Extracting activity-travel diaries from GPS data: towards integrated semi-automatic imputation

Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 22 (2014) 178 – 185

Procedia Environmental Sciences

12th International Conference on Design and Decision Support Systems in Architecture and Urban Planning, DDSS 2014

Extracting activity-travel diaries from GPS data: Towards integrated semi-automatic imputation

Tao Feng*, Harry J.P. Timmermans

Urban Planning Group, Department of the Built Environment, Eindhoven University of Technology, Vertigo 8.16, 5600MB, Eindhoven, The Netherlands

Abstract

This paper presents an integrated approach to extracting activity-travel diaries from GPS data. The imputation involves a semi-automatic procedure of transportation mode and activity type recognition, and applies full and partial consistency principles to different trip episodes of a tour. Complementing earlier work on the evaluation of this approach at the epoch level, this paper investigates the performance of the integrated imputation at the episode level. The originally imputed data were used as reference to compare the superimposed data against validated data. Results indicate that the distribution of transportation modes and activity types are similar for these data sets. The new algorithm imputes diary data that are closer to the validated data that the results of the original algorithm.

Peer-review under responsibility of the Eindhoven University of Technology, Faculty of the Built Environment, Urban Planning Group

Keywords: Activity-travel diary, Bayesian Belief Network, imputation

1. Introduction

GPS data have been increasingly used in various disciplines, including transportation, geography, computer science, health care, and environmental sciences. The spatial and temporal information encoded in GPS data

* Corresponding author. Tel.: +31-40-247-2301; fax: +31-40-243-8488.
E-mail address: t.feng@tue.nl

Peer-review under responsibility of the Eindhoven University of Technology, Faculty of the Built Environment, Urban Planning Group
doi:10.1016/j.proenv.2014.11.018

potentially allows investigating activities and movement at the microscopic and macroscopic level. For transportation modelers, one of the main reasons for using GPS data is to extract activity-travel diary data. The diary data embed plenty of information on activity types, travel distance, time, duration and route, which are essential for various branches of research in travel demand modeling.

However, extracting such quantitative information from GPS traces is not straightforward. Basically, GPS traces data are recorded at the epoch level (measured in seconds), while the required activity-travel diaries are defined at the trip (episode) level. Thus, epoch-level GPS traces need to be transformed into segmented activities and trips. The transformation procedure generally involves multiple procedures, which deal with separate problems, including pattern recognition of transportation mode/activity type, segmentation of activities and trips, geocoding of activity locations, map matching, etc. Consequently, a comprehensive, integrated, method is needed to seamlessly incorporate these components.

Two main procedures to transform GPS traces into segmented activities and trips have been proposed in the literature[6,9,10]. One common approach involves a sequential procedure, in which GPS traces are divided into activities and trips based on an empirical time gap threshold, e.g., an activity is recognized if the time gap is larger than 120 seconds[5]. Transportation modes and activity types are then detected for each portion of the segmented GPS traces. Alternatively, activities and trips can be imputed simultaneously considering their differences in the profile of the GPS traces[2,4]. In this case, segmentation between trips and activities is based on a set of predefined rules, which merges the epoch data into segmented activity and travel episodes.

Although both approaches seem feasible as has been demonstrated in several pilot projects[2,7], the consistency between different trips within tours was rarely taken into account. Yet, the transportation mode in different trip segments may be intrinsically based at the tour level in the sense that people may use the same transportation mode within the same tour. Bridging epoch and trip data involves the segmentation issue, which may result in different activity duration and travel episodes. Connecting the trip and tour level involves investigating the consistency issue, which may lead to fully or partly consistent transportation modes.

Therefore, the accurate detection of activity-travel diaries needs an integrated approach, which accounts for the above issues in a systematic way. This paper will develop such an integrated approach for GPS data imputation. The method implements a superimposing algorithm, developed before[3] to address the consistency issue within tours. Instead of evaluating the performance of this approach at the epoch level, this paper will focus on its performance at the aggregate level. Special attention will be given to the accuracy of activity/trip segmentation in terms of transportation modes, activity types, trip frequency and average travel time. The GPS traces data and the associated prompted recall data which were collected recently in the Rijnmond region, The Netherlands, will be used to examine the performance of the proposed approach.

The remainder of the paper is organized as follows. Section 2 will describe the procedure of integrated imputation of activities and trips. Section 3 will introduce the data used in this paper, while Section 4 will discuss the results. The paper will be summarized and concluded in Section 5.

