2015 Models and Technologies for Intelligent Transportation Systems (MT-ITS)
3-5. June 2015. Budapest, Hungary

Assessing the consistency between observed and modelled route choices through GPS data

Selini Natalia Hadjidimitriou, Mauro Dell'Amico
Department of Sciences and Methods for Engineering
University of Modena and Reggio Emilia
Italy
selini@unimore.it,
dellamico@unimore.it

Guido Cantelmo, Francesco Viti
Faculty of Science, Technology and Communication
University of Luxembourg
Luxembourg
guido.cantelmo@uni.lu,
francesco.viti@uni.lu

***Abstract*— In traffic engineering, different assumptions on user behaviour are adopted in order to model the traffic flow propagation on the transport network. This paper deals with the classical hypothesis that drivers use the shortest possible path for their trip, pointing out the error related to using such approximation in practice, in particular in the context of dynamic origin-destination (OD) matrix estimation. If this problem is already well known in the literature, only few works are available, which provide quantitative and empirical analysis of the discrepancy between observed and modelled route sets and choices. This is mainly related to the complexity of collecting suitable data: to analyse route choice in a systematic way, it is necessary to have observations for a large period of time, since observing trajectories for the single user on a specific day could not be enough. Information is required for several days in order to analyse the repetitiveness and understand which elements influence this choice. In this work the use of the real shortest path for a congested network is evaluated, showing the differences between what we model and what users do. Results show that there is a systematic difference between the best possible choice and the actual choice, and that users clearly consider route travel time reliability in their choice process.**

Keywords - route choice; shortest path; travel time; GPS trajectories; hierarchical cluster analysis; reliability

I. INTRODUCTION

Traffic assignment models are coarse representations of the real rational behaviour of drivers. The limits of using these approximations are already known in the literature (e.g. [1]-[2]). Several studies on route choice within traffic assignment problems deal with the development of algorithms for the generation of the choice set. Given an origin and a destination, the problem consists in the identification of the sub-set of alternatives that are the most likely to be considered by the drivers. These may vary during the day and across days, depending on habits, past experiences, information acquired before and during the trip, etc. Furthermore the alternatives to be identified by the algorithm have to be heterogeneous in order to properly represent the variety of the available choices. The goodness of the solution is usually evaluated using the overlap percentage between the alternatives generated by the algorithm and the observed paths. In [3] an overview of different algorithms for the identification of the best alternatives (i.e. in terms of cost, distance, time, etc.) is presented. Two main categories of route choice sets algorithms are identified: the link elimination [4], and the k-best paths [5]. The first approach consists in the identification of the optimal paths, from which one or more links are eliminated. For instance, when a driver wants to avoid a specific road link, this is eliminated and the new alternative is computed. The second approach consists in the identification of the k-best paths. The disadvantage of this second method is that it is not very efficient and it tends to create similar alternatives. They also propose a constrained k-path algorithm to find the set of alternative routes that are most likely selected by the drivers, through identifying the number of alternatives, which are not overlapping or that are not circular. The method proposed by [6] consists in the measurement of the spatial dissimilarity of alternatives after the k-shortest paths are computed. The dissimilarity of the paths is obtained by randomly selecting the paths from the set such that the dissimilarity, in terms of overlapping, is maximized.

Apart from the dissimilarity between (mostly unobservable) actual choice sets and the ones used in modelling the route choice, actual choices are often differing from the optimal route choices indicated by the models. Among the different works reported in the literature, [7] found that the 37% of respondents selected the shortest time paths (90% of overlapping was required). Similarly, [8] found that the travellers who selected the shortest time path were 43.3%. An empirical analysis of route choice behaviour is presented in [9], confirming the systematic difference between shortest and selected routes. The authors studied habitual driver behaviour using GPS coordinates registered during three weeks. Analyses from this database highlight how results from the deterministic route choice models do not match with the observed paths. A

complementary view is given in [10], where the authors reported that one of the factors determining the systematic difference between best choices and observed route choices is travel time variability, in combination with information reliability [11]. Another study on the perception of travel time [12], consistently with the previous analyses, found that 41% of drivers minimize time, while 80% of drivers minimize distance. They realized also that having a more direct connection or a faster route influences the perception of travel times. Finally, [13] point out that the structure of the network influences the perceived travel time too. They used a dataset of GPS coordinates and a survey to find significant differences in travel time perception based on the characteristics of the road network. Results presented in this work differ from the ones reported in the literature, since only the 26% of the paths selected by the drivers overlaps the shortest time alternative (at least 80% of overlapping), as discussed in the section that describes the results of the analysis.

