Traffic Flow Estimation Models using Cellular Phone Data

N. Caceres, L. Romero, F.G. Benitez and J.M. del Castillo

***Abstract*— Traffic volume is a parameter used to quantify demand in transportation studies, commonly collected by using on-road (fixed) sensors such as inductive loops, cameras, etc. The installation of fixed sensors to cover all roads is neither practical nor economically feasible, so they are only installed on a subset of links. Cellular-phone tracking is an emerging topic developed and investigated during the last few years to extract traffic information. Cellular systems provide alternative methods to detect phones in motion without the cost and coverage limitations associated with those infrastructure-based solutions. Utilizing existing cellular systems to capture traffic volume has a major advantage compared with other solutions, since it avoids new and expensive hardware installations of sensors, with a large number of cellular phones acting as probes. This research proposes a set of models for inferring the number of vehicles moving from one cell to another by means of anonymous call data of phones. The models contain in their functional form terms related to the users' calling behavior and other characteristics of the phenomenon such as hourly intensity in calls and vehicles. A set of inter-cell boundaries with different traffic background and characteristics were selected for the field test. The experiment results show that reasonable estimates are achieved by comparing with volume measurements collected by detectors located in the same study area. The motion of phones while being involved in calls can be used as an easily accessible, fast, and low-cost alternative to derive volume data on inter-cell boundaries.**

***Index Terms*— Vehicle Traffic Flow, Traffic Counts, Cellular Phone Data.**

I. Introduction

TRADITIONALLY, traffic counts are one type of information that can be automatically collected on a subset of links in a network, commonly by on-road (fixed) sensors, such as inductive loop, magnetometer, visual camera, etc. Measurements of on-road sensors are available with little effort, but they are not sufficient due to their limited coverage and expensive costs of implementation and maintenance.

Moreover, those sensors are subject to errors, which can degrade considerably the volume estimates. A complementary method for collecting traffic data is therefore needed. This paper proposes to utilize location data generated from cellular phones travelling along the network in order to estimate traffic flows, so that cellular systems appear as a complementary solution to fixed sensors. While this idea has its limitations, as discussed later in this paper, it has the potential of collecting data on moving travelers over a large coverage area without requiring expensive infrastructure investment.

In modern societies cellular phones have reached high penetration rates, with many countries surpassing the 90% [1]. In cellular networks, such as Global System for Mobile Communications (GSM), the service area is covered by a set of base stations whose radio coverage area is called a cell. A cell radius depends on parameters, such as antenna type, power level, or even topology and surrounding buildings. It is smaller in urban areas – where the people density is high and more antennae are necessary to provide good communication services – than in rural areas, varying from several hundred meters to several kilometers. A set of adjacent cells is grouped into one location area (LA). Mobility management includes processes that automatically keep the databases updated with the location of phones in order for phones to make or receive calls or messages. In the case of a phone involved in a call, the cellular system always knows the base station (cell) to which the phone is connected. When a phone with a call in progress moves from one cell toward another, the call is transferred to the new cell by means of the handover procedure in order to provide an uninterrupted service. In the case of a phone that is turned on but not on call, for efficiency and flexibility, the system does not need to know the precise cell all the time, but rather the LA. The Location Update (LU) procedure is the mechanism to retain the exact LA of the phone. Then, the execution of either a handover or a LU procedure inserts a record into the databases to update them with content consisting of the phone identification (ID), location (LA, cell), or event timestamp. Based on this, the motion of phones travelling along the network is monitored. More technical details regarding cellular networks can be found in [2].

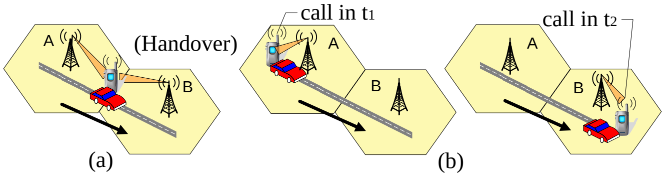


Fig. 1. In-motion calls: (a) handover and (b) two calls in t_1 and t_2, t_2-t_1≤T.

II. TRAFFIC INFORMATION FROM PHONE DATA

A. *State of the Art*

The idea of using cellular phones to collect traffic information is a decade old [3]–[5] and it has become increasingly widespread. Some works presenting analysis of the exploitation of cellular phones for traffic monitoring and reviewing the state of the practice can be found in [6] and [7]. Regarding the use of phone data, we can refer to simulated experiments carried out to evaluate the potential of using mobile data for estimation of travel demand [8] and for space-based passing time estimation [9], or even studies using information retrieved from a software platform for the evaluation of urban dynamics based on the anonymous monitoring of phone movements [10]. In terms of volume, the concept of cellular phones as probes has been explored by various researchers by means of field tests [11]–[14]. In all cases, volume data would be associated with phone transit through a boundary area, detecting crossing rates either at the inter-cell boundary level (handover) or at the location area boundary level (the LU process). The LU procedure is executed by any switched-on phone to keep the system informed about changes in the LA. Therefore, it manages a large amount of sample data. Although the use of call events does not reach a sample size as large as in the previous case, it offers greater precision with regard to the location. Thus, call data have been utilized in this research, and most of research works in literature also focuses on handovers to detect phone transit through the boundaries between two cells. The main quantitative findings show that phone flow (calls) is related to the flow of vehicles measured by loop detectors installed in the network, having similarities for most of an ordinary day [11],[12]. Obviously, there are striking differences between the two flow curves, but they exhibit peaks in the morning and afternoon rush hours. Therefore a relationship exists between vehicle flow changes and phone habits. However, those studies concluded that accurate vehicle flows cannot be obtained directly from cellular phone data due to the characteristics of this source data. The main question is how the number of crossing phones is correlated with the real number of crossing vehicles. Volume data on inter-cell boundaries provided by cellular phones does not yield information on the complete set of vehicles crossing a boundary, but only a statistical sample of all the travelling vehicles. Some vehicles may carry either phones of other operators or switched-off phones; these phones are not detected as crossing phones. These aspects imply special treatment to correlate the two measures (phone counts and vehicle counts), which may require a calibration stage to be performed using vehicle volume data. In this regard, different works have performed procedures to bring phone counts to vehicle counts based on empiric transfer functions [13] or correction factors [14] obtained by means of loop data. In both, a relatively good overall fit was found but detectors located spatially in the same monitored section were required.

In this paper different models to infer volumes of vehicles from (anonymous) cellular phone call-data are proposed in order to be used in transport applications. The calibration of these models uses volume collected by a set of loop detectors. After this, models work for any road crossing inter-cell boundaries without physical detectors.

B. Estimation of the cellular phone count

Given a vehicle travelling on a road with a phone involved in a call, the handover happens when the vehicle is moving from one cell to another (Fig. 1a). The handover record for the entry of the on-call phone into a new cell can be used to detect the mobility of a phone in order to count in-motion phones. However, the use of only handover data has certain limitations. If the call ends before it is handed over to the new cell, the handover is not performed, and hence the phone movement is not detected. Additional situations that increase the number of in-motion phones detected must be considered.

For billing purposes, when a call occurs the system always inserts a record into its databases including the parameters related to the call such as start/end time of the call, duration, caller phone number, or identification of the originating cell (cell ID). When a phone makes two consecutive calls in different cells, cells A and B (Fig. 1b), within a short period of time, the stored records are useful to identify the phone as in-motion one. Then, if these cells are neighboring, the inter-cell mobility of a phone is also detectable by analyzing those call records, as in the case of handover. The difference in the start time of these calls, called T, cannot be too large; if T is high, the phone may move to other cells before making the second call in cell B. For this field test, several periods were used to avoid such situations, and taking into account the average service area of the involved cells, finally the period of 15 minutes was chosen. This double-call event, together with the handover event, increases the number of in-motion phones detected. The calls associated with these mobility situations are named in-motion calls, classified as:

1) those in which a phone has an active call and changes from cell A to cell B, that is, a handover event (Figure 1a);

2) those in which a phone makes two consecutive calls in cell A and cell B in the time period T (Figure 1b).