2. Integrated imputation of activities and trips

2.1. Activity-trip segmentation and detection

Previously, a computer program, TraceAnnotator (TA) was developed to transform GPS traces into activity-travel diaries[4]. TA takes GPS trace data as input and generates diary data as a sequence of activities and travel. At the core of TA is the detection of transportation mode and activity episode. Based on the profile differences of activity and travel data, TA simultaneously differentiates between a certain transportation mode and an activity episode. A Bayesian belief network (BBN) model, an advanced learning-based algorithm, is used to replace commonly used ad hoc rules with a dynamic structure. The BBN represents the multiple relationships between different spatial, temporal and other factors, including errors in the technology itself (input), and the facet of activity-travel patterns that we wish to impute from the GPS traces (output). Basically, the network is represented as a directed graph, together with an associated set of probability tables. The nodes of this graph represent causal relationships between the variables. The choice for a BBN was made a priori and was based on the contention that these networks allow a flexible specification of the relationship between the considered variables and the classificatory variable.

The BBN model was applied to single epochs of 3 seconds. Because speed data and other variables used in the classification may change across successive epochs, the classification result may also differ between epochs. The challenge, therefore, is to detect whether this variability is due to inherent fluctuation in the data or that multiple transportation modes were used during the trip. A common approach to deal with this issue is to set some threshold for the length of a stop[8].

Detection of type of transportation mode is usually based on the classification results of all epochs belonging to the same trip, using predefined merge rules. These rules are based on the type of activities and trips and on their time threshold value. The threshold value was defined as 3 minutes in the sense that all trips and activities of less than 3 minutes are merged with other trips or activities.

In addition to transportation mode detection, a geo-analysis procedure was also embedded to identify activity types. More specifically, activity locations were recognized first according to the frequency of the log points in a predefined area. The point with the highest frequency was taken as the activity location. Afterwards, a spatial recognition function was implemented to search the point of interest data sets with a user-defined diameter. The point closest to the recognized activity location was selected and the category of that point of interest node was treated as the activity type.

2.2. Incorporating the full and partial consistency

Because the imputed activity-travel diary is based on epoch data, the imputation process does not guarantee the consistency of transportation modes for different segments of tours[3]. In other words, the transportation mode that people use in different segments of a tour should be the same in most cases. Even if an individual may choose different transportation modes for different trips, the majority will use the same transportation mode for the different segments of a single tour, especially if a transportation mode is difficult to leave behind.

Therefore, the challenge is how to identify the set of hierarchically related tours, and how to detect the (main) transportation mode for each tour. The identification of tours involves the comprehensive investigation of the sequence of activity and travel episodes. The activity location and activity type in conjunction with the timing information provide useful references to identify tours. Furthermore, an algorithm is needed to address the consistency of transportation modes for each tour.

For each participant and each day, we identified tours using a backward search algorithm to recognize identical locations based on location names and spatial information, such as geocoded personal profiles, land use data and a Web-based address data base. In the few cases where location data was missing or incorrect or geocoding unrecognizable, activity locations were classified as unknown. Tours were identified from the complete activity-travel sequences, excluding any missing data.

In order to guarantee consistency at the tour level, we proposed three algorithms that superimpose some degree of consistency in detecting transportation modes based on the conditional probabilities at the epoch level. The algorithms were compared using the prompted recall data as the ground truth.

Results indicated that all three algorithms improved imputation accuracy. The method, which takes the highest frequency for each mode across all trip segments, has the best performance. In spite that the full consistency principle within tours applies to the majority of activity-travel patterns, variations may exist in the whole travel sequence where people change mode during certain segments of the tour. The reason behind such change will in many cases depend on the extent travelers need to pick-up their vehicle. Therefore, a further improvement, called partial consistency, was examined. It was found that including the condition of partial consistency further improved imputation accuracy[3].

Previous studies measured the performance at the epoch level. However, this good performance at the epoch level does not necessarily guarantee good performance at the trip level. Moreover, comparison is meaningful only when timing of activities and trips of the two data sets can be matched. However, respondents differ in terms of their inclination to modify prompted recall data, which results into inconsistent segmentations. Therefore, in this study, we investigate the performance of the proposed integrated imputation algorithms at the aggregated segment level. Thus, we examine the degree of similarity between the imputed daily activity-travel schedule and the validated schedule by comparing characteristics of the schedule at the episode (trip) level.

