IET Intelligent Transport Systems

Research Article

Experimental analyses and clustering of travel choice behaviours by floating car big data in a large urban area

ISSN 1751-956X
Received on 7th January 2018
Accepted on 1st February 2018
doi: 10.1049/iet-its.2018.0015
www.ietdl.org

Gaetano Fusco[1] ✉, *Agnese Bracci*[1], *Tommaso Caligiuri*[2], *Chiara Colombaroni*[1], *Natalia Isaenko*[1]
[1]*Department of Civil Environmental and Constructional Engineering, Sapienza University of Rome, Via Eudossiana, 18, Rome, Italy*
[2]*Research Centre for Transport and Logistics, Sapienza University of Rome, Via Eudossiana, 18, Rome, Italy*
✉ *E-mail: gaetano.fusco@uniroma1.it*

Abstract: This study introduces a general methodology to process sparse floating car data, reconstruct the routes followed by the drivers, and cluster them to achieve suitable choice sets of significantly different routes for calibrating behavioural models. This methodology is applied to a large set of floating car data collected in Rome in 2010. Results underlined that routes assigned to different clusters are actually very different to each other. Nevertheless, as expected according to Wardrop's principle, the clusters belonging to the same origin–destination have rather similar average route travel times, even if there is a large range between their minimum and maximum values. A focus on drivers' behaviour highlighted their propensity to follow the same route to their usual destination, though the 12% of the drivers switched to an alternative route. However, the analysis conducted over the 1 month of observations did not reveal the existence of any systematic correlation between neither the change of route nor the change of departure time and the travel time experienced the day before.

1 Introduction

Mobile devices continuously produce a huge amount of data ('Big Data') that can be exploited to improve the knowledge about the state of the transport system and perform appropriate regulation and policy actions [1]. The extent of big data applications in transport is vast and multifaceted. It concerns the vehicle state, the transport system performances, and the users' positions and preferences [2]. With reference to the last issue, floating car data are a ubiquitous source of information of the speed distribution on the road network, which can be used for both statistical analyses and real-time operations [3]. Individual data provide detailed information on the origins and destinations of the trips as well as on intermediate stops in trip chains. They also provide unprecedented observations on the actual route choice mechanisms of the users, which can be used to calibrate and validate the numerous behavioural models that were built in the past years. Specifically, the actual routes followed by the road users were never directly observed at a large scale on the road network but were collected only for small fleets of vehicles such as taxis or small samples of drivers involved in experiments or asked in specific questionnaire-based surveys. It follows that, for example, the famous Wardrop's principle is still a theoretical statement, even if it is soundly based on the rational user theory [4]. Similarly, route choice models are so far often calibrated on small samples of data and have been usually validated on aggregated measures such as link flows instead of on the actual route choice frequencies. Floating car data can be exploited in the several different phases of calibration of multilevel path-based random utility models: identifying suitable path choice sets, specifying a mathematical structure that captures the correlation among different paths, and determining the most likely values of the coefficients of the choice model. When moving from laboratory experiments to real-size floating car data, a huge number of routes is usually observed, which often differ for very small path deviations, so that the choice set may become intractable.

The goal of our research is to provide a general methodology to deal with floating car big data in mobility and apply it to conduct a broad travel pattern analysis on a large urban area. To this goal, we collected since 2010 a set of big data of positions and speeds monitored for insurance reasons. A huge number of observations are available to study spontaneous mobility patterns of users and route choices of about 100,000 drivers. The size of the data set, the high level of detail used in current road graphs and the sparseness of point data, collected every 2 km, gave rise to several problems: the route followed between two consecutive observations has to be estimated; on a highly detailed graph, two routes may differ by negligible quantities, so that some simplifying method should be introduced to get a suitable number of routes that are significantly different among them for the purpose of mobility analysis and modelling. To face these problems, a clustering method is introduced to identify sets of similar routes and select a representative route for each set. This method is applied to a set of trips observed during the month of May 2010 in the metropolitan area of Rome to perform a statistical analysis of travel times on different routes connecting the same origin–destination (O–D) pair and investigate the day-to-day variations of route choice and departure time.

This paper is structured as follows. The next section specifies the goals of the method proposed and refers it to the related works in the literature. Section 3 explains the methodology applied for reconstruction and clustering of the routes followed by the road users. Related results are illustrated in Section 4. Section 5 reports the results of the experimental analysis carried out on the dispersion of route travel times and the day-to-day individuals' travel behaviour. Then, the conclusions follow in Section 6.

2 Related work

The issues dealt with this paper were widely studied in the literature. They can be divided into three main topics: methods of route choice set generation based on route similarity; map-matching and path reconstruction for even low-frequency vehicle sampling; and analysis of route choice drivers' behaviour.

As far as the first issue, various indicators have been proposed to reduce the size of the route choice set that measures the dissimilarity of route alternatives. A necessarily narrow selection is presented as follows. Akgün *et al.* [5] introduced the dissimilarity in terms of length of shared links between two paths. Dell'Olmo *et al.* [6] used the concept of a buffer zone to characterise heterogeneous paths. Martí *et al.* [7] proposed an indicator that overcomes the problems related to the buffer area and takes into account drivers' choice behaviour. The problem of route dissimilarity is closely related to the covariance analysis. Cascetta *et al.* [8] were the first to capture the correlation between route alternatives explicitly by introducing a commonality factor in the deterministic part of the logit model formulation, which is proportional to the overlap of each generic path with the other paths in the choice set. Further contributions are due, among others, to Bekhor *et al.* who adapted a logit kernel model to the route choice problem [9]; Marzano and Papola [10] who developed a link-based path-multilevel logit model; Cascetta and Papola [11] who implemented the strategy of viewing the choice set as a fuzzy set in an implicit model of availability/perception of choice alternatives. In this paper, instead of generating a feasible set of available routes, we follow an experimental approach and we face the complementary problem of selecting a cluster of representative routes that represent the alternative actually perceived by the drivers.

