is mandatory to synthesise the travel demand over a region, but other data regarding socioeconomic and trip-making characteristics of individuals and households are extremely valuable to further understand the relationship between trip, travel choice, and scheduling of daily activities. A survey is essentially intended to yield data on the travel pattern of the residents of the household, providing information regarding the number of trips made, their origin and destination, the purpose of the trip, travel mode, time of departure from the origin, time of arrival at the destination, and so on. Certainly, it is possible to determine these aspects (directly or indirectly) from MPD. Advanced methodologies have been developed during the last decade to infer the transportation mode (Reddy et al. 2010; Horn et al. 2017; Semanjski et al. 2017; Phithakkitnukoon et al. 2017) or the purpose of the trip (Gong et al. 2014; Alexander et al. 2015; Jiang et al. 2017) from mobile data. Nevertheless, an HTS also obtains data on the general characteristics of the household that influence trip making (e.g. family size, vehicle ownership, occupancy and so on,) from which it is possible to relate the amount of travel per household and zonal characteristics and to develop patterns for trip generation rates. In this sense, at present, mobile technology is not able to provide these kinds of characteristics which are typically available from travel surveys (Wang et al. 2018), although there are efforts in that direction (Eftekhari and Ghatee 2016; Rahul and Winter 2016; Xiao et al. 2017; Cheng et al. 2017; Yin et al. 2017; Bwambale et al. 2017). Moreover, in the context of multi-stop tours associated with trip chaining (e.g. dropping off children at school before going to work), this

technology requires extra computational efforts with a high uncertainty level to detect stops (or stages) made during the trip, particularly if the stops are short in duration, while a travel survey is one of the few data acquisition procedures able to collect such information.

## 4. Discussion

This section is devoted to synthesising the main results discussed in previous paragraphs in order to provide a clearer understanding of the potential and difficulties of using mobile phone data in the field of matrix estimation. Table 2 summarises the main strengths and weaknesses.

**Table 2.** Main strength and weakness of the MPD source in the context of matrix estimation.

| Strengths | Weaknesses |
|---|---|
| Higher sample size; wider coverage to capture the extensiveness of OD relations and connections. | Low spatial resolution for matching detailed zoning systems and sparse representation in time. |
| Reduced time and cost of data collection and processing. | Under-reporting of short trips (distances of less than 2.5 km). |
| Timely data, with valuable information on temporary mobility patterns. | Lack of socioeconomic characteristics of travellers (e.g. income, family size, etc.) or even details of trips (e.g. number of stops, occupancy, etc.). |
| Feasibility data, automatically and passively produced. | |

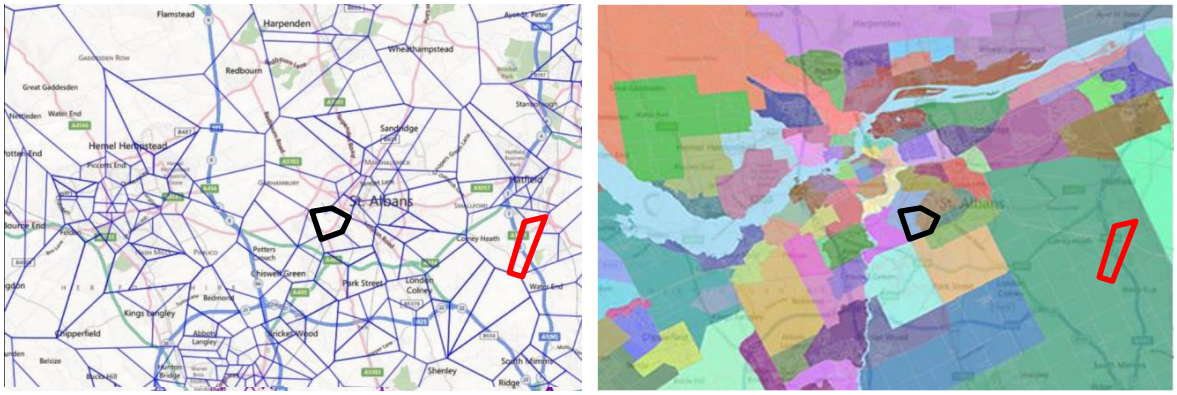
First and foremost, it is worth highlighting that the compared matrices come from sources with different technical bases and procedures for collecting, processing, and expanding data; in fact, both matrices may contain large errors (which are difficult to dimension in some cases). Moreover, although they were conducted over the same area of study and very close together in time, the ever-changing daily life and work patterns as well as possible roadway system modifications make it difficult to reproduce the exact numeric values. There is no solid support defined (i.e. ground truth), so it is difficult to say to what extent one is better than the other. However, based on the literature reviewed and the comparative analysis presented in the previous paragraphs, we can make the following observations:

- Mobile data offer a wider coverage, creating dense matrices in which a higher percentage of all possible OD movements in the studied area are covered.
- In fact, the wide representativeness of the MPD sample, which is homogeneously distributed across the territory, also provides an efficient alternative for sampling areas that are not properly represented in the HTS sample (e.g. non-populated zones embodying airports, train stations, or industrial parks) or even for visitors' movement within the region (not only residents), without additional resources or surveys.
- Moreover, many common features of travel behaviour have been detected, despite the highly diverse nature of these datasets. In terms of a structural correlation between MPD and HTS sources, the comparison has revealed homogeneity at both spatial and temporal scales; the structural patterns in matrices derived from MPD are very close to those captured in matrices derived from HTS. From the numerical point of view, both sources are highly correlated when the zoning system is rolled up to a lower granularity, not only for OD trips but also for trips by origins and by destinations.
- Focusing on non-short trips, the two sources reached similar ratios for the total number of trips and trip distribution.
- However, in the context of short trips, mobile technology in particular seems to present more difficulties in obtaining similar trip ratios than HTS sources. This fact is mainly motivated by the inherent nature of the events compounding the MPD source. As previously explained, the detection of movements using mobile technology is strongly dependent on the number of events generated by phones as they communicate with the network. For medium- and long-distance trips, the detection accuracy improves due to the fact that the number of available events increases in time and space (both active and passive events are used). But many short trips take less than 15 or even 30 minutes, presenting a reduced time window for the generation of mobile

events. Therefore, mobile events may be too sparse to determine these trips consistently. Of course, this does not mean that mobile technology is not able to capture short trips at all but merely indicates that short trips are more likely to be undetected than other trips according to the current state of this technology.

- Separately, there is a facet that can also affect the under-reporting of short-trips. This issue is associated with the matching algorithms to convert the footprints (generated by mobile events during movement) into origins and destinations within the zoning system. In the context of this comparative study, both mobility sources (MPD and HTS) are referred to a traditional zoning system based on polygonal sectors corresponding to administrative divisions (or transport zones). However, mobile technology has no information about the exact positions of handsets in the study area (like GPS technology) but only has information about regions linked to the service coverage area of antennas (or cells). This coverage area varies from site to site, ranging from hundreds of metres in urban areas to tens of kilometres in low population areas. Since the location algorithm is based on the triangulation of cell tower signals, mobility data are constrained by the density of cell towers in the studied area. In this study, focusing on a dense urban area, cell tower triangulation works pretty well. In particular, the spatial resolution is claimed to be about a few hundreds of metres, ranging from 200 to 300 metres. But here, the location is referred to the zoning system of the studied area, adding a bit more spatial uncertainty. In mobile systems, the antennas are distributed following a Voronoi tessellation to provide adequate radio coverage for communication rather than transportation planning criteria (such as homogeneous socioeconomic characteristics) used for the zoning system. Then, usually a cell (black polygon in Fig. 14a) overlaps not exactly with a typical transport zone but partially with two or more transport zones (black polygon in Fig. 14b).

Therefore, there are errors in the conversion of trips between cells and trips between transport zones. The larger the transport zone with which the cell has to be matched, the lower the matching error obtained (red polygon in Fig. 14). Moreover, mobile systems are designed to have an overlap between the cells to avoid coverage holes. Therefore, the complexity of matching 'footprints' provided by MPD with the zoning system is reduced by working at broader zones (i.e. lower granularity).




(a) (b)

**Fig. 14** Area division (a) based on Voronoi tessellation in mobile networks; (b) by zoning system based on administrative divisions.