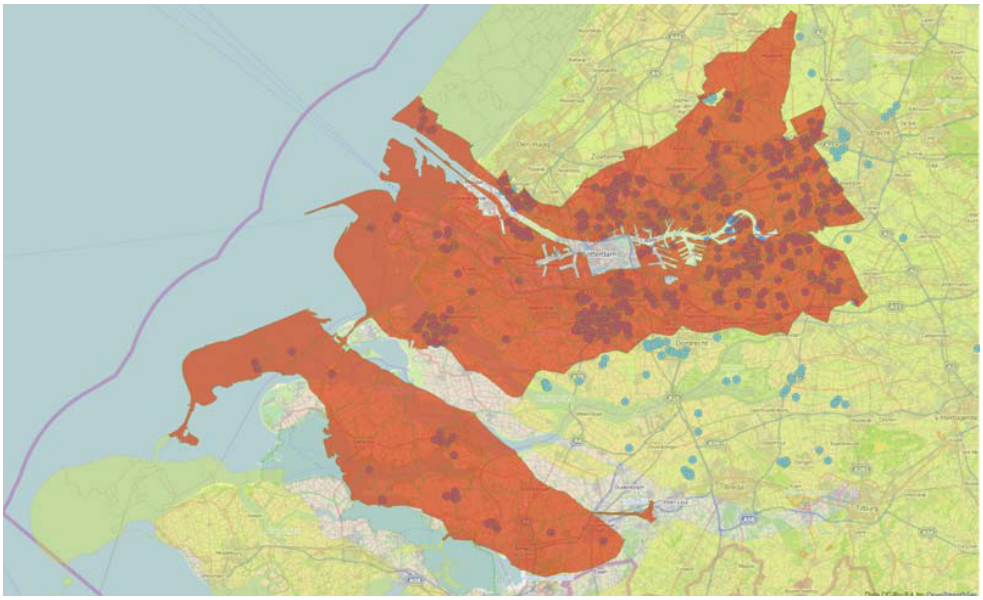

Fig.1 The Rijmond region and locations of respondents.

3. Data

In this paper, we use data collected in the Rotterdam region, The Netherlands. The survey was conducted between 2012 and 2013 in the context of a large-scale research project Participants were randomly selected to join the survey for consecutive three months. A map representing the research area and spatial distribution of residential locations is shown in Fig. 1. Most respondents are located around the center of Rotterdam city.

GPS loggers recording the timing and position of respondents for every 3 seconds were used. Their traces contain information about date, time, longitude, latitude, speed, distance, accuracy of the measurement and number of satellites. TraceAnnotator was used to derive the diaries, which were validated using a follow-up prompted recall instrument, asking respondents to fill out and/or correct inaccurate and/or missing information. The system allows changing, removing and merging the imputed data, and creating new activity/travel data. These data were considered the ground truth to evaluate the performance of different imputation algorithms. Further details about the data collection and/or the prompted recall survey are described elsewhere[1].

Several data sets were used to evaluate the performance of the integrated imputation. The first set consists of the validated data, which although not necessarily error free, are considered to perfectly describe the true activity-travel patterns. These data have been used before to investigate the accuracy of our algorithms at the epoch level. The second set is the imputed data obtained from the GPS traces. This data was also shown to participants in the prompted recall surveys for modification. Moreover, there is the data, which was obtained after applying the superimposing algorithm. It is believed that this data represents the diary of improved quality. In the subsequent discussion, we call the last diary data the superimposed data, while the one shown in the prompted recall surveys is called the originally imputed data. In the subsequent analysis, each of these data sets is used to extract the aggregate information related to activities and trips, respectively. The performance of the proposed imputation algorithms is then examined based on the level of consistency between the different data sets.

4. Results

In order to evaluate the imputation accuracy at the aggregate level, we compare the average statistics of the different data sets. The validated data is from the prompted recall survey, which is normally treated as the ground truth. We consider the validated data as a basis for comparison.

4.1. Descriptive statistics of social demography

Table 1 presents the descriptive statistics of the social demographic profile of both individuals and households. It shows that overall results are consistent with the average of the whole population. Men and women are almost equally distributed. For the age variable, the majority of respondents is in the 35~54 category (39.8%) and the 55+ category (45.3%), while 15% of respondents is under 34 years old. The percentage 55+ is slightly higher than the other categories. The number of cars in the household partly indicates the mobility pattern of the household, i.e. car-dependent or car independent. The percentage of households with 2 or 3 cars is respectively 55.4% and 24.8%. Only 5.2% of the households do not have a car.