The above results reported in the literature motivated and are used as starting point in the work presented in this paper. However, this work differentiates from the others in the use of the actual mean speeds to compute the path travel times, which is different within each time interval and for each link of the road network. In the literature, the shortest path evaluation is usually based on the maximum speed for each road link. In this work, shortest time alternatives for each time interval are evaluated based on the observed data. Secondly, since the GPS dataset is collected during a year and half (September 2010 – January 2012), real observations of the habitual driver behaviour over a significant amount of time are available. GPS coordinates are used as raw data by matching each of them to the road network. Our analyses are carried out in the morning peak period, in which congestion is observed. This allows to evaluate not only the difference between the *observed* and the *shortest* path, but also with the *instantaneous shortest path*, which is the best possible route when congestion is observed, according to the observed speeds.

The contribution of the paper is firstly to point out how drivers do not select the shortest time path, so that the probability of selecting the shortest path alternative is much lower. Then we focus on the impact of the reliability of the routes in partially explaining the systematic difference between the optimal paths and selected routes. Finally, the authors focus on the analysis of few users to analyse if the real route choice set is matching with an artificial one.

The organization of this paper is as follows. Section II describes the dataset, which is deployed for the comparison of modelled and observed alternatives, and the methodology to compute the shortest time alternatives. Section III presents the description of the methodology used for the paths comparison. The last Section IV presents the analysis of the results, where the main differences between the *observed* and the *shortest time* alternatives are showed. Finally, section V outlines the conclusions of the work and put the basis for future analysis on the consistency of traffic assignment models.

II. DESCRIPTION OF THE DATASET

Low-frequency GPS coordinates

The dataset used in this study consists of low-frequency GPS coordinates and refers to paths performed by 89 drivers in the Province of Reggio Emilia, Italy, during the period 1st September 2010 – 31st January 2012. Data were collected using a data logger installed into the vehicle. The database includes more than 57.000 observed paths distributed rather uniformly in the Provincial territory. Data were collected in the context of the European Community FP7 project TeleFOT, which aimed at testing the impact of in-vehicle and nomadic devices on usability, behaviour, incidents, safety, green driving, efficiency and the impact on the transport system. Systematic trips of individuals are detected using a hierarchical clustering approach, which run over the origin and destination pairs. The result consists of 119 clusters of repetitive trips for a total of 13.766 paths. Each cluster includes a set of paths made by the same driver during the observation period. The dataset allows to have an overview of the systematic choices made by the users in terms of route choice, travel time, day identification and departure time.

Clusters of repetitive travels in terms of similar origin and destination are generated using the single linkage method [14]. Similarities between origin and destination pairs are summarized in a matrix of distances measured through the Euclidean metric. Initially, each observation forms a cluster. Successively, step by step, the nearest observation pairs are merged into a new cluster. The *cophenetic* correlation coefficient, which is an indicator of the goodness of the clusters structure [15] and varies between -1 and 1, shows values that are above 0.7. An indication of the significance of cophenetic correlation coefficients has been provided by [16]. He found that values equal or greater than 0.8 indicated a good fit. Thus clusters with value of the coefficient included between 0.7 and 0.8 have been analysed in detail and outliers have been eliminated from the cluster.

Shortest time alternatives

Time-based shortest paths between each observed origin and destination are computed using an A* algorithm [17] that has a better performance in terms of time with respect to the Dijkstra [18]. The algorithm exploits the average velocities, using the observed data, on the road links to compute the shortest path in terms of time. For each of the 13 thousand paths, the length and the travel time have been measured. Furthermore, a GIS (Geographical Information System) shape file for each shortest path has been created to allow the visualization of the path with the GIS. One of the main advantage of the GIS is the possibility to compare observed and shortest time alternatives in terms of route choice based on the visualization of the spatial characteristics of the paths.