Due to disruptive effects of cell phone conversations on driving [15], some countries prohibit the use of cell phones while driving, unless a hands-free device is employed. So any vehicle carrying a passenger (not driver) with a cellular phone or a hands-free device can be regarded as a probe when making in-motion calls. The analysis of call data records is therefore useful for detecting the transit of phones from one cell to another, that is, phones crossing an inter-cell boundary. Those phones travel on board vehicles moving along the roads, so that a cellular system is turned into a kind of count station (hereinafter so called as virtual traffic counter, VTC) to monitor the flow of phones at inter-cell boundaries. The map of intersection areas between inter-cell boundaries and roads has to be previously identified, for example, by using coverage service information obtained from the operator or by using a laptop equipped with special software to track down the cell identity and with a Global Positioning System (GPS) receiver to record the location coordinates.

Fig. 2 illustrates possible monitored routes. For the case of the boundary between cell 1 and cell 2, a single roadway connects the two cells, so the counted phones belong to users travelling along route 3. The situation in which a unique roadway connects two cells does not occur frequently in mobile networks. In most cases, the road network is denser than the cell distribution, so multiple links connect two cells. For instance, two routes (1 and 2) connect cells 1 and 3; there are two roadways crossing the boundary between cell 1 and cell 3. In this case, it can only be asserted that the phone moves along one of the roadways crossing the inter-cell boundary but without identifying which one. Thus, the number of phones counted by the proposed approach contains those moving along all the routes connecting the two cells, and hence the volumes derived from such a number of counted phones are also provided in aggregate format.

C. *Monitored boundaries*

By analyzing the road network and cell layout, it is possible to identify the map of roadways (or routes connecting cells) that can be observed. This initial processing defines a set of boundary candidates to be monitored, then new criteria for the choice of "monitored boundaries" should be applied. For this purpose, a transport network model is used. This model is a simplified representation of the road network and consists mainly of nodes and links, which represent intersections and road sections, respectively. The road network shown in Fig. 2 is simplified into a set of nodes and links in Fig. 3. According to this representation, there are four possible VTCs located at inter-cell boundaries so that only links whose starting and ending nodes are located in different and neighboring cells are taken into account when composing a monitored boundary.

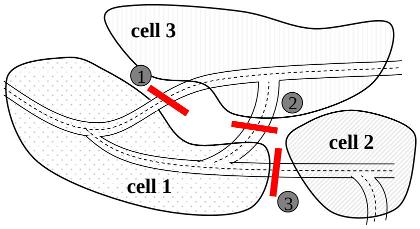


Fig. 2. Inter-cell boundaries to be monitored.

When multiple roads (links) connect two cells, the number of phones counted by using in-motion calls contains those moving along all the existing routes. The flow is measured on a group of links and hence flow measurements are provided in aggregate format for each group of links crossing inter-cell boundaries. However, some inter-cell boundaries might not be always suitable for being monitored; it is necessary for the starting and ending nodes of crossing links to be clearly located in different cells to compose a monitored boundary. Cellular systems are designed to have an overlap between the cells in order to provide a better quality of communications, avoiding coverage holes. Thus, cell overlap has to be assessed to ensure the above condition is accomplished.

Cell coverage area

The cell boundaries are not static, but dynamic to some extent; they depend on the number of phones, the direction of travel, and the network type (3G, 4G, etc.). Besides, each cell coverage area has an effective radius where a phone can communicate with a unique base station (Fig. 4a). This radius may vary according to certain random factors (weather, call load …). However, this uncertainty in radius size does not impact on the identification of a "monitored boundary" since only links whose starting and ending nodes are located inside the effective cell coverage area are taken into account when creating a valid "monitored boundary" (Fig. 4b). Fluctuations in cell boundaries do not affect as long as those nodes remain inside the effective coverage area, as can be seen in Fig. 4c.

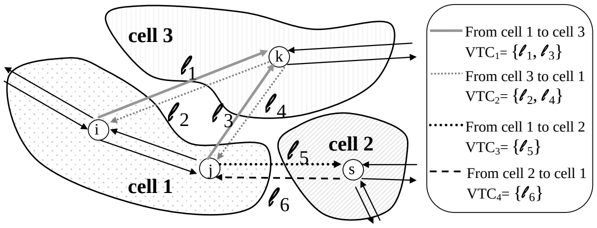


Fig. 3. Location of 'virtual' traffic counters (VTC).

Additionally, when designing the cell layout for a cellular system, the cells overlap at the edges to prevent shadows in coverage. The cell overlapping area is defined as the overlap between adjacent cells with regard to the primary coverage. The size of the overlapping area is a design parameter that depends on measurement control parameters. According to standard design criteria, most network planners agree that the typical cell overlapping area may occupy about 20–30% [16]. Since the typical radius in non-urban environments is around 1–30 km, the overlapping area between two cells is larger than the area covered by a transport node (e.g. intersection). However, when a node is located in an overlapping area (Fig. 4d), the condition of the monitored boundary is not matched; it belongs to two cells, one of them being the same as where the other node is located. Thus, this boundary, and the associated links, are not valid for observation.

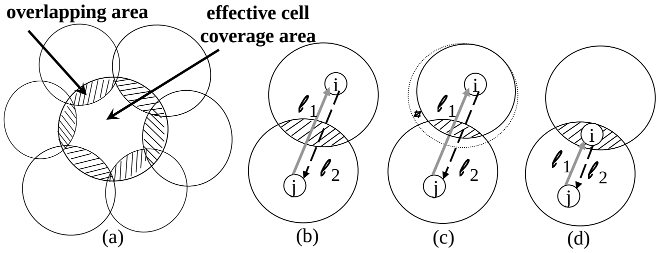


Fig. 4. Cell coverage area.

Valid monitored boundaries

Once feasible monitored boundaries have been identified, a selection process is performed to choose the set of valid ones among all the candidates. The selection process is aimed to avoid possible indeterminacies during the phone counting. The valid monitored boundaries are finally chosen regarding certain conditions such as adequate cell coverage of the road/link observed, absence of uncertainties regarding alternative roads (local roads are not included in the network model) or the absence of pedestrian and railway traffic.

Regarding this last condition, it is necessary to highlight that any phone making in-motion calls can act as a probe. In this research, the interesting probes are those located in vehicles moving along a motorway (passengers with a phone or drivers using a hands-free device). So that the identification of the means of transport used by phones (pedestrian and vehicular users) has been deliberately excluded by concentrating the attention on cells whose boundaries have mainly vehicular traffic, neither pedestrian nor train users. Different approaches based on advanced algorithms have been developed, implemented and tested in order to identify the means of transport of moving phones [17],[18]. Although these works have reached good results, there are still certain problems to be solved. In this research, the phone records are only used for reporting movement between cells to be related with vehicles. After identifying the valid set of boundaries between cells, phones moving along routes connecting the two cells, that is, roadways crossing inter-cell boundaries are monitored by a VTC located at such boundary.

III. Field Test

The absence of pedestrian and train traffic conditioned the choice of cells to be monitored, concentrating the attention on cells whose boundaries have mainly vehicular traffic. After combing information about cell distribution and road network, six points, outside of the metropolitan area of Seville, corresponding to roads with different traffic background and characteristics were selected for the field test (Fig. 5). This creates twelve monitored boundaries (every kilometre point has two monitored boundaries, each being associated with a crossing direction). The dataset used in this test consisted of all “outgoing calls” (calls initiated by the user, not received calls) recorded in the study area, which were provided by a Spanish operator. For every outgoing call, the dataset included the exact time of the call, the encrypted identification (ID) of the phone, duration, and the Cell ID to which the phone was connected while the call was active. Besides, an additional parameter is included in the collected call data which is related to the reason for the call drop in a cell, being one of them the handover event. By using this parameter the measure of handovers from the data was derived. The encrypted ID for every phone was a unique and randomized number based on original phone ID, in compliance with privacy regulations. This encryption procedure was made directly by the operator. Traffic volumes measured by detectors located next to such boundaries were also used (provided by the Spanish Traffic Management Centre, DGT). These two types of data were collected every day for 6 weeks, but only core weekdays (Tuesdays to Thursdays) were extracted. The attention only focuses on this type of day to avoid the effects of non-working days on daily traffic patterns. Hence holidays (weekend days) and days before and after a holiday were discarded from the sample data. Then, the models are calibrated using historical data collected over 18 days of regular traffic patterns.