Moreover, the management of large data sets of floating car data gives rise to some computational problems that require data pre-processing [12]. Floating car data are, in fact, collected as successive geographical coordinates and have to be matched on the road network before being applied in transport modelling [13]. Although many map-matching algorithms have been developed in the past years for navigation systems, they are not suitable for modelling analyses. In fact, floating car data are usually collected with lower frequencies than those applied by on-board navigation systems. Thus, the second aforementioned problem arises, which consists of the need to recognise the route followed by the vehicle between two successive sample points, collected even every 1 or 2 min. Rahmani and Koutsopoulos [14] developed a two-step method to individuate a set of candidate links, and therefore find the most likely path in such candidate graph. Frejinger and Bierlaire [15] introduced the concept of sub-network, which tries to capture the most important correlation among similar paths on the network. They assumed the choice set to be composed by all possible paths on the network and developed a method for building the sub-network by applying factor analysis. Bierlaire *et al.* [16] overcome the problem by implementing a probabilistic map-matching algorithm that associates a likelihood value to each of the potentially generated paths. Chen *et al.* [17] proposed a floating car data map-matching algorithm based on a local path searching. The information of the previous matched global positioning system (GPS) point is used to reduce the search space significantly by considering a square confidence area, allowing to determine vehicle moving trajectory.

Li and Xie [18] introduced a bi-level probability method that addresses the problem of missing links due to the low sampling rate of floating cars by building the path between two consecutive points with the shortest path algorithm. Liu and Liu implemented a map-matching method for low-frequency trajectories (e.g. one GPS point for every 1–2 min) [19]. The algorithm takes into account several factors such as the spatial positioning accuracy of GPS points with the topological information of the road network, the consistency of the driving direction of a GPS trajectory etc.

This paper, other than recognising the most likely routes from sparse floating car data, aims more specifically at identifying a limited number of significantly different paths that represent drivers' route choices with the level of accuracy required by traffic models. This problem was recently tackled by Kim and Mahmassani [20] who implemented a trajectory clustering method for the analysis of travel patterns which is unrestricted from map-matching procedures and analyses vehicle trajectory data without using the information of the underlying road network. However, our method aims at building routes on the road graph in order to directly use this information for route choice modelling purposes.

Finally, this paper aims at studying the day-to-day variability of drivers' behaviour in both route choice and departure time, which is the third issue mentioned at the beginning of this section. Regardless the paths reconstruction and clustering, some information has been gained from the raw data about departure time day-to-day variability. This is done by verifying if one individual switches from a cluster to another and check whether this change corresponds to a significant delay experienced the day before. This implies to consider the trip choice mechanism as a dynamic process in which the generic traveller revises his/her previous choice if he/she expects to obtain a benefit from the change. The phenomenon of day-to-day variability has already been addressed by several authors who analysed individuals' behaviour depending on variances in travel time. However, most of these studies were based on travel behaviour analysis of a small sample of drivers [21–24] or laboratory studies based on travel simulators [25, 26].

The contribution of this paper is to introduce a general method for processing, analysing, and clustering sparse Big Data in mobility for studying and modelling route choice drivers' behaviour. By applying the methodology proposed, we get trip data on a road graph and we are able to present some experimental results on route choice behaviour and day-to-day changes of departure time and route choice.

3 Methodology for route choice set identification

The methodology developed determines significantly different paths that represent the actual route choice alternatives for drivers from sparse floating car data. It consists of the following operations:

- *Map-matching* algorithm, which assigns single position points to the arcs of the road network graph they more likely belong to.
- *Route reconstruction*, which explores a reasonable number of feasible routes connecting two successive sampled points and selects the most likely route for each trip in the data set.
- *Path selection*, which analyses the whole set of the reconstructed routes, splits it into several clusters and selects the most representative path for each cluster. Such representative paths compose the final route choice set of alternatives that can be used for behavioural models.

The map-matching operation has already been addressed in a previous work using a semi-probabilistic map-matching algorithm [27]. The latter two issues are described as follows. The aim is to obtain a set of feasible paths that are representative of users' preferences and are significantly different from each other.

3.1 Route reconstruction

Data of vehicle trips are stored in a database. Each trip is described by a sequence of records that depict the instantaneous states of the vehicle and the travelled distance from the origin. Each pair of consecutive records belonging to the same trip forms a segment. For each segment, the k-shortest paths between the sampled points are computed by applying a specific algorithm developed by de La Barra [28]. To select the route that most likely represents the one actually followed by the vehicle, the path that has the minimum difference of length with the observed travelled distance has been chosen within the set of k-shortest paths. For each trip, the whole route followed by the vehicle from the origin O to the destination D is reconstructed as the sequence of most likely paths from consecutive sample points.

It is to note that the processing time is a critical issue in large databases of floating car data. The time for processing a single trip varies with the number of links that compose it; that is, with the length of the trip and the level of detail of the graph. For the k-shortest paths calculation, the value $k = 7$ has been chosen after some experimental results, which showed that larger values of k increased the processing time considerably. For each k, the route reconstruction procedure takes about 75–80 s, and thus for an entire path about 9 min by using an i7-2600K 8-core 3.4 GHz processor with 16 GB of random access memory. This means that having for instance 50 trips and increasing the number of k from 7 to 8 the algorithm would take about 60 min more. Since the increase of k from 7 to 8 did not produce a significant reduction of the error, the former value has been selected.

3.2 Route selection

The route selection procedure takes the results of the path reconstruction routine as inputs; they are the set of routes that are most likely the road users follow in their different trips. Then, the problem is to select a subset of different routes that can be

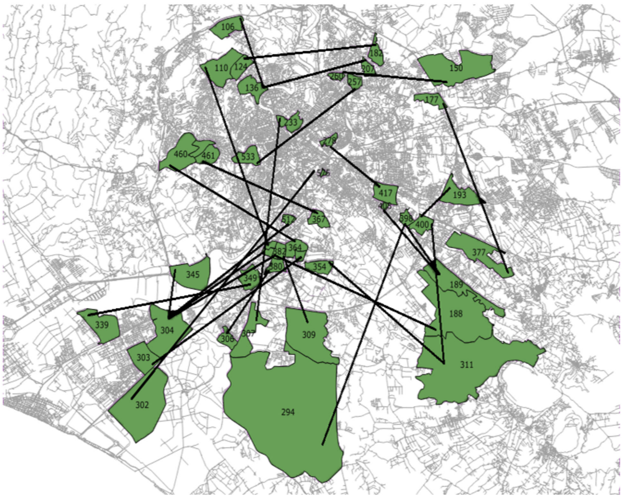

Fig. 1 *Selected O–D pairs for the drivers' behavioural analysis*

perceived by the users as different alternatives among the whole set of routes. This problem is solved by a heuristic algorithm that assigns the routes to different sets and selects the most representative route in each set. The clustering criterion consists of maximising the dissimilarity of paths belonging to different sets and minimising the dissimilarity between paths of the same set. The following dissimilarity index $D(i, j)$ between routes i and j is introduced:

$$D(i, j) = 1 - \frac{1}{2}\left(\frac{L(P_i \cap P_j)}{L(P_i)} + \frac{L(P_i \cap P_j)}{L(P_j)} \right) \qquad (1)$$

where $L(P_i)$ is the length of path i and $L(P_i \cap P_j)$ is the length of the overlapping part of paths i and j. Low index values indicate highly overlapping routes while unit values denote completely distinct routes. More complex indicators that introduce travel time, number and category of links can be introduced. However, they require an extensive knowledge of the traffic speed on all the links of the network in different hours of the day and to take into account weekly and seasonal effects, which are very difficult to achieve. That is the reason why the pure distance-based indicator has been used.

After the final route choice set has been obtained, a representative route is chosen for each set. Such a route should represent drivers' choices and also the most relevant on the graph model; then, a simple rule is applied that maximises a weighted function of users' frequency of choice with the hierarchy of the links travelled.

The route selection procedure applies the following steps.

Step 0 (Initialisation): Get the set $\boldsymbol{P} = \{P_i;\ i = 1, 2, \dots, n\}$ of n reconstructed paths and take it as the initial set of representative paths: $\boldsymbol{S} = \boldsymbol{P}$. Initialise the number of path sets $m = 1$. Let M be the maximum desired number of path choice sets.

Step 1 (Dissimilarity analysis): For each pair of paths P_i and P_j of $\boldsymbol{S}$, identify the road links shared by P_i and P_j, compute their cumulative length $L(P_i \cap P_j)$, and evaluate the dissimilarity between P_i and P_j through the dissimilarity index $D(i, j)$ of (1).

Step 2 (First split): Find the most dissimilar pair of paths in the set $\boldsymbol{S}$, denoted here as h, h'

$$(h, h') = \arg\max_{i,\, j} \{D(i, j)\} i \quad j = 1, 2, \dots, n \qquad (2)$$

Take each of these two paths as the first item of two new distinct sets of paths $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Update the number of sets $m = m + 1$.

Step 3 (Path classification): For each of the paths P_i, $i = 1, \dots, n$, find the set $\boldsymbol{S_l} \in \{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_m\}$ of minimum dissimilarity with P_i

$$l_i = \arg\min_{j_k} \{D(i, j_k)\}$$
$$i = 1, 2, \dots, n \quad j_k = 1, 2, \dots, n_k \quad k = 1, 2, \dots, m; \quad k \neq i \qquad (3)$$

and put P_i in $\boldsymbol{S}_l$. If $m < M$, go to step 4. Otherwise, go to step 5. If P_i is completely distinct from all the identified sets, it is assigned to a temporary cluster S_0.

Step 4 (New set identification): For each set $\boldsymbol{S}_k \in \{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_m\}$, compute the dissimilarity index $D(i_k, j_k)$ between each pair of paths within the same set (internal dissimilarity), find the path P_{q_k} of maximum D within each set k, and select that with the maximum value among all sets, denoted by the index p in the equation below: (see (4)) If the dissimilarity index D_p is greater than a given threshold η, define a new path set $\boldsymbol{S}_\mathrm{p}$ and put P_p in $\boldsymbol{S}_\mathrm{p}$. Update the number of sets $m = m + 1$ and go to step 3. Otherwise, put $m = m$ and the go to step 5.

The threshold value $\eta = 0.5$ has been chosen because it is the median in the definition set of the internal dissimilarity index. If $D_k < 0.5$, there is no need for a further split since the paths of the cluster are classified as similar.

The number of path sets is variable and increases until either a sufficiently low value of internal dissimilarity is reached for all clusters ($\eta < 0.5$) or the maximum number M is reached.

Step 5 (Selection of representative routes): For each set of paths $\boldsymbol{S}_k \in \{\boldsymbol{S}_1, \boldsymbol{S}_2, \dots, \boldsymbol{S}_m\}$, find the path P_{r_k} that maximises the following function:

$$r_k = \arg\max \frac{\sum_{a \in P_k} w_a f_a l_a}{\sum_{a \in P_k} l_a} \quad k = 1, 2, \dots, M \qquad (5)$$

where f_a and l_a are the frequency of choice and the length of the arc a, respectively, and w_a is a weight depending on the hierarchy of the arc a on the network. Select the path as the representative route of the path set $\boldsymbol{S}_k$ end.

4 Application of the route choice set identification methodology

The experimental analysis on route choice has been carried out on a data set of about 100,000 users travelling within the metropolitan area of Rome composed by about 100 million single positions and speed points, collected every 2 km. Each record contains the timestamp, the coordinates of the vehicle, its instantaneous speed, and its state (switched on, moving, and turned off) together with the quality of the signal. Since data are collected every 2 km, to get a sufficient number of position points from the sparse set of data, an initial process of data skimming was applied to select, from all the O–D pairs, the trips done during the morning peak period (7:00–10:00 am), containing at least 30 trips, having a length of at least 6 km, and a travel time of at least 20 min, as well. In this way, a suitable sample of about 600 drivers, performing 1450 trips between 28 O–D zones has been obtained.

The procedures of route reconstruction and route selection have been applied to each trip between the 28 O–D pairs corresponding to the selected criteria (Fig. 1) to obtain the travel times and distances for each route, the cluster it belongs to and the (internal and external) dissimilarity indices.

The procedure of route reconstruction retraces the whole route followed by a vehicle from the origin O to the destination D as the sequence of most likely links from consecutive observed points.

$$D_{q_k} = \arg\max_{i_k,\, j_k} \{D(i_k, j_k)\}$$
$$(i_k, j_k) = 1, 2, \dots, n_k \quad k = 1, 2, \dots, m; \quad k \neq i \quad p = \arg\max_{k} \{D_{q_k}\} \qquad (4)$$

An example of the algorithm result is reported in Fig. 2. The picture on the top highlights the positions of the sparse floating car points.

After the whole set of 1450 routes has been reconstructed, the clustering algorithm has been applied for each O–D pair to form the clusters of similar routes and identify the representative route in each cluster.