A map matching over the road network would, therefore, allow to identify the selected paths [19]. In this work, GPS coordinates and the geographical representation of the road network are used to compare modelled and observed alternatives.

III. METHODOLOGY

This section outlines the methodology proposed to compare the observed alternatives to the shortest time paths and describes how the modelled paths are selected.

The comparison between the observed and shortest time alternatives has been performed using a spatial query that was run using PostgreSQL 9.0 with the PostGIS 2.0 extension, which allows to perform queries on tables containing geometry information. All the shape files of the shortest time alternatives and the coordinates characterizing the observed routes have been imported in the database. The query counts the number of GPS coordinates (for each observed path) that intersect the corresponding shortest time alternative path, represented as a polyline. The procedure to compare the two datasets consists in counting the number of points, for the observed path – which are included within few meters from the shortest time alternative (50 meters). The first step of the analysis consists in the comparison between observed paths and shortest routes, obtained exploiting speed measurements.

Further, a part of the network has been implemented on one of the most popular planning tools, PTV Visum [20]. The goal was to compare results from the Route Choice Model (RCM) within Visum, with the observed alternatives. In order to quantify the contribution of the RCM with respect to the Traffic Assignment (TA), a synthetic demand has been assigned on the network to reproduce the proper level of congestion, using a Deterministic User Equilibrium (DUE) assignment. Then, keeping constant the demand, few vehicles have been loaded using a Stochastic Assignment (SA), in order to generate the set of alternatives insulating the RCM model.

Since the average velocities were available for all time ranges and most of the observed alternatives were performed during the morning peak, the calibration phase focused on the 8-9 AM time period. Information about speeds and observed links flows have been used, aiming at simulating the congestion and reproducing realistic link flows. As first step, artificial OD flows have been loaded on the network in order to reproduce reasonable link flows, consistent with the observed link flows. This configuration has been used as starting point for the calibration phase, which has been performed using measured speeds, derived by GPS coordinates, and using observed link flows to verify the solution. The output is a calibrated network, in which the error between modelled and observed speeds is less than 6% (according to the RMSE metric). During the calibration phase, a "Deterministic User Equilibrium" (DUE) approach has been used. To evaluate the modelled route choice subset, a SA model has been used. SA models "assume that traffic participants in principle select the best route, but evaluate the individual routes differently due to incomplete and different information" [20]. In this part we describe the stochastic model used to obtain our results, according to Visum. While according to the classic DUE, users instantaneously switch on the shortest path with a homogeneous behaviour, in the stochastic assignment the shortest path is not uniquely defined for all the demand. The demand for each route is thus distributed according to a distribution model which establishes the share of demand. In the experiments, we used the Kirchhoff distribution model to model the route choice and compare it to the observations. Anyway, in this study we do not focus on the amount of vehicles assigned to each route, but only on the route set generation aspects.

The output of the model consists of 171 paths for all the three OD pairs analysed in this paper. The set of modelled alternatives have been compared to their corresponding observed paths in terms of length, travel time and overlapping percentage. The modelled alternatives have also been loaded into the GIS using the node coordinates for the selected links.

The last part of our analysis focuses on the reliability index, which has been computed for each path. The indicator used is the one proposed by [21], which provides a measure of how early/late a driver arrives at destination by providing two measures: the lateness and the earliness reliability indexes. The lateness reliability index is computed as follows by using the average and variance velocities of the road links selected by the driver:

$$r(l) = exp[1/2 * T_{log}(l) - z_{\alpha/2} * \sqrt{T_{log}(l)}] \quad (1)$$

Where $T_{log}(l)$ is a dimensionless variation logarithm, computed using the day-to-day average and variance travel time. The z-score is set to 1.645. The main characteristic of the index is that it does not depend on the length of the link. Therefore if a link is divided into two without changing its physical characteristics, the measure of reliability remains unaffected. In this work, only the lateness reliability index is considered to evaluate observed and modelled route choices because it should be more important for the driver how late the alternative route allows to arrive at destination. The objective of this analysis is to explore the possibility to consider into the route choice model the reliability explicitly as parameter, evaluating the influence of this element on the user route choice.