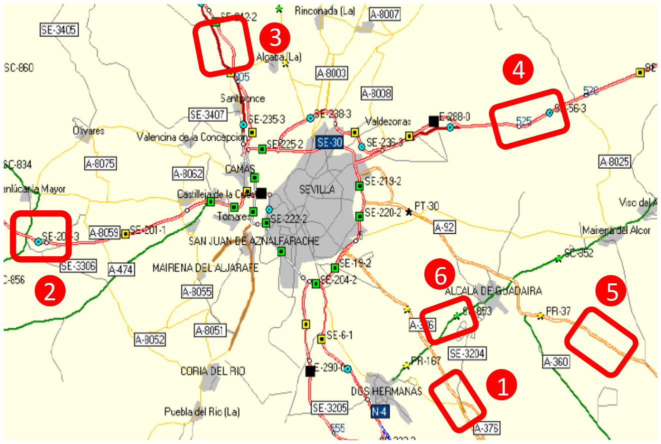


Fig. 5. Point location of studied roads where there are inter-cell boundaries.

Cellular phone activity strongly depends on time. It is clear that there are periods when users are more likely to make calls; for example, daytime hours as opposed to night-time hours. From the data collected for this research it may be noted that the number of calls is significantly high in the period from 08:00 to 21:00 hours according to the coefficient of variation. This means that outside this period call data have no statistical significance to be able to make any inferences. This fact is not a problem at all given that traffic flow estimates outside that period is not of great interest, with the exception perhaps of the rush-hour period between 07:00 and 08:00 hours. Although the traffic flow starts to become significant in this period, the number of calls before work (especially before 07:30) is very low since there is less demand for people to call. Hence, the observed time period is focused on 08:00 to 21:00 hours.

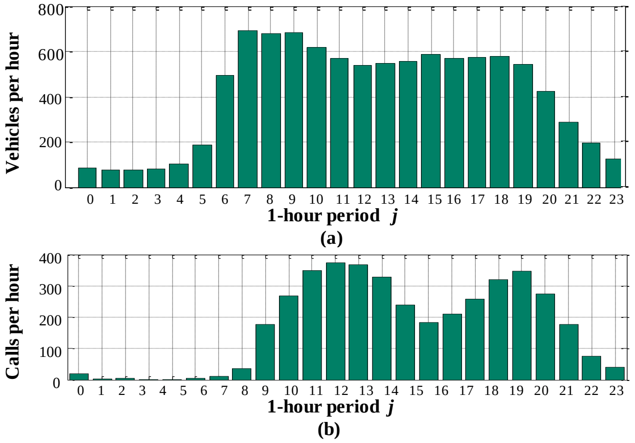
Finally, a sample was available for each day comprising the pair of data: the number of calls and the number of vehicles recorded for 13-hour periods existing in the period from 08:00 to 21:00 hours. By virtue of the transferability of findings to other locations where there are no loop detectors, a larger sample was obtained for each point by aggregating the 18 days into a single sample. The sample was divided into two sets: calibrating data and testing data. The calibrating set was used to estimate the model’s parameters for the complete definition of the models. A significant relationship between typical calling behaviour and patterns of human activity is present. The calibrating stage aims to incorporate implicitly mobility patterns and call activity associated with the typical daily routine in the region under study (the start of working hours, lunch time...) in the functional form of the models with the purpose of application to any boundary without requiring a recalibration process for each one. The testing set was employed for the evaluation and comparison of those models. The notation $Y(i,j,k)$ was used to designate the number of vehicles counted at kilometre point k on day i during 1-hour period j, and $X(i,j,k)$ the number of calls recorded on day i during 1-hour period j between the cells of which the boundary corresponds to kilometre point k, for the purposes of estimating the vehicle traffic flow. Note that Spanish laws prohibit the use of cell phones while driving, unless a hands-free device is employed. So, the study was developed assuming that the probes, that is, the phones that makes the "in-motion" calls detected, are those belonging to passengers or drivers using a hands-free device.

IV. Model Definition

A. Selection of variables

The scope of this approach is to correlate the vehicular volume (Y) in a given road cross-sectional point with the number of in-motion calls (X) generated at the associated inter-cell boundary. A set of models is proposed for providing vehicle volumes based on phone data in a manner similar to that of classic counting stations. It is necessary to highlight that there is no high linear dependence between the two magnitudes (Pearson's correlation coefficient, R=0.34). However, it can be expected that an increase in call volume is directly proportional to the number of phone users, although this relationship is not constant during different hours of a day; for instance, the number of phone calls in the late afternoon may increase due to cheaper "price plans". So additional terms must be considered in designing models. Logically, volumes inferred from phone call data are strongly affected by the calling users' behaviour; furthermore, there are periods when users are more likely to make calls. It is important to consider the dependence with the time period in the formulation of the models since the two variables implicated in the process, calls and vehicular flow, vary with time in a differentiated manner.

Fig. 6 shows the temporal intensity variation in the number of calls and the number of vehicles crossing a given boundary on a specific day. The time periods with high intensity are close in time in both cases, but do not coincide. Such behaviour is extendable to other boundaries. In particular, users have differing call-making habits according to the time of day due to criteria such as "price plans" or admissible call-making times. In terms of vehicular flow, this fluctuates over time as a function of various criteria, such as the rush hour at the start of the working day, lunch hours, et cetera. Vehicular flow starts to become significant in the 6:00–8:00 period, although the number of calls is relatively low due to it being regarded as a non-admissible call-making time (too early). Something similar occurs in the 14:00–16:00 interval but in that case it seems to be explained by the lunch break (this activity depends much on the country habits). Thus, the availability of phone call data (input data for the model) significantly differs at different daytime hours. In this sense predictions must include a term associated with such hourly variability of both call intensity and vehicular intensity. The models presented below introduce time-associated factors that capture temporal dependence. These factors f_j and g_j are obtained by means of the observed data $X(i,j,k)$ and $Y(i,j,k)$, aggregating for all days and boundaries, eliminating spatial dependence and giving rise to more versatile models.



(a)

(b)

Fig. 6. Specific one-day sample of the numbers of calls and vehicles counted.

First, the vehicle intensity factor, f_j, associated with time variation in the vehicular flow is defined in (1) as the ratio between the mean number of vehicles counted in a 1-hour period and the mean number of vehicles counted for the entire observed time period (08:00-21:00). To establish a factor common for typical Spanish traffic patterns, their numerical values are obtained using hourly traffic counts provided by the DGT. In particular, a total of 315 points observed continuously (24 hours a day, 365 days per year) were used.

$$f_j = \frac{\sum_{k=1}^{K}\sum_{i=1}^{D} Y(i,j,k)}{\frac{1}{H}\cdot\sum_{j=1}^{H}\sum_{k=1}^{K}\sum_{i=1}^{D} Y(i,j,k)};\ \begin{cases} H = 13\text{-hour periods}\\ D = 365\text{ days}\\ K = 315\text{ observed points}\end{cases} \tag{1}$$

The call intensity factor, g_j, is defined in (2) as the ratio between the probability of a vehicle making a call in the 1-hour period j, P_j, and the probability averaged for the entire observed time period. Note that the call data used come from a unique operator that has delivered them. So that the value of P_j refers not to make any call, but a call supported by the concrete operator. The explanation of how P_j is calculated is given in the next section.

$$g_j = \frac{P_j}{\frac{1}{H}\cdot\sum_{j=1}^{H} P_j}\quad \text{with } H = 13\text{-hour periods of interest} \tag{2}$$

Then, the numerical values of factors f_j and g_j and the probability P_j are used as coefficients in the functional form of the models. Their hourly values are shown in Fig. 7 and 8. Those coefficients make the volume estimations dimensionless for all the time periods since they weigh the observations of in-motion calls in each time period according to their relative importance within the considered time periods. It should be remarked that these factors f_j and g_j, and the probability P_j are obtained empirically from historical data. Their numerical values can be considered valid for a certain time (eg: 6 months, 1 year, etc.) since the characteristics of the sample are more or less stable. However, it would be desirable to perform periodically a calibration of these factors and probabilities in order to remodel the underlying changes in individual behaviour, such as travel patterns or calling activity. It occurs because, even within a 6-month period, people may change their home or job location, car ownership level, income and other factors, as well as changing phone use styles (penetration rate, daily call volumes, market shares, etc.). Each of these changes requires or enables them to alter the factors and probabilities, which will be used in the functional form of models.