For instance, the 131 trips between the O–D zones number 106 and number 136 are grouped into 5 clusters; for each of them, a representative route is selected, as depicted in Fig. 3.

The green and the black routes follow the ring road expressway for almost half of their length, the red one follows a shorter path along a different urban road, whereas the blue and the orange routes follow local roads for about a half of their length. The 50% of trips are assigned to the cluster with the blue representative, the 42% of trips belong to the cluster with the red representative, the 6% of trips are in the cluster of the orange representative, while the clusters of the black and green representative paths have been chosen only once and are composed only by the representatives themselves.

The values of the dissimilarity indices for the O–D pair taken as an example highlight the results of the clustering procedure. The external dissimilarity is the average dissimilarity of the routes of each subset with respect to the routes of other subsets, while the internal dissimilarity is the average dissimilarity of the routes of each subset with respect to the other routes of the same subset. The values related to the representative paths depicted in Fig. 3 are reported in Table 1.

Taking the blue representative route as an example, the average internal dissimilarity of 0.25 means that its cluster is homogenous and that this route stands for a large number of similar alternatives. The average external dissimilarity of 0.90 indicates that the clustering procedure is actually effective since the clusters are significantly different from each other.

The same procedure has been applied to all the 28 selected O–D pairs. The corresponding average results are reported in Table 2.

In the third column of Table 2, the average dissimilarity between the generic path and the representative path of its cluster is reported. A low value means that the representative path of each cluster is actually suitable to characterise all the others assigned to it. The last column reports the average dissimilarity between the cluster representatives. The average value of 0.73 indicates that the choice of the representative paths is actually effective to cluster the trips into considerably distinct groups.

5 Experimental analysis of travel behaviour

5.1 Investigation on route travel time variability

In the previous section, a method has been introduced to cluster routes according to their topology characteristics. However, travel behaviour is mainly related to travel times. Traffic network itself is based on the Wardrop's principle that states that ‘under equilibrium conditions traffic arranges itself in congested networks such that all used routes between an O–D pair have equal and minimum costs, while all unused routes have greater or equal costs’. Now, we want to analyse the statistical distribution of route travel times, their differences, and the effect produced by clustering the routes on their statistical distribution.

This analysis is conducted on a subset of data made during the peak hours (7:00–9:00 am) of working days and repeated at least once over the 1 month observation interval. By applying the clustering method, 685 trips have been selected and arranged into 73 clusters.

In the next section, we tackle the following question: ‘Are the travel times of different used routes almost equal among them?’.

5.1.1 Individual route travel time differences: To test how different are the travel times of the diverse routes connecting the same O–D, the relative travel time difference expressed as the percentage of average travel time on the same O–D is computed for each couple of individual trips within the same O–D pair. This leads to obtaining a new set of $\sum_{i=1}^{N}\binom{n_i}{2}$ individual points, where n_i is the number of individual trips within the ith O–D pair and N is the total number of O–D pairs. The frequency distribution of the relative difference between route travel times is reported in Fig. 4. As expected, the distribution assumes a bell-shaped curve with an average value of −6.8% and standard deviation of 38.6%. The confidence interval of the average value is [−7, 5%; −6.1%] at a 95% confidence level. It is worth noting that 5% of average travel time corresponds to 110 s, which can be considered as a hardly perceptible deviation for trips longer than 6 km, requiring an average travel time of 34 min.

The evidence of a small average travel time difference between the routes chosen by the drivers is in agreement with Wardrop's first principle. On the other hand, the experimental data highlight that some drivers choose twice longer routes than the quickest, as it is expected according to probabilistic behavioural models. Two questions arise: ‘How frequently are worst routes chosen?’ ‘What is the approximation introduced by the zonal aggregation that is usually introduced in transport models?’ ‘Is it comparable with the day-to-day variability of route travel time?’ These questions are the objects of the next two sections.

5.1.2 Route choice frequency distribution: To assess how frequently are worst routes chosen, we assume the average O–D travel time as a reference; then, we compute the frequency distribution of the relative difference between the route travel times and the corresponding quickest O–D value, which is shown in Fig. 5. The results show that 15% of the routes chosen by the drivers have travel times that exceed the minimum value by <20% and that the 50% of the route travel times exceed the minimum by <50%.

On the other hand, if we consider the frequency distribution of the relative difference between the route travel times and the corresponding average O–D values, we observe that the 60% of the routes differ from the average by <20% whereas only the 4% differ from it by more than 50%.

Thus, the variability of route travel times within each O–D and within each cluster of the O–D is worthy of further investigation.

5.1.3 Travel time variability within O–D pairs: To derive statistics about the overall travel time variability within each O–D pair, the average relative standard deviation (that is, the average coefficient of variation) is introduced as the ratio between the travel time standard deviation of all O–D trips and the average O–D travel time. An average value of 23% on all O–D pairs was obtained with a confidence interval of ±12% at a 90% confidence level.

It is worth mentioning that these values include different starting and ending points within the same zone and the variability within different working days. To assess the contribution of the day-to-day variability, we consider the average coefficient of variation of all trips repeated by the same drivers on the same routes. Since it resulted in 20% and the overall variability was 23%, we can deduce that the contribution of the day-to-day variability is prevalent on the space approximation.

5.1.4 Travel time variability for clusters of routes: It is of interest now to evaluate the approximation introduced into the statistical representation of route travel times when introducing clusters of routes. To this goal, we computed the internal variability within clusters, defined as the average coefficient of variation of the travel time over all the routes within each cluster. The average internal variability for all clusters is 15%. Then, we computed the external variability between clusters, defined as the coefficient of variation of the average travel time of each cluster with respect to the average O–D travel time. The value obtained is 17%. These values can be compared with the corresponding value of 23% computed by considering the variability of all trips without clustering.