IV. RESULTS

Results allow to make considerations on the preferences of habitual drivers, when they make decisions on alternative routes during the peak hour, for a specific OD pair.

Fig. 1 and 2 show the results of the first analysis, which takes into account the entire dataset of clustered paths comparing them to their shortest time alternative. The dots indicate the low frequency GPS coordinates, while the line represents the shortest time path. The figures show the selection made by the same driver during two different days and between the same origin and destination pair. The corresponding shortest path, in terms of travel time, has been computed using the hourly average velocities, therefore the timestamp of the origin zone for the observed path is used as reference point by the algorithm when velocity hourly time range is selected. In the two examples, however, the shortest time alternatives do not differ from each other. This is often the case since drivers tend to travel between the same OD during

similar hours of the day. The same analysis has been done for all the observed paths inside the study area.

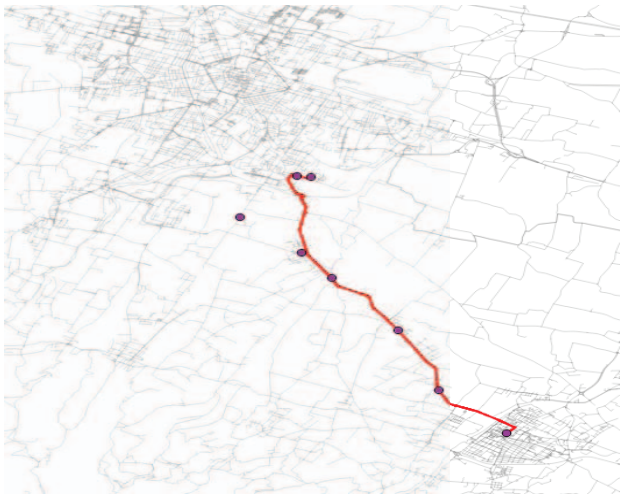

Fig. 1. Overlap 75%

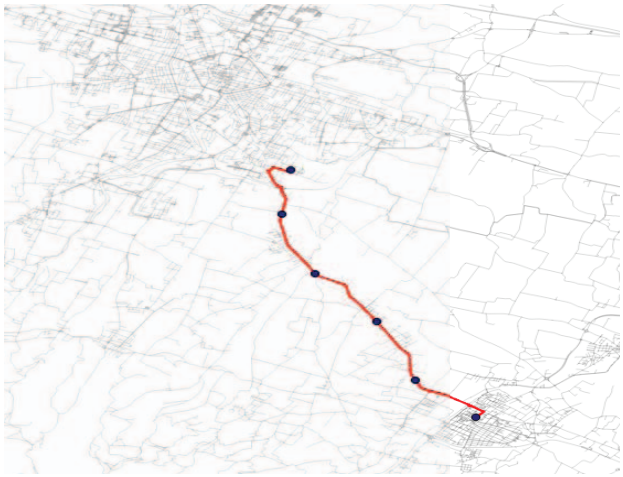

Fig. 2. Overlap 100%

Table 1 shows results for different overlapping percentages for all the 13.766 paths. The values indicate that only ¼ of the alternatives coincide with the shortest path for more than the 80% of its links. This result is considerably different from the values reported in the literature, in which it was reported that about 40% of the users select paths that overlap the shortest route in terms of travel time for more than 90% of the times. It is relevant to point out that, to the best of our knowledge, this is the first work in which the comparison between shortest time and observed alternatives is performed within a congested network, and the real shortest path time is computed using observed speeds.

Results of Table 1 highlight that drivers use different routes with respect to the shortest path.

Table 1 Overlaps percentage of paths

Overlaps percentage of paths	
Overlap	***%***
100%	15.07%
90-99%	1.46%
80-89%	9.62%
70-79%	9.57%
60-69%	11.10%
50-59%	4.89%
40-49%	12.17%
30-39%	13.63%
20-29%	11.11%
10-19%	4.52%
0-9%	0.03%

To gain insight into the systematic difference between the observed and optimal routes in terms of travel time, we investigate two elements: 1) for each user, the average travel time has been computed and normalized with respect to the shortest path travel time, in order to evaluate how much the real path is longer than the shortest on average (Fig. 3). Then, 2) the same operation has been done to analyse the delay per km travelled, obtained as the ratio between travel time and route length (Fig. 4).