B. *Model formulation*

Regarding the model formulation, the existence of a relationship between typical call activity and traffic mobility is considered. The curves of phone flows (calls) and vehicle flows (measured by loop detectors) have similarities for most of an ordinary day (Fig. 6). Both curves exhibit peaks in the morning as well as in the afternoon rush hours (e.g. the vehicle flow increase in the morning precedes the increase in the number of calls). After analysing one of the most typical families, a set of models combining the coefficients f_j and g_j with the in-motion calls by means of exponential and polynomial functions is proposed. In all cases, the dependent variable is $y(i,j,k)$, which reproduces the number of vehicles estimated that have crossed boundary k in the 1-hour period j on day i. The independent variable, $X(i,j,k)$, is the number of in-motion calls observed in the hour period j on day i at boundary k. The following models are defined:

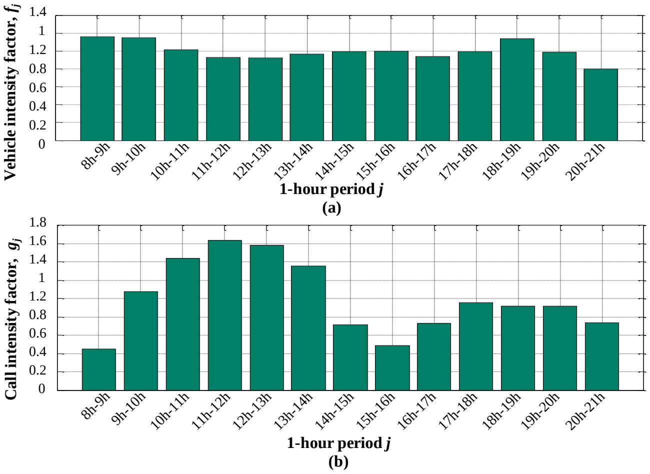


Fig. 7. Intensity factors f_j and g_j, (a) and (b) respectively, at 1-hour period j.

M1) Cobb–Douglas Model $y_1(i,j,k) = a \cdot f_j^{\phi} \cdot g_j^{\beta}$

This model depends on the intensity factors by means of three parameters $\{a, \phi, \beta\}$, and lacks data regarding the number of in-motion calls observed. It has been proposed for the purpose of evaluating the importance of in-motion calls in inferring volume.

M2) Modulated Cobb–Douglas Model

$$y_2(i,j,k) = \left[a + bX(i,j,k)\right] \cdot f_j^{\phi} \cdot g_j^{\beta}$$

This model with four parameters $\{a, b, \phi, \beta\}$ introduces a difference with respect to the previous one: a first-order dependence on the number of in-motion calls observed, providing a prediction of the number of vehicles moving from one cell to another as a function of the number of in-motion calls observed.

M3) Second-Order Modulated Cobb–Douglas Model.

$$y_3(i,j,k) = \left[a + bX(i,j,k) + cX^2(i,j,k)\right] \cdot f_j^{\phi} \cdot g_j^{\beta}$$

This model with five parameters $\{a, b, c, \phi, \beta\}$ is similar to the previous model; however, it establishes a second-order dependence on the number of in-motion calls.

In these three models, f_j and g_j are fixed hourly coefficients integrated into their functional form; the unique input variable is $X(i,j,k)$. In order to study the importance of incorporating intensity factors f_j and g_j into the formulation, the following two models without both factors are also considered:

M4) Linear Model $y_4(i,j,k) = a + bX(i,j,k)$, parameters $\{a, b\}$

M5) Quadratic Model: $y_5(i,j,k) = a + bX(i,j,k) + cX^2(i,j,k)$, parameters $\{a, b, c\}$.

Finally, a last model has been defined based on an approximation of the physical phenomenon of in-motion calls. In this, the probability of a vehicle with a given distribution of passengers and phones making in-motion calls is quantified.

M6) Physical model: $y_6(i,j,k) = f\left(X(i,j,k), P_j, Q_{j,k}\right)$ with five parameters $\{a, b_1, b_2, c, d\}$.

It depends on the number of in-motion calls observed, $X(i,j,k)$, the probability of a vehicle making a call in the hour period j, P_j, and the probability of handover in the hour period j at boundary k, $Q_{j,k}$. The detailed derivation of this model is developed in the next section.

Model based on the physical phenomenon

The previous models contain some of the most used functions in their functional form for vehicle volume inference by means of call data. Those models make use of terms introduced to capture features of the studied phenomenon such as the hourly variability of both call and vehicular intensity. However, a model that reliably represents the physical relationship between the number of in-motion calls and the number of vehicles may play an important role in achieving accurate estimations of vehicle volumes using such input data. The formulation of the physical model is founded on the hypothesis that a cellular phone makes an in-motion call when either the user makes a call in each of the two cells that form the boundary within a short period of time (event M, Fig. 1b) or the user has an active call and changes from one cell to another (event L, Fig. 1a). With regard to the event M, although it requires the making of two calls, only one is counted as an in-motion call. The concept of mobility is only associated with the phone from which the calls needed for the detection of movement were made.

In view of these situations, it seems logical that the observations of in-motion calls are affected by the tendency to make calls. A typical user tends to make calls during a certain time periods, but does not call in others, such as late night or early morning. In the case of handovers, the duration of the calls will also affect these observations, given that the handover is more likely to occur as the call duration increases when the phone is moving. Therefore, the time dependence in user behaviour when making calls must be taken into account when formulating a model that allows the estimation of the number of vehicles using a given number of in-motion calls.

In order to formulate this physical model, an expression that models each of the aforementioned situations is proposed. The nature of the cells under study, i.e. those away from urban environments, allows the assumption that practically all of the observed in-motion calls will be made by users on board vehicles travelling along the existing roads at that boundary. These users therefore constitute a sample of the vehicle population that has crossed the boundary *k* in the hour period *j* on day *i*, $Y(i,j,k)$. With regard to the first type of in-motion call, the success probability of event M can be modelled in a simplified manner as P_j^2, that is, the making of two calls during the hour period *j* taking into account that P_j is defined as the probability of "making a call on board a vehicle, using the monitored operator" in the hour period *j*. One should bear in mind that the process of making a call is not independent of the fact of having made a previous call recently. However the sample does not deal with the whole population of phone users, but with a particular group of users: those travelling on board a vehicle. The calling activity patterns of this sample of users may be considerably different from those of the entire population. In order to simplify the modeling of this type of in-motion calls, we assume that the two calls are uncorrelated in spite of a bias may arise with this assumption. Having assumed that the population that generates the in-motion call is the set of vehicles crossing a boundary, $Y(i,j,k)$, the number of in-motion calls generated by event M is given by:

$$X_M(i,j,k) = Y(i,j,k)\cdot P_M \approx Y(i,j,k)\cdot P_j\cdot P_j = Y(i,j,k)\cdot P_j^2 \qquad (3)$$