The number of people in the household influences the number of activities and/or trips in that more people may conduct more activities. Table 1 shows that two persons households almost represent half of the sample (48.0%), followed four person households (21.1%) and three person households (13.9%). The percentage households with a single person or more than four persons are 9.0% and 8.0%, respectively.

Since commuter trips contribute substantially to traffic congestion, it is important to ensure that the selected sample is representative in terms of work status. Results show that the majority of the respondents is employed, with 45% of the people is employed by agencies, besides the 3.4% independent contractors and 6.1% governmental workers. 21.7% of the respondents are retired. In addition, 3.7% of the respondents is studying, while 12.5% of the respondents does not have paid work (slightly higher than the national average unemployment rate in 2013 of 8.6%).

4.2. Transportation mode

Fig. 2 presents the frequency distribution of the five defined transportation modes. It shows the distribution of transportation modes is similar across different data sets.

Table 1. Descriptive statistics of social demography.

Variables		Frequency	Percent
Gender	Male	317	48.5
	Female	337	51.5
Age	18-34	98	15.0
	35-54	260	39.8
	55+	296	45.3
Number of cars in the household	0	34	5.2
	1	73	11.2
	2	362	55.4
	3	162	24.8
	4	20	3.1
	5	3	.5
Number of people in the household	1	59	9.0
	2	314	48.0
	3	91	13.9
	4	138	21.1
	5+	52	8.0
Work status	Independent contractor	22	3.4
	Employed by agencies	294	45.0
	Employed by government	40	6.1
	Incapacitated	32	4.9
	Unemployed/job-seeking/Assistance	16	2.4
	Retired	141	21.6
	Studying	24	3.7
	Housewife/househusband	82	12.5
	Missing	3	.5
Total		654	100.0

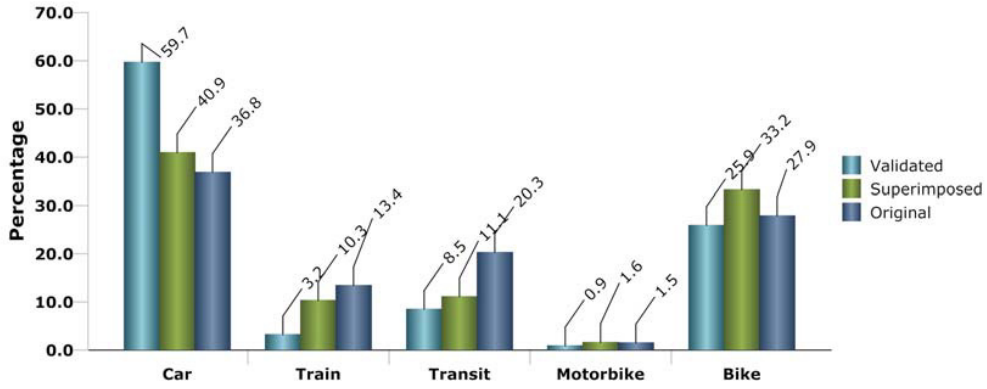

Fig. 2. Frequency of transportation modes.

It also shows that the proportion of different transportation modes differs between the data sets. The originally imputed and superimposed data both have less car trips and more train and transit trips than the validated data. This is mainly due to the wrong imputation of car for transit and car for train. The travel speed may become similar on a congested road (for car and bus) and on a free-flow highway (for car and train), leading to possible mis-classifications.

The figure also shows that the superimposed data are more similar to the validated data than the originally imputed data for all transportation modes, except bike. Especially for motorbike, the proportions between superimposed data (1.6%) and the originally imputed data (1.5%) are similar. This indicates that the superimposed data has a better quality than the originally imputed data in terms of transportation modes.

4.3. Activity type

Fig. 3 shows the results for the different activity types. The table includes eight types of activities for comparison. The shopping activity combines grocery shopping, which people conduct on a daily basis and non-daily shopping, e.g. go to a shopping center to buy clothes. Results show that, for all data sets, the activities related to paid work and shopping compromise the main portion of all activity categories. The imputed data (originally imputed and superimposed) over-represent the proportion work activities, while underrepresenting the proportion shopping activities. In addition, the proportion social activity is substantial (13.6%), relative to the other two data sets. This is perhaps because geo-location information only cannot uniquely identify social activities because social activities can be conducted at different locations, e.g. restaurant, home, which may have multiple functions.