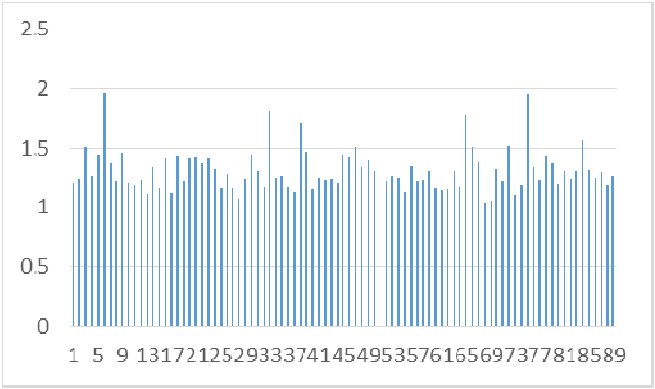

Fig. 3. Normalized Average Travel Time for each User

Figure 3 shows that drivers have the tendency to select routes with a higher travel time - on average 1.3 times longer - than the shortest time path. In Figure 4, it is possible to see that, in some cases, users choose routes with a lower delay/km with respect to the shortest path. Specifically, the lowest observed delay is 0.73, where 1 represents the delay for the shortest route. If this element shows that users can take longer routes to avoid congestion, this cannot be considered a general rule, since the average value of the delay among all the users and the observations is 1.15. This element points out that, on

average, people use routes which present a higher delay/km with respect to the shortest path.

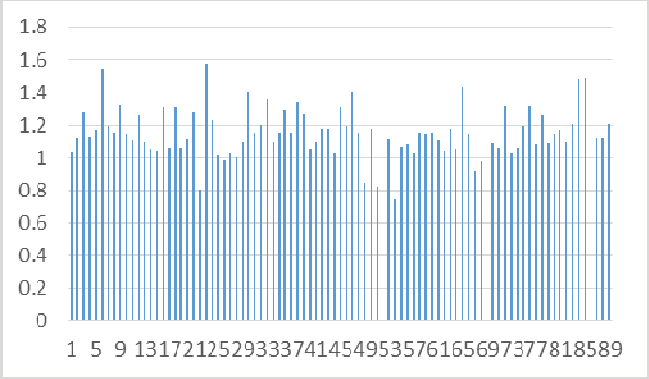

Fig. 4. Normalized average delay/km for each User

In an attempt to identify the factors determining this systematic error in identifying the best alternative route by the drivers, in the next step, the correlation between route choice and reliability of the path is investigated using the reliability indicator (1). The hypothesis is that reliability plays a relevant role in the route choice, as previously reported in [10] but using a Stated Preference survey.

To perform this analysis, we evaluate the reliability of all routes in the database according to equation (1). For each user, the most reliable route has been identified, computing the percentage of using this alternative for each user. For illustration's sake, these percentages have been sorted in decreasing order, i.e. from the driver who selects always the most reliable route until the least performing user in terms of finding the most reliable route, and results are shows in Fig.5.

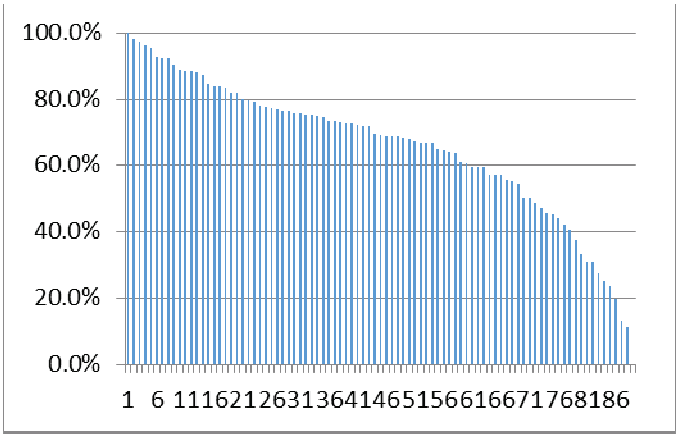

Fig. 5. Normalized average delay/km for each User

Analysing the trend displayed in figure 5, we can observe that more than 80% of the users were able to select the most reliable route for more than 50% of the times. This highlights that there is a systematic influence of the reliability in the way users choose a route.