For the second type of in-motion call (event L), the proposal is similar but a term is introduced related to the likelihood that the call performs a handover. Thus, its probability of occurrence requires two terms: on one hand, P_j, the probability of "making a call on board a vehicle, using the monitored operator", and on the other, the probability that the call requires a handover, so-called probability of handover, *Q*. Regarding the latter, the probability that a handover is performed on a call comes from the fact that its duration, t_d, exceeds the time of permanence in the cell, t_p. Assuming that the population that generates the in-motion calls is the set of vehicles crossing the boundary *k* during the hour period *j* on day *i*, $Y(i,j,k)$, the number of in-motion calls created under conditions associated with event L is defined as:

$$X_L(i,j,k) = Y(i,j,k)\cdot P_L = Y(i,j,k)\cdot P_j\cdot Q \qquad (4)$$

After modelling the situations involving in-motion calls, the total number of observed in-motion calls in the time period j corresponds to the sum of both sets of calls, $\{X_M$ and $X_L\}$.

$$X(i,j,k) = Y(i,j,k)\cdot P_j^2 + Y(i,j,k)\cdot P_j\cdot Q = Y(i,j,k)\cdot[P_j^2 + P_j\cdot Q] \qquad (5)$$

In order to formulate a theoretical model, the relationship between the number of in-motion calls and the number of vehicles is represented by $P(j,k)$, that is the probability that a phone makes an in-motion call (not any call), which may be of type 1 (associated with handover) or type 2 (associated with consecutive calls in different cells within a short time period). According to (5), the probability $P(j,k)$ can be approximated by $P(j,k)= P_j^2+P_j\cdot Q$, by considering certain events as negligible, such as the making of a call in each cell while a handover also occurs. Although these events are feasible, they did not occur on any occasion in the sample. Therefore, the number of vehicles that cross a boundary *k* during the hour period *j* on day *i* will be given by:

$$X(i,j,k) = P(j,k)\cdot Y(i,j,k) \Rightarrow Y(i,j,k) = \frac{X(i,j,k)}{P_j^2 + P_j\cdot Q} \qquad (6)$$

This expression provides a value for the volume of vehicles that cross the boundary between cells in terms of the detected in-motion calls and other variables related to the calls. The number of in-motion calls associated with the crossing of a boundary during the given time period, $X(i,j,k)$, is obtained by analysing the records, provided by the network operator, of calls made in the cells involved in the boundary *k*. For the other terms, i.e. the probability that a vehicle makes a call and the probability of a handover, P_j and *Q*, respectively, an additional statistical treatment is required to take into account the dependence relation of time with the call characteristics. The following points explain in detail the method developed to obtain the aforementioned terms.

1) Probability of making a call on board a vehicle

According to the literature, a fair number of works has successfully explored the calling activity of cellular phone users by measuring the inter-event time distribution (phone calls and SMS, sent or received) [19],[20], without regard to the users' mobility status. However, in our context, the studied event is not to make any call, but a call on board a vehicle. In this regard, the findings documented in the literature might not be applicable to this particular event. So that, we preferred to estimate a term of probability of occurrence by an empirical procedure in which the assignment of the probability of the event is based on the observed information. Using this empirical focus, the probability is determined based on the proportion of times in which a favourable or successful event occurs with regard to the total number of possible results. In this case, the studied event was "making a call on board a vehicle supported by the operator providing call data".

On one hand, as it is shown in Fig. 6b, there are periods when users are more likely to make a call. Therefore, the particular case of "making a call on board a vehicle, using the monitored operator", P_j, also varies with time. On the other, in order to investigate possible dependencies on space (location), this empirical approach was performed separately for each pair of cells. The successful event considered takes place when an in-motion call occurs, being the number of in-motion calls directly extracted from the collected call data. Besides, the total number of possible results, which is related to the number of vehicles moving along each inter-cell boundary, is known by means of counting stations that were near such a boundary. The outcome showed that the estimated values of the probability P_j for each location remained within similar ranges for the same time period. This coincidence was attributed to the similarity of calling activity of users who travel along the associated roadways. The cells mainly support freeway traffic, away from sites such as residential areas, shopping areas, commuter hubs, etc. In these last cases, the behaviour of making a call drastically changes from site to site since each of them serves users with considerably different call patterns. By contrast, no significant factor of those above mentioned influences the calling activity of users over the studied cells, showing a regular, stable behaviour. In these cells the calling behaviour is not affected by the site activity, so that it seems reasonable to assume that the locations may be aggregated. Then, the process was repeated but aggregating all locations into a single sample. Logically, an aggregate scheme provokes masked errors, but this alternative was selected for the transferability of findings to other locations with similar traffic features, and for increasing the sample size.

Finally, and bearing in mind that an in-motion call originating from event M implies the making of two calls from the handset that is travelling on board a vehicle, the probability of “making a call on board a vehicle, using the monitored operator” during the time period j is defined as:

$$P_j = \frac{\sum_{i=1}^{D}\sum_{k=1}^{K}\left[X_L(i,j,k) + 2\cdot X_M(i,j,k)\right]}{\sum_{i=1}^{D}\sum_{k=1}^{K} Y(i,j,k)};\begin{cases} j = 8h, 9h, ..., 21h \\ D = 18 \text{ days} \\ K = 12 \text{ points}\end{cases} \quad (7)$$

It is necessary to comment that those calls are made by only a sample of all the phones on board vehicles. Some vehicles may carry more than one phone, either from other operators or switched-off, although either various calls or none may be made from the same vehicle. This feature is already included in the calculation of P_j since its expression is based on the ratio of calls to vehicles, not phones to vehicles. The strong time dependence when a user making any call logically influences this probability of "making a call on board a vehicle". Fig. 8 plots the variation over time of probability P_j empirically obtained. This probability P_j is valid for any of the monitored boundaries seeing as the event “making a call on board a vehicle, using the monitored operator” is independent of the cell in which it occurs, except perhaps in zones where driving is difficult for making a call. Fig. 8 also reflects that the probability of an in-motion user (on board a vehicle) making a call follows the trends of daily call activity shown in Fig. 6b (for any user) for most of the time periods. There are also two pronounced peaks during rush hours. The morning peak centres around the same time period in both cases, while the evening peak for calls made by in-motion users is less pronounced and wider. This lower tendency may originate from the fact that an in-motion user only makes a call outside of working hours when it is necessary to do so. However, the tendency to make calls seems to be independent of user mobility during working hours.

2) Probability of handover

The likelihood of a handover being performed must be modelled. In order to obtain an expression for the probability of handover, this work assumes a simplified scenario for handover where a cellular system covering a road network is divided into regular cells. The distance that a phone travels within a cell before crossing the boundary of the said cell is modelled as a random variable with uniform distribution throughout the interval [0, *L*] metres. The permanence time of a cellular phone in the cell, t_p, if the phone moves at a constant speed of *V* m/s will also be a random variable with uniform distribution throughout the interval t_p [0, *L*/*V*] seconds. Likewise, the call duration, t_d, can be modelled as an exponential random variable of mean T_c seconds. The probability density functions of t_p and t_d will be given by:

$$f(t_p) = \frac{1}{L/V}, \quad t_p \in \left[0, \frac{L}{V}\right]; \quad f(t_d) = \frac{1}{T_c} e^{-\frac{t_d}{T_c}}, \; t_d \in \left[0, +\infty\right) \quad (8)$$

A call requires the execution of a handover procedure when its duration exceeds the permanence time of a phone in the cell. Thus, the probability of a handover can be calculated using a conditional probability.

$$Q = \int_0^{\frac{L}{V}} P(t_d > t_p \mid t_p) f(t_p) dt_p, \text{ with } P(t_d > t_p \mid t_p) = \int_{t_p}^{\infty} f(t_d) dt_d = e^{-\frac{t_p}{T_c}} \quad (9)$$

Defining a dimensionless factor *α*, we obtain:

$$Q = \int_0^{\frac{L}{V}} e^{-\frac{t_p}{T_c}} f(t_p) = \int_0^{\frac{L}{V}} e^{-\frac{t_p}{T_c}} \frac{1}{L/V} dt_p = \frac{1}{\alpha}\left[1 - e^{-\alpha}\right], \quad \alpha = \frac{L}{V \cdot T_c} \quad (10)$$

The factor *α* depends on the length *L*, the speed *V* and the mean call duration T_c. The parameter *L* represents the distance that a phone must travel within a cell until it enters another cell (crosses the boundary). The smaller this distance the easier is that the call must execute a handover. This length depends on the roads within the origin cell of the boundary. Therefore the dependence on the distance is captured through this factor. Something similar occurs with the speed: the value depends on the type of road in question (motorway, main road, etc.). Although speed can vary along the same road according to the time period, due to the level of saturation for example, it will be considered uniform within the origin cell of the boundary. Therefore, the value of *α* will very much depend on the boundary *k* in terms of the length and speed associated with the type of road that runs through the origin cell. Similarly, bearing in mind the call duration’s dependence relation on time, *α* will also depend on the analysed time period. The handover probability, *Q*, is therefore a function of the time period *j* and the boundary *k*; that is, $Q_{j,k}$.