Thus, the introduction of clusters of routes reduces the travel time dispersion, as expected. Nevertheless, this is not an obvious

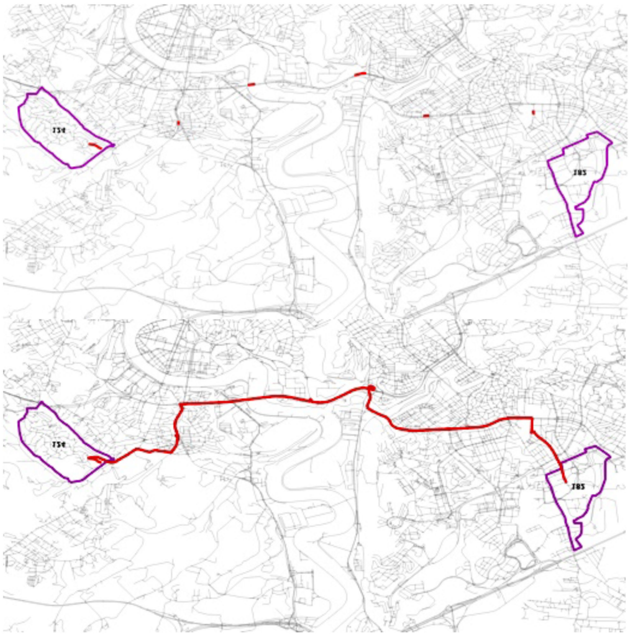

Fig. 2 *Example of route reconstruction*

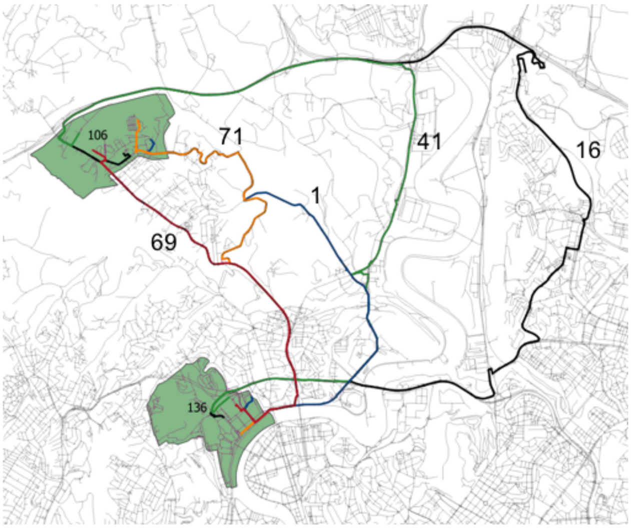

Fig. 3 *Representative paths between O–D zones 106 and 136*

Table 1 Representative routes and dissimilarity Indices of the O–D pair in Fig. 3

Representative ID	Number of choices	Length, km	Average internal dissimilarity	Average external dissimilarity
1 (blue)	65	9.8	0.25	0.90
16 (black)	1	22.9	0	0.93
41 (green)	1	15	0	0.87
69 (red)	56	7.2	0.24	0.91
71 (orange)	8	8.8	0.24	0.7

Table 2 Average values for the Dissimilarity Indices obtained for the 28 O–D pairs

Average internal dissimilarity	Average external dissimilarity	Dissimilarity between generic and representative paths	Dissimilarity among representative paths
0.33	0.80	0.28	0.73

result, because clusters do not aggregate routes according to their travel times but according to their topological characteristics. Hence, introducing clusters simplifies the representation of the observed trip patterns but implies losing a part of their variability, which in our case is around 6%.

A clear picture of the approximation introduced by clustering routes is provided by Fig. 6, which depicts, for each cluster, the maximum, the minimum, and the average value of the route travel times as well as their standard deviation. Some clusters, containing only one route, have been deleted in order to reduce the clutter. Few O–D pairs contain only one cluster of routes. This figure highlights that their standard deviation is rather limited even if the range between their minimum and maximum values is large. The other O–D pairs contain several clusters of routes. As expected according to Wardrop's principle, they have rather similar average travel times: the absolute difference between the average travel

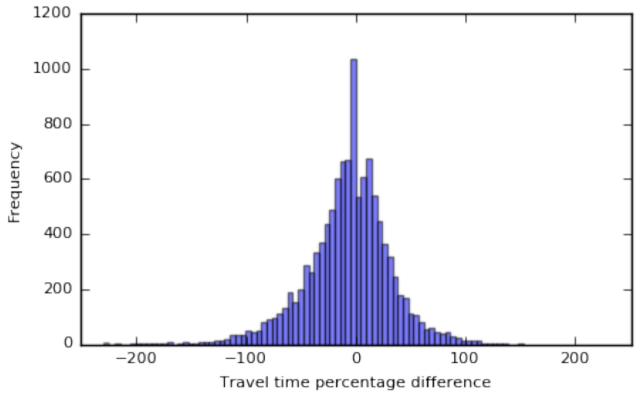

Fig. 4 *Observed frequency of travel time percentage difference computed for pairs of individual trips for all O–D couples*

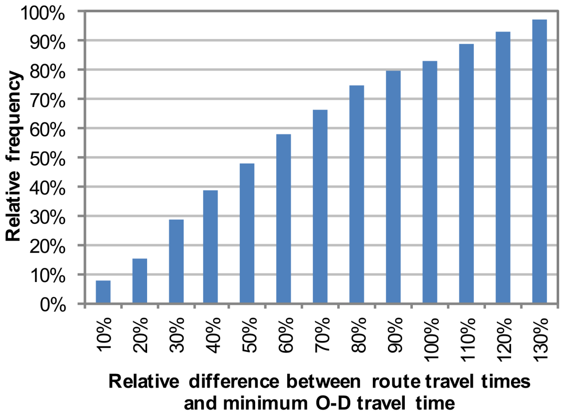

Fig. 5 *Observed frequency distribution of the relative differences between route travel times and the corresponding O–D average travel time*

times of the clusters belonging to the same O–D is on average 6%. However, the standard deviations are even different, because of the different number of observed trips for each route belonging to each cluster, is the average coefficient of variation 17%.

5.1.5 Statistical significance of cluster travel times: Clustering operations reduce travel time dispersion and introduce representative routes instead of many slightly different alternatives. However, because of the time–space dimension of the mobility, from an even large set of trips, some clusters may include few or very few routes. Thus, a question arises about the statistical significance of the results of travel time analysis.

Table 3 reports the results of the statistical analysis conducted on all the available O–D groups. Although the drivers' route choices are distributed on many routes and very few observations can be collected on the least used routes, a statistical significance of the results for 80% confidence level is achieved on 27 out of 53 available O–D cluster groups.