The final step of this work consists in the comparison between the observed and modelled route choices. In this section and due to the time-consuming operations needed to elaborate the large database, we analyse only some of the observations on the entire database and compare them with the modelled information. Hence, we focus here on the observed route choice behaviour of three users and compare them to the modelled route choices using the Visum calibrated model.

Fig. 6 shows a subset of the network, which has been used to compare the observed routes with the modelled ones.

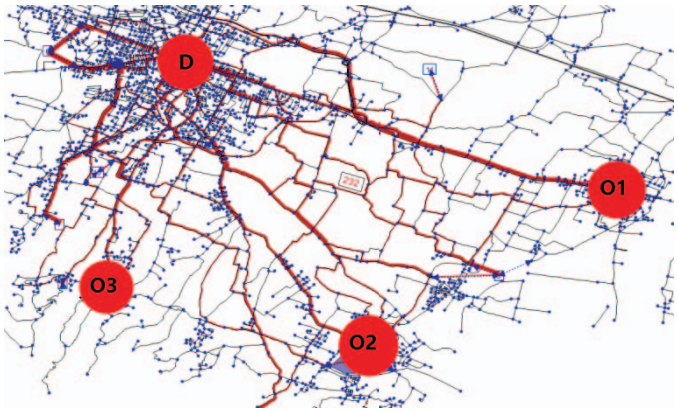

Fig. 6. Test Network

320 observed paths are used in this phase. For each origin, it was possible to observe the day-to-day behaviour of the users. Once the routes have been generated in Visum, the travel time has been obtained using the average speeds, in order to have a proper comparison with the observed routes. Fig. 7 shows, for each user, the difference between average travel times for observed and modelled routes. Specifically, the average travel time has been normalized with respect to the shortest path travel time, as for the results in Fig. 3.

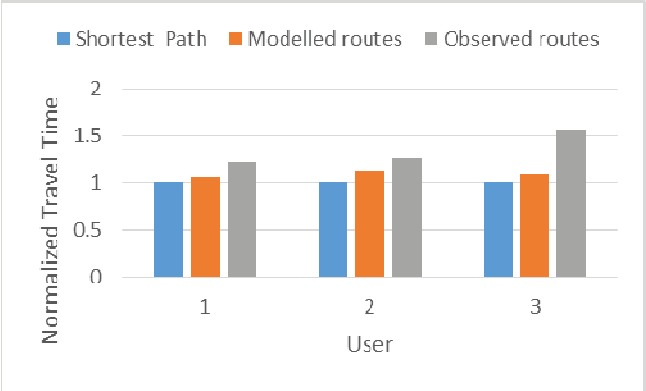

Fig. 7. Comparison between shortest time, modelled and observed routes

As expected, modelled alternatives do not overlap the observed ones, but on the other hand the travel time is similar to the shortest time path.

With reference to user 3, the difference of travel time between the observed and modelled alternatives is much higher comparing to the others. The reason could be that all the observed alternatives are much longer than any of the modelled ones: the user decided to drive around the city centre instead of selecting the straightest routes. Consequently, the travel time is also higher. Fig. 8 shows the three observed alternatives and the corresponding shortest time path (continuous line).