In order to characterize the origin cell of each boundary *k* in terms of length *L* and speed *V*, this case study used basic functionalities in the cellular system that allow us to know the cell identifier (antenna) to which a phone is connected. With the help of a GPS and a simple application implemented from mobile devices, a correspondence between the position on the road and the cell that provides it with service was established. Based on this, the length of the road along which a phone travels within a cell was found. However, this measure can be directly provided by the operator, using a GIS tool and its coverage information. For the speed *V*, the average speed for each observed road were considered. Finally, it is possible to obtain empirically the call duration through the analysis of call data provided by the operator, determining a mean value of call duration according to the considered time period, $T_c(j)$.

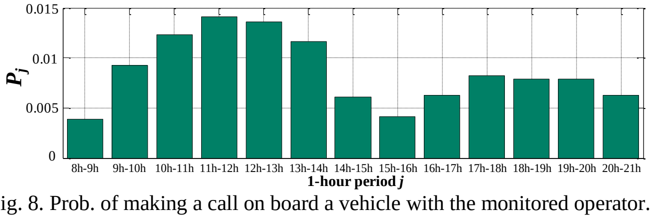


Fig. 8. Prob. of making a call on board a vehicle with the monitored operator.

This duration, which depends on time period j, and the values of speed and length that correspond to the road that passes through the origin cell of the boundary, allow the definition of the handover probability as a function that is variable over time j and the boundary k, $Q_{j,k}$. Fig. 9 shows the temporal distribution of the probability of handover calculated using (10) for the characteristics of L and V at a specified boundary. This plot reveals that there are periods when calls are more likely to perform handover. These periods are generally associated with larger call duration due to cheaper "price plans" or higher speed due to the lack of traffic congestion.

By substituting the expression of $Q_{j,k}$, the equation (6) can now be written as:

$$Y(i,j,k) = \frac{X(i,j,k)}{P_j^2 + P_j \cdot \frac{1}{\alpha(j,k)}\left[1-e^{-\alpha(j,k)}\right]}, \quad \alpha(j,k) = \frac{L_k}{V_k \cdot T_c(j)} \qquad (11)$$

where $\alpha(j,k)$ introduces the mean characteristics of the highways related to inter-cell boundary k – average speed, V_k, distance travelled in the cell, L_k, and average call duration in interval j, $T_c(j)$ – into the functional form. Finally, the expression of the physical model is obtained by introducing a set of parameters $\phi=\{a,b_1,b_2,c,d\}$ to address the indeterminacies introduced by the hypotheses required to generate the model, being formulated as:

$$y_6(i,j,k) = \frac{a \cdot X(i,j,k)}{P_j^2 + P_j \cdot \frac{b_1}{\alpha(j,k)}\left[1-e^{-b_2 \cdot \alpha(j,k)}\right] + c} + d \qquad (12)$$

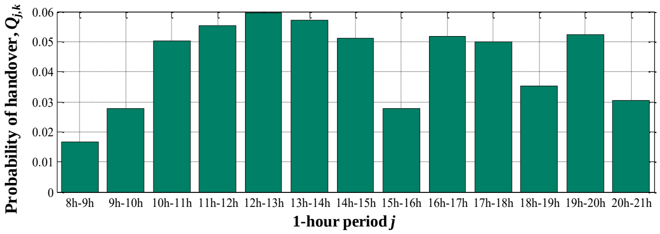


Fig. 9. Probability of handover for a specific boundary in each 1-hour period.

The term P_j is the fixed hourly coefficient associated with the probability of “making a call on board a vehicle, using the monitored operator” at time period j, determined by (7), and $\alpha(j,k)$ is a coefficient related to the probability of handover in the hour period j at boundary k, $Q_{j,k}$. Details of the formula derivation may be seen in [21].

C. Parameter estimation

Once the models proposed to infer vehicle volumes have been formulated, a calibrating stage is carried out for the parameter estimation. The calibration process uses only the sample data corresponding to the calibration set $\{(X^1,Y^1), ..., (X^N,Y^N)\}$. Another subset of the data, so-called testing dataset, is reserved for the contrast of the models. There are different procedures for estimating the parameters that define each model. The adjustment criterion based on the principle of minimization of the sum of the absolute relative error between the observed and the modeled values has been utilized, by solving the following optimization problem for each model: where Y^i is the observed volume, $y^i=f(X^i; \Phi)$ is the modeled one using observed in-motion calls, X^i, and $\Phi=\{a,b_1,b_2,c,d\}$ is the set of the model parameters. Subsequently, the models were evaluated and compared in order to select those providing accurate estimates of vehicles crossing a boundary, Y, as a function of the known value of X (in-motion calls at the said boundary). After estimating model parameters, the models were completely defined.

V. Model Selection

A set of six models has been proposed to estimate vehicle volume using phone call data. The estimates from all models for a number of observed in-motion calls were evaluated using real vehicle volumes measured by detectors, whose values are contained in the "testing dataset". The assessment was carried out on the basis of different criteria, such as error measures expressed in absolute values, percentiles, and correlation between the volumes estimated by each model and the observed by a detector that were near such a boundary. Next, the models were compared between them for selecting those that produced the best results. The most appropriate models were those providing the best balance between all the criteria. Table I shows the measurements achieved for each model. The mean absolute error and the mean absolute relative error, *MAE* and *MARE* respectively, allow a comparison between the estimates and the real values using a classical error analysis. Absolute values are used in order for the prediction error to have the same significance either upwards or downwards. Based on these measures, the best models are 6, 2, and 3, seeing as the ranges are of the same order in all three. Another criterion examined is the cumulative distribution function of the absolute relative error using percentiles. A percentile is the value of a variable below which a certain percentage of observations fall. The best models are those that show the smallest values for each one of the percentiles. The achieved percentiles also show that models 6, 3, and 2 are better than the others. The 50th percentile or median of the absolute relative error (*MedARE*) shows error levels around 17%.

Another criterion evaluated is the correlation between the estimates and the real volume values. For this purpose, the linear correlation coefficient, or Pearson coefficient, and the Spearman rank correlation coefficient were studied. The Pearson coefficient is the most widely used measure of linear relationship between two variables, while the Spearman rank correlation coefficient is a measure of the monotone association between two variables using the relationship between ranks (e.g. a positive Spearman coefficient corresponds to an increasing monotonic trend between

TABLE I
Error measurements for each model

	M1	M2	M3	M4	M5	M6
MAE	237.2	210.66	210.02	223.98	223.24	203.6
MARE	0.23	0.20	0.20	0.22	0.22	0.20
MedARE	0.20	0.17	0.17	0.18	0.18	0.16
RC	0.25	0.51	0.51	0.40	0.40	0.57
LC	0.28	0.47	0.48	0.34	0.35	0.53

RC: Rank Correlation (Spearman); *LC*: Linear Correlation (Pearson) variables). In terms of rank correlation, models 6, 2, and 3 clearly stand out from the others, especially model 6, which achieved high values for the sample size examined. A similar ranking is achieved using the linear correlation coefficient.