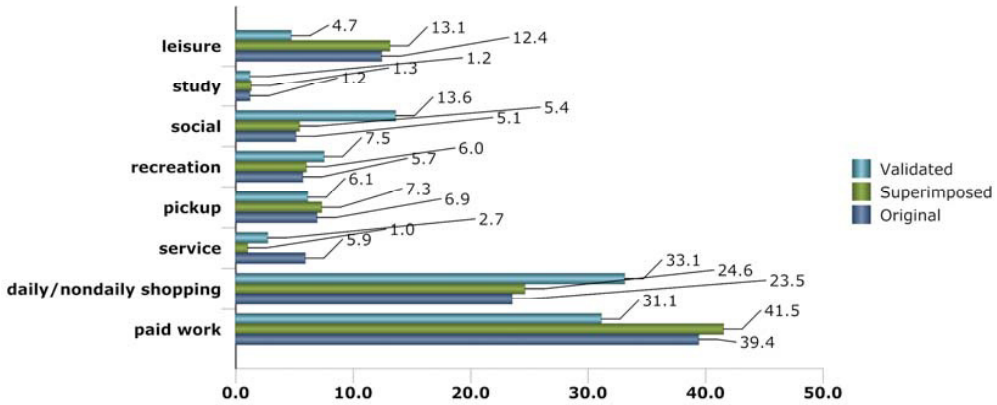

Fig. 3. Frequency of activity types.

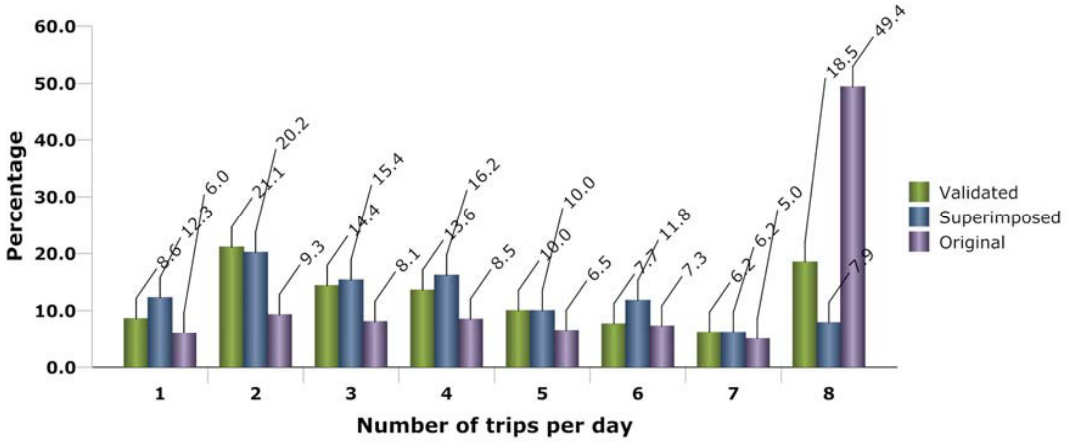

Fig. 4. Frequency of number of trips per day.

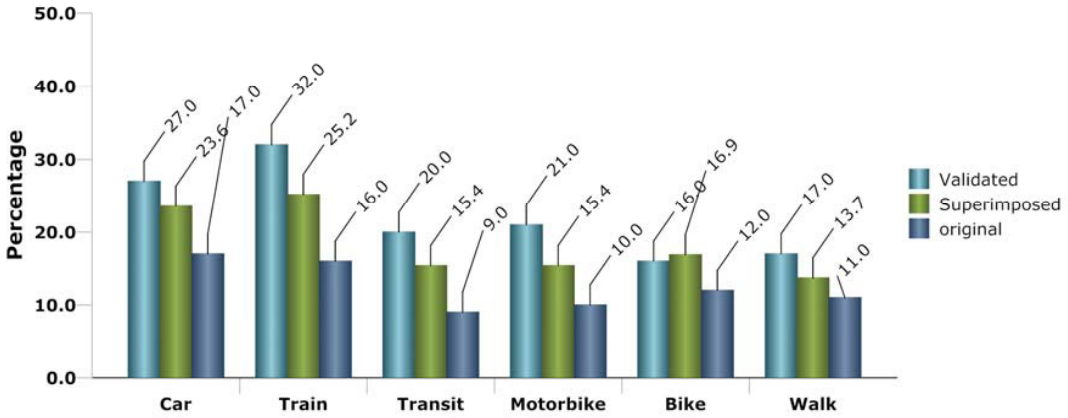

Fig. 5. Average travel time in minutes by different transportation modes.