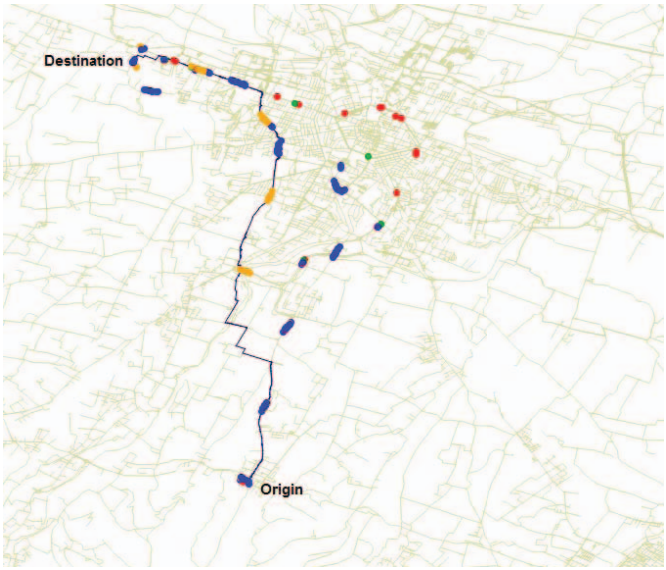

Fig. 8. OD 3 – observed alternatives and shortest time path

Fig. 9 shows three observed alternatives, where all the alternatives have a common section. While in Fig. 8 we observed that drivers may take into account longer alternatives in order to avoid the congestion, here three paths, performed in a very similar time range (10-10.40 AM) during different days, are selected to analyse more in detail the relation between drivers' route choices and the reliability of a path.

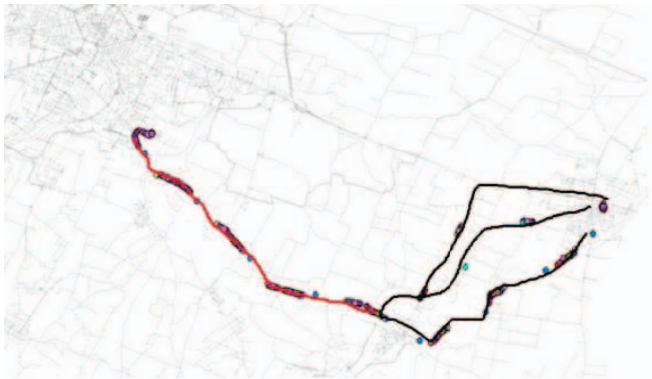

Fig. 9. OD1 observed alternatives

As one can easily note, the observed paths differ from each other only in the last section. More specifically, the first section (Part 1) has three different alternatives that have been selected by the driver in different days of the period under analysis; the last section that arrives at destination (Part 2) is in common for all the alternatives.

Table 2 Reliability of the three alternatives

Part 1	Reliability
1_1	59.36%
1_2	60.42%
1_3	61.69%

Table 2 reports the reliability of each of the three alternatives (part 1). The common part (part 2) has a considerably higher reliability value of 73.72%. Although these results are not generalizable, they indicate that the reliability of the alternative routes is, in all three cases, lower comparing to the common section. Moreover the higher value of the reliability (61.69%) corresponds to the most frequent observed alternative. Thus, the paths that are closer to the urban centre seem to be less reliable while the driver prefers alternatives that are in general more reliable. A relevant observation after this preliminary analysis, worth further elaboration that will be presented in future works, is that a possible additional rule to help identifying route alternatives in the choice set generation model, is to look for route segments which have a higher values of the reliability.

V. CONCLUSIONS

This paper provided empirical tests on the assumption of using the shortest path in terms of travel time as the most likely choice for road traveller in the context of repetitive, habitual, trips. In addition, we also extend the analysis by investigating the validity of using shortest time routes to generate the considered route set.

In the first part, results are obtained analysing a significant number of different paths, for different OD pairs. The analysis, performed using more than 14000 repetitive trips, shows that only a small part of the users (about 25%) select the shortest path for their trip.

By comparing the shortest and observed paths, it was possible to argue that in reality drivers spend systematically more time to reach the destination This element has been quantified in a tendency to observe 30% additional travel times. This is partly due to perception errors, which can partly be modelled using a stochastic route choice modelling approach.

However, drivers tend to use routes that strongly differ with respect to the "best" ones, while models tend to identify all possible alternatives starting from the shortest paths. Hence, while drivers have the tendency to choose totally independent alternatives, the modelled ones are similar and strongly correlated.

Another relevant contribution in this work is the adoption of a reliability index, which was calculated based on the day-to-day travel time variance. Looking at the results presented in the previous section, users have the tendency to accept some more delay in their travel time if the route presents a higher reliability. Results clearly show that the greatest share of observed users tends to choose the most reliable path.