Consequently, models 1, 4, and 5 are discarded for inference purposes. Model 1 depends exclusively on the factors related to time variability in the intensity of calls and vehicles. No information about the number of in-motion calls appears in the functional form, losing information regarding a proportion of vehicles crossing a boundary. Something similar occurs for models 4 and 5 but losing significant information on time variability due to the absence of intensity factors. Thus, models 2, 3, and 6 are regarded as the most suitable for the estimation of vehicle flow using the number of in-motion calls. The graphical comparison between the vehicle flows observed and estimated for these three models at various inter-cell boundaries is shown in Fig. 10. The plot reveals that the estimates follow the peaks and valleys of the observed flow curve within an admissible error level (*MARE*<20%, *MedARE*<17%). The three models show very similar estimates of vehicle flow for a known number of in-motion calls and the differences between the estimates of both models are barely appreciable graphically.

Note that higher error levels are obtained from late afternoon/early evening. This result might be explained by several factors, for example, the beginning of cheaper "price plans". This factor has a strong effect on calling behavior (duration and number of calls) which fluctuates greatly in these time periods from day to day; hence error levels increase. The vehicle occupancy is another important factor affecting the accuracy of the estimation model. There are significant differences in vehicle occupancy by time of day; the morning peak period has lower average vehicle occupancy than the mid-day period and the evening peak period. The vehicle occupancy for going to work during the morning peak period is regular on a daily basis. From evening hours, aside from returning home, there are numerous other activities in which people engage on a less-than-daily basis, such as visiting friends or relatives, shopping, entertainment, fitness, and so on. People may engage in these activities alone or accompanied, and no stable trends exist in vehicle occupancy for these time periods. Since it is difficult to exactly detect the vehicle occupancy using only call data, the model has been developed using an average vehicle occupancy rate. The model accuracy may change when vehicle occupancy is drastically different from this rate. By contrast, these error levels decrease to 6–8% in periods when the trends in making calls and vehicle occupancy are more stable (9:00–14:00).

Another aspect to be considered in selecting a model for inference is its complexity. The fit of any model can be improved by increasing the number of parameters; however, variance (uncertainty) increases as the number of parameters in a model increases. In statistics, the Bayesian information criterion (*BIC*) [22] and Akaike information criterion (*AIC*) [23] are well-known methods of assessing model fit penalized by the number of parameters. These criteria can be viewed as measures that combine fit and complexity. Then, several competing models may be ranked according to their *AIC* and *BIC*. Then, although the error measurements are of the same order of magnitude in those three models, the ranking reveals that the model 2 and 6 are the best ones. A more comprehensive analysis [21] shows that these models yield the best vehicle-flow estimates. Then, models 2 and 6 are the ones selected for volume inference. However, it is difficult to establish a clear priority between them since:

i) model 6 reaches a slight improvement with respect to model 2 in terms of *MAE*, *MARE* and *MedARE*, although their values are of the same order of magnitude;

ii) graphically, estimates from the models reflect similar behavior in terms of reproducing the peaks and valleys in the observed volume curves;

iii) the linear and rank correlation coefficients reveal that model 6 reaches a better fit to the real data than model 2.

Additionally, model 6 achieves reasonable flow estimates although it previously requires the boundary characterization in terms of the speed and length of roads running through the cell of origin. This model is less flexible than model 2 since it requires the cell-boundary characterization (average speed and

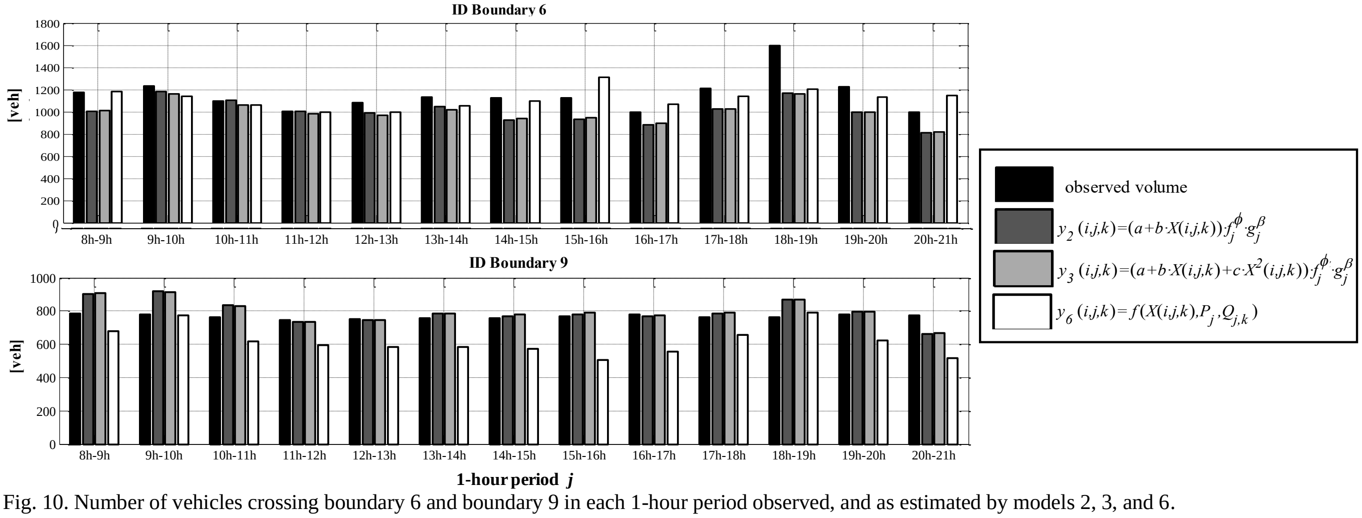


Fig. 10. Number of vehicles crossing boundary 6 and boundary 9 in each 1-hour period observed, and as estimated by models 2, 3, and 6.

length); however, its predictive capacity, evaluated in Table I, also qualifies it as a suitable model. As a consequence models 2 and 6 will be those finally selected for estimating vehicle flows due to their provision of the best results.

VI. Conclusions

The use of traditional on-road sensors (e.g. inductive loops, cameras, etc.) for collecting traffic flows is necessary but not sufficient because of their limited coverage and expensive costs of implementation and maintenance. This paper presents anonymous call data generated from moving phones as an alternative or rather complement source of high quality data for estimating traffic flows. In particular, the paper proposes a methodology for estimating vehicular volumes crossing an inter-cell boundary, that is, the number of vehicles moving from one cell to another. The main advantages of using cellular systems are that phone data can be acquired widely, and no additional implementation within the cellular network infrastructure as well as the phone is necessary. Besides, no additional costs arise, being more cost-efficient than other techniques such as local loop-data or video-based systems. These classic infrastructure techniques have limited economic conditions, so that only a restricted infrastructure and no area-wide availability is established. After discussing the required data for this purpose, six models have been developed. The data permit the determination of the said volumes by employing i) the in-motion calls generated in each 1-hour period, ii) additional information associated with the characteristics of the vehicular traffic, iii) the call features, such as hourly intensity, call duration, and iv) even the characteristics of the highways crossing the boundary.

Adjustment of the parameters tasked with modeling dependence between the variables implied has proceeded, seeking minimization of the sum of the absolute relative error. The work has been completed with a comparative analysis of the models using criteria such as error measurements, rank and linear correlation coefficients and statistical criteria for assessing model fit. Based on the foregoing reasoning models 2 and 6 have finally been considered as equally viable in estimating vehicular flows, highlighting:

i) the need to incorporate data in the functional form of the models regarding the time variability in the behavior of users travelling, in making calls, or other characteristics associated with vehicular traffic, and

ii) the applicability of the methodology to a broad set boundary within admissible error levels in comparison with measurements provided by counting stations.

To sum up, cellular phones can be regarded as a complementary solution to fixed sensors in order to enhance the available information for mobility monitoring. It is necessary to emphasize that the procedure performed to process the incoming data and infer volumes requires non-negligible computing time. Hence, the models are intended to be used for applications in which the estimation process does not need to be performed in real time. In addition, the error levels achieved by the proposed methodology impede the use of these models for applications in which accurate volume measurements are required. For real-time applications or traffic management purposes such as incident detection it is recommended that traffic monitoring system based on cellular phones should be merged with other systems to get a reliable and complete advanced traffic information system; whereas cellular phones can provide accurate enough information for applications without real time requirements. An example is in the field of Origin–Destination matrix estimation.