Despite this last remark, matrices derived from MPD can be regarded as a complementary source of information that synthesises the overall picture of mobility over a region. In particular, when the zoning system is rolled up to a lower granularity, many common features in travel behaviour can be detected by cross-checking sources. Transportation studies have to be continuously adapted to the changing lifestyles in our society but also have to become much more economical to perform. In this sense, mobile data represent a valuable technology for transport modelling. The time and cost of data collection and processing are visibly reduced as data are automatically and passively generated by mobile users. Furthermore, despite the criticisms discussed in Sections 3.2.2 and 3.2.3 regarding sampling problems, this technology provides a means of sampling people's mobility at large population scales. Therefore, this data collection method allows

transportation studies to mitigate the effects of the major sources of errors in travel surveys: sampling, non-coverage, non-response, and measurement errors (Ortuzar and Willumsen 2011). This fact is also of paramount importance due to the growing difficulty of contacting and interviewing citizens regarding their travel behaviour. But nowadays, public administrations are demanding not only more updated information to analyse transportation strategies but also more detailed data. In this regard, surveys remain one of the most important ways of obtaining rich details about socioeconomic information, trip-making characteristics, or even stops. These pieces of information are extremely valuable for transport planning and decision making, and therefore it is envisaged that surveys will continue to be an integral part of transportation studies.

### *4.1. Data fusion approach*

Based on the previous discussion, the challenge is to develop data fusion methodologies able to obtain information with the optimum accuracy from a variety of heterogeneous sources. This involves taking advantage of the strengths of mobile data (e.g. extensiveness of OD relations and connections, representativeness), minimising their weaknesses (in particular inaccuracies for short trips) by means of other information sources like travel surveys. However, surveys do not have to come from exhaustive procedures intended to yield data on origin–destination flows, but on general characteristics of travel behaviour like total trips over the study region or trip distribution by distance, which are less costly. One possible way of addressing this issue can be seen as an OD matrix estimation problem, for which entropy maximisation and information minimisation principles have been commonly used (Wilson 1970; Willumsen 1978; Van Zuylen 1978; Van Zuylen and Willumsen 1980). These works demonstrated that by maximising the entropy, the most likely trip matrix could be estimated subject to a set of constraints. An attractive feature originally proposed by van Zuylen (1978) consisted of incorporating extra information (like a prior trip matrix) that might lead to a more

realistic estimate of the trip matrix. With this choice, the relative entropy $H$ of the estimated matrix $G$ with respect to a prior matrix $M$ is defined as follows:

$$H(G,M) = \sum_{i,j} g_{ij} \left( \log\left( \frac{g_{ij}}{m_{ij}} \right) - 1 \right) \tag{4}$$

where $g_{ij}$ is the estimated number of trips from transport zone $i$ to $j$, and $m_{ij}$ is the number of trips between transport zones $i$ and $j$ from the prior matrix. In this case, the daily MPD-based matrix is selected as the prior matrix due to the strengths mentioned in the previous sections. The estimated matrix $g_{ij}$ will have the same structure as the MPD-based matrix $m_{ij}$, conserving the extensiveness of OD relations and connections monitored by mobile technology. The entropy maximisation principle seeks to identify the most likely trip matrix consistent with the information available (Ortuzar and Willumsen 2011). However, additional information has to be included in the problem to correct inaccuracies detected in mobile data, specifically for short trips (which tend to be undercounted or completely overlooked in this source). This scheme is modelled by means of constrains based on information derived from the available travel survey. In the developed framework, the constraints are based on the total number of trips and the trip distribution by distance, provided by the HTS-based matrix. Therefore, the problem can be reformulated based on this information and the objective function, in a simplistic way as follows:

$$\begin{aligned}
&\min_{g_{ij}} \quad \sum_{i,j} g_{ij} \left( \log\left( \frac{g_{ij}}{m_{ij}} \right) - 1 \right) \\
&s.t. \\
&\sum_{ij \in Bin(b)} g_{ij} = P_b \cdot T \qquad \forall b \in \{1,...,|B|\} \qquad (a) \\
&\sum_{i \in I} \sum_{j \in J} g_{ij} = \sum_{i \in I} \sum_{j \in J} m_{ij} \qquad \forall I, J \in \{1,...,|MZ|\} \qquad (b)
\end{aligned} \tag{5}$$

Problem (5) resembles the classical double constrained distribution problem. In the expressions, the notation adopted uses the uppercase indices to denote macro-level zones,