The average route travel time for each cluster is estimated within a confidence interval as 12% at a confidence level of 90%. The absolute travel time difference between the average cluster travel time and the O–D travel time is 11.3%, corresponding to 245 s and 10.6%, corresponding to 174 s, when limiting to statistically significant data. The largest relative travel time difference, which is 46%, has been observed on a cluster consisting of longer routes following the ring road expressway of Rome, while the other three clusters connecting the same O–D pair are passing through the city centre.

5.2 Day-to-day route choice variability

In addition to the statistical analysis of the observed travel times of different routes, we are interested in studying the day-to-day route choice process and in investigating the conditions that produce possible significant changes in route choice: that is, changes between routes belonging to different clusters. The results (Fig. 7)

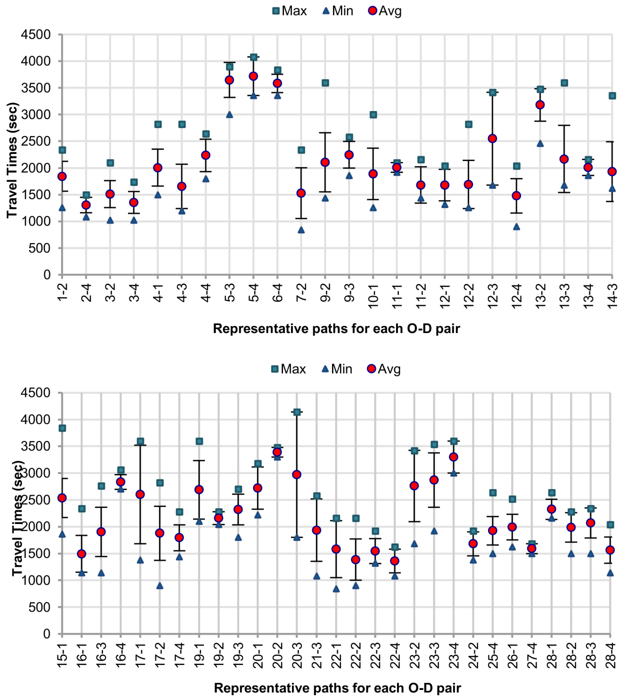

Fig. 6 *Results of maximum, minimum, and average travel time computations for each representative route of all the O–D pairs. The first number denotes the O–D pair and the second number identifies the cluster of routes*

show that the 88% of road users confirm their usual choice and always use a route of the same cluster, whereas the 12% of drivers test other paths during the 1 month observation period. The 10% of the users always choose the same alternative path with respect to the usual one, whereas the 2% choose more than one alternative route. Among them, the 7% of the users make a route change just once; the 3% choose more than several different routes with respect to the usual one.

After the identification of the route choice variability, we want to verify if a correlation exists between the route switching behaviour and possible delays experienced by the drivers the day before the switch while travelling along the usual path.

From the floating car data available, the difference between the average usual path travel time and its value collected the day before the route switch is computed

$$t_{x-1} - \frac{\sum_{i=1}^{x-2} t_i}{N} = \Delta_t \qquad (6)$$

The chart in Fig. 7 sorts the cases in which the path has been changed depending on the individual delay (positive values) or the earlier arrival (negative values) experienced the day before compared with the average travel time.

Some individuals switch route after having experienced a delay, as expected; some others, however, change even after an anticipated arrival the day before. Road users are equally distributed in these two categories. Results refer to only a small sample of route switches (130 cases) that have been observed among repeated trips. However, they do not confirm the expectation that drivers mainly change their previous route if they have experienced a significant delay before. Instead, they highlight the dominance of a very high random component in day-to-day route changing behaviour.

A wider analysis is ongoing to extend the observation period with the aim of identifying meaningful patterns if any, which have not been noted clearly in the cases objects of the present analysis. Indeed, it is likely that some route changes were due to personal reasons, while it is reasonable that drivers change their usual commuting routes after some relevant event or after having experienced systematic delays on their usual routes.

Table 3 Comparison of cluster travel times with average O–D travel times and corresponding statistical tests of 10% accuracy with 80% confidence level, after 1-route clusters have been removed

O–D pair	Cluster	Number of trips	Average cluster travel time, s	Route travel time standard deviation, s	10% Accuracy with 80% confidence level	Average O–D travel time, s	Average route travel time difference, s	Relative route travel time difference, %
1	2	40	1844	284	yes	1844	0	0
2	4	10	1308	152	yes	1308	0	0
3	2	27	1511	256	yes	1453	59	4
	4	16	1354	210	yes	1453	−99	7
4	1	16	2006	357	yes	1870	137	7
	3	17	1655	428	yes	1870	−214	11
	4	4	2235	351	no	1870	365	20
5	3	5	3648	368	yes	3698	−50	1
	4	2	3720	509	no	3698	23	1
6	4	4	3585	198	yes	3585	0	0
7	2	19	1528	488	yes	1528	0	0
9	2	21	2106	566	yes	2349	−244	10
	3	7	2246	270	yes	2349	−104	4
10	1	14	1890	497	yes	3000	0	0
11	1	2	2010	127	yes	2060	−50	2
	2	3	1680	416	no	2060	−380	18
12	1	3	1680	360	no	1742	−62	4
	2	11	1691	474	no	1742	−51	3
	3	2	2550	1230	no	1742	808	46
	4	8	1478	345	no	1742	−265	15
13	2	8	3180	324	yes	2626	554	21
	3	7	2169	682	no	2626	−457	17
	4	2	2010	212	no	2626	−616	23
14	3	12	1930	585	no	1930	0	0
15	1	33	2536	368	yes	2536	0	0
16	1	9	1493	364	no	2120	−627	30
	3	7	1903	494	no	2120	−217	10
	4	4	2835	158	yes	2120	715	34
17	1	3	2600	1126	no	2028	572	28
	2	10	1878	532	no	2028	−150	7
	4	9	1793	257	yes	2028	−235	12
19	1	5	2688	608	no	2436	252	10
	2	2	2160	170	no	2436	−276	11
	3	7	2323	309	yes	2436	−113	5
20	1	3	2720	481	no	3210	−490	15
	2	2	3390	127	yes	3210	180	6
	3	2	3220	1248	no	3210	10	0
21	3	4	1935	671	no	1935	0	0
22	1	6	1580	581	no	1459	121	8
	2	19	1386	395	yes	1459	−73	5
	3	4	1545	270	no	1459	86	6
	4	3	1360	271	no	1459	−99	7
23	2	4	2760	770	no	2894	−134	5
	3	11	2869	533	yes	2894	−25	1
	4	2	3300	424	no	2894	406	14
24	2	3	1680	275	no	1789	0	0
25	4	46	1924	269	yes	1924	0	0
26	1	19	1993	243	yes	2355	0	0
27	4	2	1590	127	no	1560	0	0
28	1	4	2325	216	yes	1811	515	28
	2	9	1987	292	yes	1811	176	10
	3	6	2070	308	yes	1811	260	14
	4	21	1563	248	yes	1811	−248	14
sum		*519*	*114,420*	*21,824*	*27*	*116,014*	*−40*	—
average		*8.58*	*2178.17*	*444.08*	—	*2174.36*	*16.08*	*11.3*
average values significant >80%		*5.50*	*893.72*	*112.42*	—	*856.44*	*47.39*	*10.6*