4.4. Number of trips per day

The number of trips per day is one of the popularly used indicators to examine the level of travel demand. Previous studies have indicated that, relative to GPS surveys, people tend to report a smaller number of trips in a conventional survey[10]. Because the GPS data was collected during multiple weeks, the number of trips was averaged by the number of days for each individual. Results are shown in Fig. 4.

It is evident that for the validated data the highest frequency for the number of trips is two times a day (21.1%). The percentage trips that is higher than eight times per day is 18.5% for the validated data, and 49.4% for the original imputed data. This means that the original imputation algorithm splits too many trips into smaller segments. One expects that the number of cases with more than eight trips per day is rare. In that sense, the superimposed data is more reasonable than the originally imputed data.

4.5. Average travel time

The average travel time by different transportation modes is presented in Fig. 5. It appears that the travel time of the originally imputed data is always the lowest across the different transportation modes for all data sets. This is

consistent with the above finding that the originally imputed data may have over split trips into smaller segments, which have relatively short travel times.

Regarding differences in average travel time, train and bike result into the longest and the shortest travel time among all transportation modes, respectively. This is as expected considering that trips by train in general involve a long travel distance, while bike is normally used for short trips. The superimposed data is more realistic than the originally imputed data.

5. Summary and conclusions

The problem how to transform GPS traces into activity-travel diaries has evoked the development of different algorithms and procedures. The success of data imputation inevitably seems to involve the formulation of complicated rules or implementation of multiple post-processing procedures. Semi-automatic integrated imputation is necessary to ensure a certain level of accuracy.

In spite of the fact that existing approaches succeed in generating activity-travel diaries, it seems that the issue of consistency at the tour level has not been fundamentally addressed. Nevertheless, incorporating the consistency principle within tours seems extremely important, leading to improved imputation accuracy of activity-travel diaries. Different from the results at the epoch level that have been explored in a previous study[3,4], results at the segment level provide another view on the quality of the imputed data. Comparative analyses between the imputed data and the validated data conducted in this study further confirmed our contention that superimposing the condition of consistency leads to more accurate activity-travel diary data.

Although the principle of consistency within tours seems successful, one can argue that it is violated in some special cases, e.g. multimodal trips, use of park and ride facilities or vehicle sharing. However, solutions for these special cases based on the consistency principle are straightforward. For example, one may divide a tour into multiple sub-tours where parking or waiting can be treated as an intermediate node, splitting trip segments because in general people need to park and/or pick up a vehicle at a same place. In these cases, the consistency principle can be directly applied to the identified sub-tours. We consider such refinement in our future work.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 230517 (U4IA project). The views and opinions expressed in this publication represent those of the authors only. The ERC and European Community are not liable for any use that may be made of the information in this publication.

References

1. Feng T, Timmermans HJP. Travel survey using GPS devices: Experiences in The Netherlands In: Rasouli S, Timmermans HJP, editors. *Mobile Technologies for Activity-Travel Data Collection and Analysis*. IGI Global; 2014.
2. Feng T, Timmermans HJP. Transportation mode recognition using GPS and accelerometer data. *Transp Res C* 2013; **37**:118-130.
3. Feng T, Timmermans HJP. Enhanced imputation of GPS traces forcing full or partial consistency in activity-travel sequences: Comparison of algorithms. *Transp Res Rec* 2014 (in press).
4. Moiseeva A, Jessurun AJ, Timmermans HJP. Semi-automatic imputation of activity-travel diaries using GPS traces: Prompted recall and context-sensitive learning algorithms. *Transp Res Rec* 2010; **2183**:60-68.
5. Schuessler N, Axhausen KW. Processing raw data from global positioning systems without additional information. *Transp Res Rec* 2009; **2105**:28-36.
6. Stopher PR, Collins A. Conducting a GPS prompted recall survey over the Internet. *Proc 84th Ann Meet Transp Res Board,* 2005, Washington, D. C., USA.
7. Stopher PR, Wargelin L. Conducting a household travel survey with GPS: Reports on a pilot study. *Proc 12th World Conf Transp Res,* 2010, Lisboa, Portugal.
8. Stopher P, Jiang Q, Zhang Y. Tour-based analysis of multi-day GPS data. *Proc 12th WCTRS Conf,* 2010, Lisbon, Portugal.
9. Tsui SYA, Shalaby AS. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transp Res Rec* 2006; **1972**:38-45.
10. Wolf J. Applications of new technologies in travel surveys. *Proc Int Conf Transp Survey Quality and Innov*, 2004, Costa Rica.