whereas transport zone levels are represented by means of lowercase indices. The optimisation problem aims to find the OD matrix $g_{ij}$ with the closest similar structure to the mobile phone matrix, $m_{ij}$, using the criterion of minimising its relative entropy, which can be considered as a distance function between matrices *G* and *M*. Restrictions in (5) impose that the resulting matrix must fulfil a set of state constraints. The constraints defined by (a) will be referred to hereafter as 'histogram restrictions', and they are aimed at maintaining the trip distribution by distance provided by the HTS-based matrix. For its definition, it is necessary to distribute OD pairs according to the distance travelled (based on the network shortest distance between origin and destination centroids). Therefore each OD-pair is classified in a discrete number of intervals or bins, $|B|$, where $P_b$ is the proportion of trips in the distance range identified by bin *b*, and *T* is the total number of trips, both magnitudes provided by the HTS-based matrix. With the product of the proportion of trips $P_b$ and the total number of trips *T*, the number of trips for each bin is obtained directly. During the estimation procedure, the cells in the estimated matrix $g_{ij}$ are modified to fulfil the imposed 'histogram restrictions'. However, the prior matrix (MPD-based matrix) contains valuable information regarding the 'true' matrix, thus the estimation method pays careful attention to the distortion of the information contained in it. To control the distortion at each cell, of the estimated matrix (number of trips from transport zone *i* to *j*) with respect to the prior one, an additional set of constraints (b) is imposed. In this case, the constrains are built to maintain the number of trips at macro-zone level. As mentioned in section 3.1.2, the number of trips between macro-zones is more accurate when the zoning system is rolled up to a lower granularity. In fact, the two sources (HTS and MPD) reflect a strong linear relationship between the numerical values (trips) at macro-zone level, not only for OD trips but also for trips either by origins or by destinations. Hence, trips at macro-zone level have to be maintained during the estimation procedure. Thus, the number of trips between macro-zones *I* and *J* in the estimated matrix are

forced to be equal to those trips between the same macro-zones in the prior matrix, where *I*, *J* belong to the set *MZ* of 46 macro-zones defined in the study area. The analytical solution to the optimisation problem (5) is well-known, and it is expressed as:

$$g_{ij} = \alpha_{b(ij)} \cdot \beta_{IJ(ij)} \cdot m_{ij} \tag{6}$$

where the lengths of vectors of coefficients $\alpha$ and $\beta$ are respectively $|B|$ and $|MZ|$. The solution obtained corresponds to a multiplicative model, which implies that original zeros in the prior matrix $m_{ij}$ are kept as zeros in the final matrix $g_{ij}$, preserving the original structure of $m_{ij}$. According to this, the total number of trips $T$ is now distributed among the number of OD connections captured by mobile technology, which is clearly higher than those captured by surveys. As mentioned in Section 3.1.1, HTS-based matrices have to distribute this total number of trips among a substantially fewer number of elements; consequently, the number of HTS-based trips tends to be overestimated. The problem (5) can be solved by means of an iterative proportional fitting procedure, largely used in transportation studies (Furness 1965; Lamond and Stewart 1981; Erlander and Stewart 1990), whose proof of the convergence is reported in Bregman (1967). The estimated matrix G obtained preserves the strong linear relationship with the prior matrix regarding the numerical values (trips) contained in OD pairs at traffic zone level ($R_p = 0.91$); due to the imposed constraints (b), the Pearson's coefficient at macro-zone level is 1. At this point the inaccuracies related to short trips have been overcome in the estimated matrix with regard to the prior one. Fig. 15a presents the trip distribution versus distance, the estimated matrix (in cyan) gets similar travel rates to those from HTS source (in red), correcting the under-reporting of short trips from MPD (in blue). This correction is more easily appreciated in Fig. 15b.

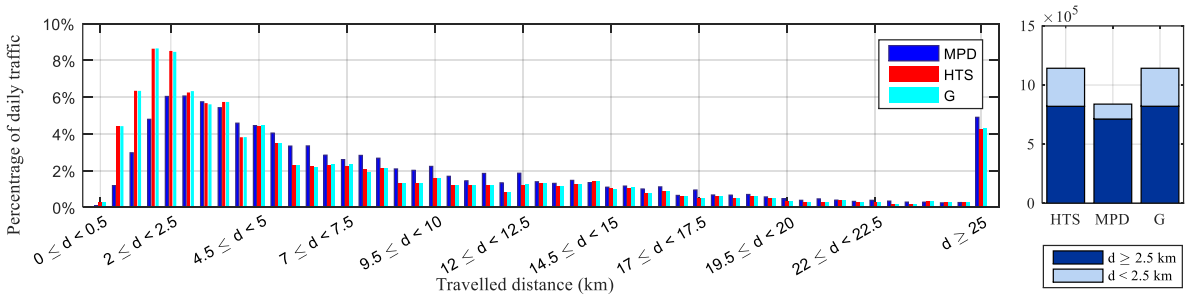



**Fig. 15** Trip distribution from the estimated matrix (G) and the two sources (MPD and HTS-based matrix): (a) percentage of daily trips by ranges (bins) of the distance travelled; (b) total daily trips.