Finally, results show preliminary analysis of how to use the reliability index to identify route alternatives which are more representing the observed routes. It was found that routes segments overlapping are characterised by a significantly higher value of the reliability index.

Results from this work are very relevant in understanding the limits of the actual route-choice within well-established assignment models. More elements that may have an influence on the route choices should be investigated: among others, the overlapping between round trip routes and symmetric ODs, the repetitively in the day to day route choice, or if route choices are influenced by trip chains, or if there are topological elements that influence decisions.

ACKNOWLEDGMENT

The authors would like to acknowledge networking support by the COST Action TU 1004 Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems.

REFERENCES

[1] Abdel-Aty, M., R. Kitamura, and P. Jovanis (1997). Using stated preference data for studying the effect of advanced traffic information on drivers' route choice. Transportation Research Part C 5 (1), 39-50.

[2] Jan, O., A. Horowitz, and Z. Peng (2000). Using global positioning system data to understand variations in path choice. Transportation Research Record: Journal of the Transportation Research Board 1725 (-1), 37-44.

[3] Scott, K., Pabon Jimenez, G., Bernstein, D. (1997). Finding alternatives to the best path. Presented at the 76th Annual Meeting of the Transportation Research Board, Washington, DC.

[4] Azevedo, J., et al., An algorithm for the ranking of shortest paths. European Journal of Operational Research, 1993. 69 (1): p. 97-106.

[5] Van der Zijpp, N.J.and Catalano, F. S. (2005). Path enumeration by finding the constrained k-shortest paths. Transportation Research Part B, 39, pp. 545–563

[6] Akgun, A., Erkut, E., Batta, R. (2000). On finding dissimilar paths. European Journal of Operation Research 121, 232–246.

[7] Bekhor, S., M. Ben-Akiva, and M. Ramming (2006). Evaluation of choice set generation algorithms for route choice models. Annals of Operations Research 144 (1), 235-247.

[8] Prato, C. and S. Bekhor (2006). Applying Branch-and-Bound Technique to Route Choice Set Generation. Transportation Research Record 1985, 19-28.

[9] Zhu, S. and Levinson, D. (2012). Do people use the shortest path? An empirical test of Wardrop's first principle. 91st annual meeting of the Transportation Research Board. Washington D.C.

[10] Viti F., Bogers E.A.I., Hoogendoorn S.P. (2005). Day-to-day learning under uncertainty and with information provision: model and data analysis. Presented at the 16th International Symposium of Transportation and Traffic Theory (ISTTT16), Maryland, US.

[11] Bifulco G., Di Pace R., Viti F. (2013). Evaluating the effects of information reliability on travellers' route choice. European Transport Research Review (2014) 6:61–70.

[12] Vreeswijk, J. D., Thomas, T., Berkum, E. C. van & Arem, B. van (2013). Drivers' perception of route alternatives as indicator for the indifference band. In TRB (Ed.), Proceedings of 92th annual meeting of the Transportation Board, Washington D.C.

[13] Parthasarathy, R., Levinson, D., and Hochmair, H. (2013). Network structure and travel time perception, 53 presented at the 92nd annual meeting of the Transportation Research Board, Washington D.C.

[14] Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (British Computer Society) 16 (1): 30–34.

[15] Jain, A. K. and Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall, Inc.

[16] Rohlf, F. J. (1998). NTSYSpc: Numerical Taxonomy and Multivariate Analysis System. Version 2.02. Exeter Software, Setauket, New York.

[17] Hart, P. E.; Nilsson, N. J.; Raphael, B. (1968). "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". IEEE Transactions on Systems Science and Cybernetics SSC4 4 (2): 100–107.

[18] Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". Numerische Mathematik 1: 269–271.

[19] Miwa, T., Kiuchi, D., Yamamoto, T., Morikawa, T. (2012). Development of map matching algorithm for low frequency probe data, Transportation Research Part C: Emerging Technologies, 22: 132-145.

[20] PTV, A., 2012. VISUM 12. 5 User Manual.

[21] Kaparias, I., Bell, M.G.H., Belzner, H. (2008) A new measure of travel time reliability for in-vehicle navigation systems, Journal of Intelligent Transportation Systems,12,4,202-211,Taylor & Francis.