The most commonly used models for updating travel demand make use of volume data observed on links in the transport network and other available information (often contained in a prior matrix), so the prior matrix may be “adjusted” or “changed” to reproduce observed volumes. Customarily, observed volumes are expressed in terms of the mean number of vehicles per type of day (working, weekend, or even all days), which come from permanent loop detectors embedded in the road or vehicle identification technologies. The loop data have two main types of errors [24]; first, the detectors tend to undercount vehicles. In most cases this error is less than 10% of the real volume. Secondly, detectors tend to count vehicles in neighboring lanes in addition. In some cases the share of the additionally counted vehicles has a 15% of divergence. The standards defined are that the total traffic volume should not vary from reality by more than 20% [24]. Then, the error levels obtained using the estimation model are within the limits for fulfilling the standards. Moreover, demand matrices are studied for time periods representing one or two hours, usually morning peaks, and the achieved error levels are lower during these morning hours. So, volumes inferred from cellular phones may also be regarded as an attractive option for matrix estimation [25]. In that case, the usage of flow inferred from cellular phone data for updating the matrix requires a reformulation of the model notation to use volume on groups of links (the most common case using cellular system criteria for selecting the observed location), rather than on single links (the traditional format of detectors). Indeed, an estimation algorithm combining volume data from cellular phones with automatic traffic counts based on traditional techniques (e.g. detectors) will allow the achievement of more realistic matrices.

Acknowledgment

This project was financed by the Spanish Ministry of Science through R&D National Programmes (ENE2008-05552) and by Vodafone Spain through the Minerva Project (1C-021). The authors acknowledge the collaboration of A.L. Jimenez from the Spanish Traffic Management Centre (DGT).

References

[1] CIA World Factbook. List of countries by number of mobile phones in use. [Online] Available: https://www.cia.gov/library/publications/the-world-factbook/index.html.

[2] M.D. Yacoub. *Wireless Technology: Protocols, Standards and Techniques*. Boca Raton, FL. CRC Press, 2002.

[3] R. Bolla and F. Davoli. "Road traffic estimation from location tracking data in the mobile cellular network," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2000, vol. 3, pp. 1107–1112.

[4] Yilin Zhao. Mobile phone location determination and its impact on intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 55 - 64, Mar. 2000.

[5] C. Drane and J.L. Ygnace. "Cellular telecommunication and transportation convergence: A case study of a research conducted in California and in France on cellular positioning techniques and transportation issues," in *Proc. 4th Int. IEEE Conf. Intell. Transp. Syst.*, Oakland, CA, 2001, pp. 16–22.

[6] Z. Qiu, J. Jin, P. Cheng, and B. Ran. "State of the art and practice: Cellular probe technology applied in ATIS," in *Proc 86th Transp. Res. Board Annu. Meeting*, Washington, D.C., Jan. 2007, paper no. 07-0223.

[7] N. Caceres, J.P. Wideberg, and F.G. Benitez. "Review of traffic data estimations extracted from cellular networks," *IET Intelligent Transport Systems*, vol. 2(3), pp. 179–192, Sept. 2008.

[8] K. Sohn and D. Kim. "Dynamic origin-destination flow estimation using cellular communication system," *IEEE Transaction on Vehicular Technology*, vol. 57, no. 5, pp. 2703–2713, Sept. 2008.

[9] K. Sohn and K. Hwang, "Space-based passing time estimation on a freeway using cell phones as traffic probes," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 559–568, Sept. 2008.

[10] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata and C. Ratti. "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141-151, Mar. 2011.

[11] K.U. Thiessenhusen, R.P. Schäfer, T. Lang. "Traffic data from cell phones: A comparison with loops and probe vehicle data," presented at the ITS World Congress, 2006, CD-ROM, paper no. 1550.

[12] M. Höpfner, K. Lemmer, and I. Ehrenpfordt. "Cellular data for traffic management – First results of a field test," presented at the ITS Europe Conference, 2007, CD-ROM paper no. 2407.

[13] S. Bekhor, M. Hirsh, S. Nimre, and I. Feldman. "Identifying spatial and temporal congestion characteristics using passive mobile phone data," in *Proc 87th Transp. Res. Board Annu. Meeting*, 2008, paper no. 1534.

[14] M. Friedrich, P. Jehlicka, T. Otterstätter, J. Schlaich. "Mobile phone data for telematic applications", presented at International Multi-Conference on Engineering and Technological Innovation, IMETI 2008.

[15] Y. Liang, M.L. Reyes and J.D. Lee. "Real-time detection of driver cognitive distraction using Support Vector Machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 340-350, June 2007.

[16] UMTSWorld. UMTS Network Coverage Planning. [Online]. Available: http://www.umtsworld.com/technology/coverage.htm

[17] S. Reddy, J. Burke, D. Estrin, M. Hansen, M. Srivastava. "Determining transportation mode on mobile phones," in *Proc. 12th IEEE International Symposium on Wearable Computers*, 2008.

[18] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, Funchal Portugal, 2010, pp. 318-323.

[19] M. Gonzalez, C. Hidalgo, A.L. Barabasi. "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779-782, June 2008.

[20] F. Calabrese, F. C. Pereira, G. Lorenzo, L. Liu, C. Ratti. "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events," *Pervasive Computing*, LNCS 6030, Springer, 2010, pp. 22-37.

[21] N. Caceres. "Mobility matrix estimate by using mobile phone data (In Spanish)". Ph.D. Thesis Dissertation, University of Seville (Spain), 2010. [Online]. Available: http://www.esi2.us.es/GT/docs/TesisNCS.pdf

[22] G. Schwarz. "Estimating the Dimension of a Model," *Annals of Statistics*, vol. 6, pp. 461–464, Mar. 1978.

[23] H. Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proc. 2nd Int. Symposium on Information Theory*, Budapest, Hungary, 1973, pp. 267-81.

[24] N. Lehnhoff. "Quality of automatic data collection with loop detectors," in *Proc. of the 2nd International Symposium Networks for Mobility*, 2004, Stuttgart, Germany.

[25] N. Caceres, L.M. Romero, F.G. Benitez. "Inferring origin–destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data," *Journal of Advanced Transportation*, Oct. 2011, doi: 10.1002/atr.187.

Noelia Caceres was born in Don Benito, Spain, in 1980. She received the Ph.D. degree in Telecommunication Engineering from the University of Seville, in 2010. She is currently a Senior Researcher with the Department of Transportation Engineering, University of Seville. Her research interests include transport demand modelling, traffic flow, software design and ITS.

Luis M. Romero was born in Badajoz, Spain, in 1971. He received the Ph.D. degree in Industrial Engineering from the University of Seville, in 2007. He is currently a Senior Researcher with the Department of Transportation Engineering, University of Seville. His research topics are transport demand modelling, traffic flow, numerical methods, ITS and software design.

Francisco G. Benitez was born in Seville, Spain, in 1956. He received the Ph.D. degree in Industrial Engineering from the Technical University of Madrid, in 1981. He was a Research Fellow at the Department of Engineering Science, University of Oxford in 1981-83, and a Fulbright Scholar and Visiting Associate at the California Institute of Technology 1984-1986.

He is currently a Professor with the Department of Transportation Engineering, University of Seville, and Head of Transportation Engineering and Infrastructure Division. His research topics are transportation modelling, transmissions, GIS, numerical methods, ITS.

Jose M. del Castillo was born in Seville, Spain, in 1965. He received the Ph.D. degree in Industrial Engineering from the University of Seville, in 1994. He was a Postdoctoral Researcher at the Institute of Transportation Studies at the University of California-Berkeley, in 1995. He is currently an Professor with the Department of Transportation Engineering, University of Seville. His research topics are applied statistics, transportation modelling, heuristic logistics and traffic flow.