## 5. Conclusions

In today's world, new technologies offer effective options for complementing or even replacing traditional transport data collection methods. In particular, the pervasive use of mobile phones has made this technology emerge as a promising alternative for generating OD matrices, traditionally extracted from survey-based approaches. Through passive and active events, footprints are generated that reveal the 'approximate' locations where people have been and the times at which they were there. A number of techniques have already been developed in the literature for converting these data into trips. But, in order to provide a clearer understanding of the potentialities and challenges of mobile phone data regarding traditional survey methods, this work has conducted a comparison among matrices derived from both types of sources over the same area of study: the urban agglomeration of Malaga, in the south of Spain. The conclusions have been derived by using statistical techniques and other methodologies like *MSSIM*.

As a result of the comparative analysis presented in this work and other experiences with mobile data that have been investigated and reviewed, one can conclude that mobile data can be used to draw a complete and representative picture of mobility flows over a region, especially when the zoning system is rolled up to a lower granularity. In particular, this comparative study has not only demonstrated the existence of many common features in the

travel characteristics reflected in traditional survey sources but also shown that the huge representativeness of this technology allows inherent problems in survey sampling frames to be overcome, capturing mobility in OD connections extensively with independence of the socioeconomic characteristics of the studied area. This fact is of paramount importance for monitoring mobility in non-populated areas such as mass transportation facilities, industrial parks, educational campuses, or hospitals, which are regularly excluded from the sampling frame (when no households are available to be interviewed). Hence, mobile technology can be used as a complementary information source for generating trip matrices using larger zones. However, as finer zoning systems are used, the use of mobile technology may raise questions about the accuracy of estimated trips, especially for short trips (which imply shorter distances or displacements made at faster speeds), which tend to be undercounted or completely overlooked in mobile data. Obviously, the importance that this fact may have in estimating the origin and destination flows depends on the nature of the studied problem itself. The severity is not the same in mobility studies at a national/regional level as in an urban context where the occurrence of this segment of mobility cannot be neglected. In this sense, the magnitude of the total number of trips is a concern when using mobile phone data, which should be addressed in conjunction with location-based services data generated by smartphone applications (with larger spatial precision) or other available data (e.g. outdated matrices or traffic sensor data) in order to extend and validate the collected data. The findings of this work form the basis for further research on developing data fusion methodologies to obtain the optimum accuracy from these heterogeneous sources. In such a framework, there are modelling tools and optimisation techniques that can also be implemented to minimise inaccuracies for short trips. The use of mobile data augmented with less costly travel surveys with the main aim of, for example, capturing the magnitude of short trips or even total trips over the study region (instead of exhaustive origin–destination flows) also deserves further

research. In any case, traditional surveys still constitute an extremely valuable source of information to be used in transportation studies because of the rich data they provide, especially in terms of socioeconomic characteristics. Only a full consideration of all available sources can lead to high-quality data collection results.

**Acknowledgments.** This work was supported by the ERDF of the European Union for financial support via the project DIURMOVIL under 'Programa Operativo FEDER de Andalucia 2011-2015', and by the Public Works Agency and Regional Ministry of Public Works and Housing of the Regional Government of Andalusia. One of the authors also acknowledges funding provided by the Spanish Ministry of Economy and Competitiveness through the Torres Quevedo Programme (PTQ-13-06428). The authors would like to thank two anonymous referees who provided valuable constructive criticism on an earlier version of this paper.

**Disclosure statement.** No potential conflict of interest was reported by the authors.

## References

Abrahamsson, T. (1998). Estimation of origin–destination matrices using traffic counts—a literature survey. Report IR-98021, International Institute for Applied Systems Analysis (IIASA).

Alexander, L.P., Jiang, S., Murga, M., and Gonzalez, M.C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C* 58(2), 240–250.

Ampt, E.S., and Ortuzar, J. de D. (2004). On best practice in continuous large-scale mobility surveys. *Transport Reviews* 24(3), 337–363.

Anda, C., Erath, A., and Fourie, P.J. (2017). Transport modelling in the age of big data. *Int. J. Urban Sci.* 21, 19–42.

Bonnel, P. (2003). *Postal, telephone and face-to-face surveys: how comparable are they?* In: Stopher, P.R., and Jones, P.M. (eds.), Transport Survey Quality and Innovation, Elsevier, London, pp. 215–237.