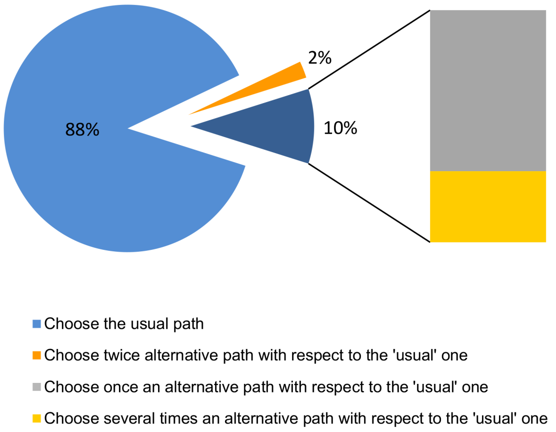

Fig. 7 *Results of route choice variability analysis*

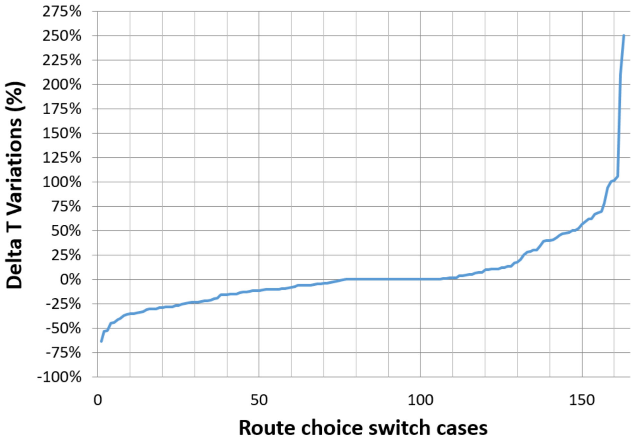

Fig. 8 *Travel time variations in correspondence to route switching cases*

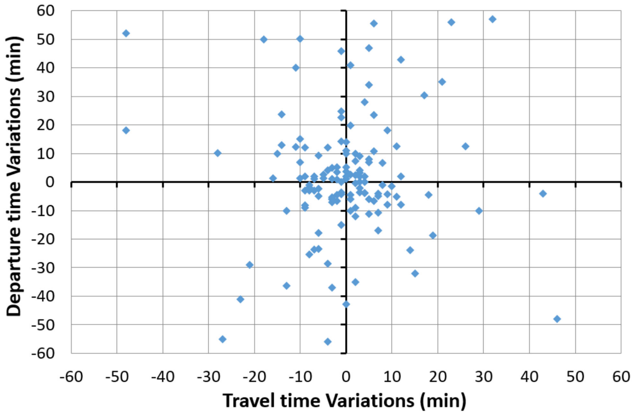

Fig. 9 *Results of departure time variability analysis*

5.3 Day-to-day departure time variability

A similar analysis is conducted to investigate if the choice of departure time can be affected by small variations in travel time experienced the day before.

This analysis requires identifying the most travelled path which has been classified as the 'usual path' for the individual and comparing its travel time with the alternative paths that were chosen by the driver. In Fig. 8, the results are reported.

There are many users who, as expected, anticipate their departure time if they have experienced a significant delay the day before and some who delay it if they arrived too early. There is nevertheless the evidence of many other users who have a counterintuitive behaviour and delay their departure time in correspondence to a delayed arrival the day before or anticipate their departure time if they arrived earlier the day before.

Thus, no evidence confirms the expected behaviour. Data seem to be dominated by the random component that encompasses any other specific reason that leads the users changing their route. Of course, more observations have to be analysed before reaching any conclusion. However, the analysis highlights the evidence of a very large random component that hides the theoretical expectations only based on a strictly rational behaviour (Fig. 9).

6 Conclusions

This paper presented a methodology for route identification from sparse floating car data and apply to analyse the drivers' behaviour in route choice and departure time. Results of a travel pattern analysis of Big Data on urban mobility collected in Rome have been reported.

The routes followed by road users have been reconstructed from a series of sparse positions. Once the routes have been reconstructed, they have been clustered through the path selection algorithm in a limited number of clusters which represent dissimilar alternatives for the individuals. The results of the experimental application have highlighted the effectiveness of the clustering procedure in discriminating significantly different routes and aggregating the similar ones.

The relatively small standard deviation (23%) of the travel times between the same O–D pair and the smaller difference between the average values of the travel times of paths belonging to the different clusters of the same O–D pair (17%) indicate that the clustering procedure simplified the problem modelling and reduces its variability by only 6%.

The travel time distribution of different routes has shown that the 50% exceed the minimum by <50%. However, the 60% of the routes differ from the average by <20% while only the 4% differ from it by more than 50%.

The results of the day-to-day variability confirm the propensity of users to follow their 'usual route' to get to their destination, though the 12% of the users switched from it to an alternative route. However, observations did not reveal the existence of any systematic correlation between neither the change of route nor the change of departure time and the travel time experienced the day before. In fact, the analysis of departure time and route choice behaviour highlighted the predominance of a random component that hides any expected correlation between the choice changes and the travel times experienced the day before. This is probably due to the need of a very long period of analysis that allows observing a sufficient number of users' decision changes, which are expected to be very rare and related to the occurrence of exceptional events. To this goal, further analyses are ongoing on a larger data set of O–D pairs to investigate possible correlation that can be observed over a longer observation period, provided that route changes are unusual events for commuters.