Bonnel, P., Fekih, M., Smoreda, Z. (2018). Origin-Destination estimation using mobile network probe data. *Transp. Res. Proc.*, 32, 69–81.

Bregman, L.M. (1967). Proof of convergence of Sheleikhovskii's method for a problem with transportation constraints. *USSR Comput. Math. & Math. Phys* 1, 191–204.

Brög, W., and Erl, E. (1999). Systematic errors in mobility surveys. In: Proceedings of the 23rd Australasian Transport Research Forum: Perth, Western Australia.

Bwambale, A., Choudhury, C.F., and Hess, S. (2017). Modelling trip generation models using mobile phone data: a latent demographics approach. *J Transp Geogr.*, https://doi.org/10.1016/j.jtrangeo.2017.08.020

Caceres, N., Romero, L.M., and Benitez, F.G. (2013). Inferring origin-destination trip matrices from aggregate volumes on groups of links: A case study using volumes inferred from mobile phone data. *J. Adv. Transp.* 47(7), 650–666.

Caceres, N., Wideberg, J.P., and Benitez, F.G. (2007). Deriving origin-destination data from a mobile phone network. *IET Intell. Transp. Syst.* 1(1), 15–26.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, Jr. J., and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transp. Res. Part C* 26, 301–313.

Cascetta, E., Papola, A., Marzano, V., Simonelli, F., and Vitiello, I. (2013). Quasi-dynamic estimation of O-D flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transp. Res. Part B* 55, 171–187.

Castillo, E., Jimenez, P., Menendez, J.M., and Nogal, M. (2013). A Bayesian method for estimating traffic flows based on plate scanning. *Transportation* 40(1), 173–201.

Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis . *Transp. Res. Part C* 68, 285–299.

Cheng, X., Fang, L., Hong, X., and Yang, L. (2017). Exploiting mobile Big Data: sources, features, and applications. *IEEE Network* 31(1), 72–79.

CNMC (2015). Spanish National Authority for Markets and Competition. Statistics on mobile communications. <http://data.cnmc.es/datagraph/jsp/inf_trim.jsp>

Cools, M., Moons, E., and Wets, G. (2010). Assessing quality of origin–destination matrices derived from activity and travel surveys: results from a Monte Carlo experiment. *Transp. Res. Rec.*, 2183, 49–59.

Day-Pollard, T., and Van Vuren, T. (2015). When are origin-destination matrices similar enough? In: Transportation Research Board 94th Annual Meeting, 15-1074, Washington, D.C.

De Regt, K., Cats, O., van Oort, N., and van Lint, H. (2017). Investigating potential transit ridership by fusing smartcard data and GSM data. *Transp. Res. Rec.*, 2652, 50–58.

Demissie, M. G., Correia, G., and Bento, C. (2015). Analysis of the pattern and intensity of urban activities through aggregate cellphone usage. *Transportmetrica A-Transport Science* 11(6), 502–524.

Diao, M., Zhu, Y., Ferreira, J., and Ratti, C. (2016), Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design* 43(5), 920-940.

Djukic, T., Hoogendoorn, S.P., and van Lint, JWC. (2013). Reliability assessment of dynamic OD estimation methods based on structural similarity index. In: Transportation Research Board 92nd Annual Meeting, 13-4851, Washington, D.C.

Doblas, J., and Benitez, F.G. (2005). An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix. *Transp. Res. Part B* 39(7), 565–591.

Eftekhari, H. R., and Ghatee, M. (2016). An inference engine for smartphones to preprocess

data and detect stationary and transportation modes. *Transp. Res. Part C* 69, 313–327.

Erlander, S., Stewart, N.F. (1990). *The gravity model in transportation analysis: theory and extensions*. Utrech,VSP.

Furness, K.P. (1965). Time function iteration. *Traffic Engng. and Control* 7, 458–460.

Frias-Martinez, V., and Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. Proceedings of the 5th International Conference on Information and Communication Technologies and Development, New York, pp. 76–84.

Gadzinski, J. (2018). Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. *Transp. Res. Part C* 88, 74–86.

Ge, Q., and Fukuda, D. (2016). Updating origin–destination matrices with aggregated data of GPS traces. *Transp. Res. Part C* 69, 291–312.

Gerike, R., Gehlert, T., and Leisch, F. (2015). Time use in travel surveys and time use surveys – Two sides of the same coin? *Transp. Res. Part A* 76, 4–24.

Global mobile consumer trends (2017). Second edition. Deloitte global mobile consumer survey.https://www2.deloitte.com/global/en/pages/technology-media-and-telecommunications/articles/gx-global-mobile-consumer-trends.html

Gong, L., Morikawa, T., Yamamoto, T., and Sato, H. (2014). Deriving personal trip data from GPS data: a literature review on the existing methodologies. *Procedia Soc Behav Sci.* 138, 557–565.

Herrera, J., Work, D.B., Herring, R., Ban, X., Jacobson, Q., and Bayen, A. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment. *Transp. Res. Part C* 18(4), 568–583.

Hofer C., Jäger G., and Füllsack M. (2018). *Generating realistic road usage information and origin–destination data for traffic simulations: augmenting agent-based models with*

*network techniques*. In: Cherifi C., Cherifi H., Karsai M., Musolesi M. (eds) Complex Networks & their Applications VI. Complex Networks 2017. Studies in Computational Intelligence vol. 689. Springer, pp. 1223–1233.

Horn, C., Gursch, H., Kern, R., and Cik, M. (2017). *QZTool – automatically generated origin-destination matrices from cell phone trajectories*. In: Stanton, N., Landry, S., Di Bucchianico, G., Vallicelli, A. (eds), Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing, vol. 484. Springer, Cham.

Iqbal, M. S., Choudhury, C. F., Wang, P., and Gonzalez, M.C. (2014). Development of origin–destination matrices using mobile phone call data. *Transp. Res. Part C* 40, 63–74.

Jensen, P., Rouquier, J.B., Ovtracht, N., and Robardet, C. (2010). Characterizing the speed and paths of shared bicycles in Lyon. *Transp. Res. Part D* 15(8), 522–524.

Jiang, S., Ferreira, J., and Gonzalez, M.C. (2017). Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. *IEEE Trans. Big Data* 3(2), 208–219.

Lamond, B., and Stewart, N.F. (1981). Bregman's balancing method. *Transp. Res. Part B* 15(4), 239–248.

Lee, R.J., Sener, I.N., and Mullins III, J.A. (2016). An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. *Transportation Letters* 8(4), 181–193.

Lu, S., Fang, Z., Zhang ,X., Shaw, S. L., Yin, L., Zhao, Z., Yang, X. (2017). Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. *Int. J. of Geo-Inf.*, 6, 7.

Malleson, N., Vanky, A., Hashemian, B., Santi, P., Verma, S.K., Courtney, T. K., Ratti, C. (2018). The characteristics of asymmetric pedestrian behavior: A preliminary study

using passive smartphone location data. *Transactions in GIS*, 2018(22), 616–634.

Mellegard, E., Moritz, S., and Zahoor, M. (2011). Origin/destination-estimation using cellular network data. In: IEEE 11th International Conference on Data Mining Workshops, Vancouver, Canada, pp. 891–896.

Meng, F., Wong, S.C., Wong, W., and Li, Y.C. (2017). Estimation of scaling factors for traffic counts based on stationary and mobile sources of data. *Int. J. ITS Res.* 15(3), 180–191.

Milne, D., Watling, D. (2019). Big data and understanding change in the context of planning transport systems. J. Transp. Geography (in press). Doi:10.1016/j.jtrangeo.2017.11.004.

Ni, L., Wang, X.C., and Chen, X.M. (2018). A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. *Transp. Res. Part C* 86, 510–526.

Nigro, M., Cipriani, E., and del Giudice, A. (2018). Exploiting floating car data for time-dependent Origin–Destination matrices estimation. *J. Intell. Transp. Syst.*, 22(2), 159–174.

Ortuzar, J. de D., and Willumsen, L.G. (2011). *Modelling Transport*. 4th ed. John Wiley & Sons Ltd, Chichester, West Sussex, UK.

Ortuzar, J. de D., Armoogum, J., Madre, J.-L., and Potier, F. (2011). Continuous mobility surveys: the state of practice. *Transport Reviews* 31(3), 293–312.

Parry, K., and Hazelton, M.L. (2012). Estimation of origin–destination matrices from link counts and sporadic routing data. *Transp. Res. Part B* 46(1), 175–188.

Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., and Bento, C. (2017). Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Sci.* 6: 11, 1–29.

Pollard, T., Taylor, N., van Vuren, T., and MacDonald, M. (2013). Comparing the quality of

OD matrices: in time and between data sources. In: Proceedings of European Transport Conference, Frankfurt, Germany.