The methodology proposed in this paper is general and can be applied to process floating car data and derive feasible sets of representative routes of the actual drivers' choices. Experimental results of drivers' behaviour are limited to the specific case under study, though the conditions observed are typical of many large European towns.

7 References

[1] Fusco, G., Colombaroni, C., Isaenko, N.: 'Short-term speed predictions exploiting big data on large urban road networks', *Transp. Res. C, Emerg. Technol.*, 2016, **73**, pp. 183–201

[2] De Felice, M., Baiocchi, A., Cuomo, F., *et al.*: 'Traffic monitoring and incident detection through VANETs'. Proc. 11th IEEE Annual Conf. Wireless On-demand Network Systems and Services (WONS), 2014, pp. 122–129

[3] Fusco, G., Colombaroni, C., Isaenko, N.: 'Comparative analysis of implicit models for real-time short-term traffic predictions', *IET Intell. Transp. Syst.*, 2016, **10**, (4), pp. 270–278

[4] Fusco, G., Colombaroni, C., Isaenko, N.: 'Dynamic traveler information systems', in Fusco, G. (Ed.): '*Intelligent transport systems (ITS): past, present and future directions*' (Nova Science, Hauppauge, NY, 2017)

[5] Akgün, V., Erkut, E., Batta, R.: 'On finding dissimilar paths', *Eur. J. Oper. Res.*, 2000, **121**, (2), pp. 232–246

[6] Dell'Olmo, P., Gentili, M., Scozzari, A.: 'On finding dissimilar Pareto-optimal paths', *Eur. J. Oper. Res.*, 2005, **162**, (1), pp. 70–82

[7] Martí, R., González-Velarde, J.L., Duarte, A.: 'Heuristics for the biobjective path dissimilarity problem', *Comput. Oper. Res.*, 2009, **36**, pp. 2905–2912

[8] Cascetta, E., Nuzzolo, A., Russo, F., *et al.*: 'A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks'. Proc. 13th Int. Symp.

Transportation and Traffic Theory, Pergamon, Lyon, France, 1997, pp. 697–711

[9] Bekhor, S., Ben-Akiva, M., Ramming, M.S.: ‘Adaptation of logit kernel to route choice situation’, *Transp. Res. Rec.*, 2002, **1805**, pp. 78–85

[10] Marzano, V., Papola, A.: ‘A link based path-multilevel logit model for route choice which allows implicit path enumeration’. Proc. European Transport Conf., Strasbourg, France, 2004

[11] Cascetta, E., Papola, A.: ‘Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand’, *Transp. Res. C, Emerg. Technol.*, 2001, **9**, (4), pp. 249–263

[12] Fusco, G., Colombaroni, C., Comelli, L., *et al.*: ‘Short-term traffic predictions on large urban traffic networks: applications of network-based machine learning models and dynamic traffic assignment models’. Proc. Fourth IEEE Int. Conf. Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2015, pp. 93–101

[13] Isaenko, N., Colombaroni, C., Fusco, G.: ‘Traffic dynamics estimation by using raw floating car data’. Fifth IEEE Int. Conf. Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017, pp. 704–709

[14] Rahmani, M., Koutsopoulos, H.N.: ‘Path inference from sparse floating car data for urban networks’, *Transp. Res. C, Emerg. Technol.*, 2013, **30**, pp. 41–54

[15] Frejinger, E., Bierlaire, M.: ‘Capturing correlation with subnetworks in route choice models’, *Transp. Res. B, Methodol.*, 2007, **41**, (3), pp. 363–378

[16] Bierlaire, M., Chen, J., Newman, J.: ‘A probabilistic map matching method for smartphone GPS data’, *Transp. Res. C*, 2013, **26**

[17] Chen, F., Shen, M., Tang, Y.: ‘Local path searching based map matching algorithm for floating car data’. Int. Conf. Environmental Science and Information Application Technology (ESIAT 2011), Beijing, China, June 2011, pp. 704–709

[18] Li, J., Xie, L.H., Lai, X.J.: ‘Route reconstruction from floating car data with low sampling rate based on feature matching’, *Res. J. Appl. Sci. Eng. Technol.*, 2013, **6**, (12), pp. 2153–2158

[19] Liu, X., Liu, K., Li, M., *et al.*: ‘A ST-CRF map-matching method for low-frequency floating car data’, *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (5)

[20] Kim, J., Mahmassani, H.S.: ‘Spatial and temporal characterization of travel patterns in a traffic patterns in a traffic network using vehicle trajectories’, *Transp. Res. Procedia*, 2015, **9**

[21] Meneguzzer, C., Olivieri, A.: ‘Day-to-day traffic dynamics: laboratory-like experiment on route choice and route switching in a simple network with limited feedback information’, *Soc. Behav. Sci.*, 2013, **87**, pp. 44–59

[22] Selten, R., Chmura, T., Pitz, T., *et al.*: ‘Commuters route choice behavior’, *Games Econ. Behav.*, 2007, **58**, (2), pp. 394–406

[23] Avineri, E., Prashker, J.N.: ‘Sensitivity to travel time variability: travelers’ learning perspective’, *Transp. Res. C, Emerg. Technol.*, 2005, **13**, (2), pp. 157–183

[24] Vacca, A., Meloni, I.: ‘Understanding route switch behavior: an analysis using GPS based data’, *Transp. Res. Procedia*, 2015, **5**, pp. 56–65

[25] Mahmassani, H.S., Liu, Y.: ‘Dynamics of commuting decision behavior under advanced traveler information systems’, *Transp. Res. C, Emerg. Technol.*, 1999, **7**, pp. 91–97

[26] Tawfik, A.M., Rakha, H.A., Miller, S.D.: ‘An experimental exploration of route choice: identifying drivers choices and choice patterns and capturing network evolution’. 13th Int. IEEE Conf. Intelligent Transportation Systems (ITSC), Funchal, Portugal, September 2010

[27] Rambaldi, S., Marchioni, M., Bazzani, A., *et al.*: ‘Traffic global analysis on the whole Italian road network’. IEEE MIPRO, Proc. 35th Int. Convention, Opatija, Croatia, May 2012, pp. 1678–1682

[28] de la Barra, T., Perez, B., Anez, J.: ‘Multidimensional path search and assignment’. 21st PTRC Summer Annual Meeting, University of Manchester, UK, 1993