Rahul, R.D, and Winter, S., (2016). Automated urban travel interpretation: a bottom-up approach for trajectory segmentation. *Sensors* 16(11), 1962.

Ranjan, G., Zang, H., Zhang, Z.L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mob Comput Commun Rev* 16(3), 33–44.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Trans. on Sensor Networks* 6(2), 1–27.

Richardson, A.J., Ampt, E.S, and Meyburg, A.H. (1995). *Survey Methods for Transport Planning.* Eucalyptus Press, Melbourne.

Rojas, M.B., Sadeghvaziri, E., and Jin., X. (2016). Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data. *Transp. Res. Rec.*, 2563, 71–79.

Ryley, T.J. (2008). The propensity for motorists to walk for short trips: evidence from West Edinburgh. *Transp. Res. Part A* 42(4), 620–628.

Santos, A., McGuckin, N., Nakamoto, H.Y., Gray, D., and Liss, S. (2011). Summary of travel trends: 2009 National Household Travel Survey. <http://nhts.ornl.gov/2009/pub/stt.pdf>

Semanjski, I., Gautama, S., Ahas, R., and Witlox, F. (2017). Spatial context mining approach for transport mode recognition from mobile sensed big data. *Comput. Environ. Urban Syst.*, 66, 38–52.

Seo, T., Kusakabe, T., Gotoh, H., and Asakura, Y. (2017). Interactive online machine learning approach for activity-travel survey. *Transp. Res. Part B.* doi:10.1016/j.trb.2017.11.009.

Sohn, K., and Kim, D. (2008). Dynamic origin–destination flow estimation using cellular communication system. *IEEE Trans. Veh. Technol.*, 57(5), 2703–2713.

Steenbruggen, J., Tranos, E., and Nijkamp, P. (2015). Data from mobile phone operators: A tool for smarter cities? *Telecommun. Policy* 39, 335–346.

Stopher, P., and Greaves, S. (2007). Household travel surveys: where are we going? *Transp. Res. Part A* 41, 367–381.

Toole, J.L., Colak S., Sturt B., Alexander L.P., Evsukoff A., and Gonzalez M.C. (2015). The path most traveled: Travel demand estimation using big data resources. Transp. Res. Part C 58, 162–177.

Van Zuylen, H. (1978). *The information minimising method: validity and applicability to transportation planning*. In: New Developments in Modelling Travel Demand and Urban Systems, G.R.H. Jansen et al. (eds.), Saxon, Farnborough, United Kingdom.

Van Zuylen, H.J., and Willumsen, L.G. (1980). The most likely trip matrix estimated from traffic counts. *Transp. Res. Part B* 14(3), 281–293.

Wang, F., and Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transp. Res. Part C* 87, 58–74.

Wang, Y., Correia, G.H.D.A., van Arem, B., and Timmermans, H.J.P.H. (2018). Understanding travellers' preferences for different types of trip destination based on mobile internet usage data. *Transp. Res. Part C* 90, 247–259.

Wang, Z., Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P. (2004). Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Image Processing* 13(4), 600–612.

Wang, Z. He, S.Y., and Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11, 141–155.

Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., and Buckee, C.O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *J R Soc Interface* 10(81), 20120986.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., and Gonzalez, M.C. (2015). Discovering urban activity patterns in cell phone data. *Transportation* 42(4), 597–623.

Willumsen, L.G. (1978). Estimation of an O-D matrix from traffic counts: A review. Institute for Transport Studies, Working paper no. 99, Leeds University.

Wilson, A.G. (1970). *Entropy in Urban and Regional Modelling*. Pion, London.

Wu C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., and Bayen, A. (2015). Cellpath: fusion of cellular and traffic sensor data for route flow estimation via convex optimization. *Transp. Res. Part C* 59, 111–128.

Xiao, Z., Wang, Y., Fu, K., and Wu, F. (2017). Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *ISPRS Int. J. Geo-Inf.* 6(2), 57.

Yin, M., Sheehan, M., Feygin, S., Paiement, J-F., and Pozdnoukhov, A. (2017). A generative model of urban activities from cellular data. *IEEE Trans. Intell. Transp. Syst.*, 99, 1–15.

Zhang, Y., Qin, X., Dong, S., and Ran, B. (2010). Daily O–D matrix estimation using cellular probe data. In: Transportation Research Board 89th Annual Meeting, 10–2472, Washington, D.C.

Zhou, X., and Mahmassani, H.S. (2006). Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Trans. Intell. Transp. Syst.*, 7(1), 105–114.