Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 109C (2017) 10–17

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Ambient Systems, Networks and Technologies (ANT 2017)

Methodology for identifying activities from GPS data streams

Vladimir Usyukov*

WSP Group, 100 Commerce Valley Drive West, Thornhill, L3T 0A1, Canada

Abstract

When the global positioning system became available for civil uses in the early 1990s, there was an enthusiasm and anticipation that information stored in GPS data streams would replace the traditional data collection methods, especially in the transportation field. Despite the wealth of GPS surveys available to practitioners to work with, the existing studies have not made much progress to deliver models for identification of activities from GPS data streams. The lack of models for identifying activities prevents the reconstruction of activity patterns stored in GPS data streams. The present study proposes a methodology for the identification of activities using a rule-based and discrete choice modeling. This novel approach uses a rule-based model that implements the properties of home-based tours in the form of the feedback loop in order to allow identification of home activities. This model is inert to the presence of travel characteristics as it can be applied to most multi-day GPS data sets, and not just prompted recall surveys. In regard to the non-home activities, a discrete choice model is calibrated to Transportation Tomorrow Survey (TTS), for identification of work and other activities. The estimated results are positive, as they are compared against the TTS, and are consistent with the observed patterns.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: GPS/GLONASS data; activity identification; travel demand modeling; home-based tours; discrete choice modeling

1. Introduction

The application of global satellite positioning systems, e.g. the US GPS and Russian GLONASS, in the transportation field has opened new and exciting opportunities for the researchers. The transportation field has

* Corresponding author. Tel.: 1-416-826-5702;
E-mail address: vusyukov.at.gmail.com

Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.05.289

commonly used global positioning systems for obtaining data on travel time or speed from the objects carrying the GPS/GLONASS receivers. At the same time, the transportation field has not made much progress in developing better models for identifying activities from the GPS/GLONASS data streams.

The reasons for having better models for identifying activities from the GPS/GLONASS data stream are numerous. First, these models can provide more accurate data than the traditional surveys made by paper or phone recall service[9]. Specifically, these models can remove human errors and make the data available in the digital form immediately, thus improving the turnaround time. One study suggests that the information obtained using these models will replace, rather than supplement, the traditional data collection methods in future[13]. Second, the abundance of activity patterns stored in the data streams will bring benefit to the quality of the travel demand models. The existing travel demand models are based on relatively small samples since data collection is a costly undertaking for most agencies. Unlocking activity patterns will increase data samples used to develop travel demand models and so improve their forecasts. Finally, these models provide a useful insight into the understanding of travel behavior and the driving force behind the travelers' decisions expressed in the form of activity patterns[11].

The travel behavior research for identifying activities from the GPS/GLONASS data streams is still in early stages. Much of the published studies apply some form of the rule-based models for identifying activities[4, 5, 6, 7, 8, 9, 10]. The common limitation of the existing models is the implementation of the rules that are either specific to the study area or the method of data collection. This is a limitation because the results from these models may not be easily replicated or transferred from one area to another and this prevents models reuse. For instance, several studies[5, 6, 7] have developed models that are limited to data collected using prompted recall survey. This means that the actual choices of travelers must be recalled in order for the activities to be matched against the GPS data. The quantity of information that can be obtained using this method is limited as each survey participant needs to be interviewed separately which can be costly in man-hours as well as other areas.

The novelty of our research is in proposing a rule-based model for identifying home activities. This model uses the property of home-based tours, implemented in the form of the feedback loop, to estimate and then validate home activities. Our model overcomes the limitations of the existing models since it does not require travelers to recall their activities or any other array of travel data.

In brief, we used data from the GPS survey of cyclists collected in Waterloo, Canada. Using the procedures outlined by Wolf and others[5, 6, 7, 8, 9], we removed outliers from our sample and then split continuous GPS data streams of each cyclist into individual trips. The individual trips were then processed using our model for identifying home activities. The entire process from data cleaning to identifying home activities was automated in Python to process over two million data points. In regard to non-home activities, the multinomial logit model was estimated from the travel survey and then applied to the trips for identifying work and other activities. As a measure of validity, we compared the results obtained by our model to the travel survey that was not used in the model calibration.

2. Literature Review

There are three main directions used for identifying activities from the global positioning data: rule-based, discrete choice, or machine learning. Table 1 presents a substantive list of research, models, and factors used for identifying activities.

Table 1. Summary of models for identifying activities

Models	Modelled activities	Factors	References
Rule-based	home, work, pick up/drop off, private business, school, shopping, leisure, other	home and work addresses, land use, points of interest, activity start time and duration, frequently visited places	4, 5, 6, 7, 8, 9, 10
Discrete choice modeling	home, work, school, shopping and other	activity start time and duration, land use	11, 12
Machine learning	banking, shopping, gas, meeting, friend, work, school, daycare, pick up/drop off, other	socio-economic attributes, activity start time and duration, land use	13, 14, 15

The application of the rule-based models has been tested in a few small and large sample studies[4, 5, 6, 7, 8, 9, 10]. In a

study using 186 vehicles equipped with in-dash navigation systems, Axhausen et al.[4] collected data for thirty days. In the end, they collected data for nearly a quarter-million of trips. Their model was using home and work addresses, land use, points of interest, activity start time and durations for identifying home, work, school, other, leisure, drop off and pick up, short and long-term shopping activities. One of the assumptions of the model is that home addresses have to be known in order to identify home activities from the GPS data. We believe this is a limitation of the model that should be overcome by making the model less reliant of such an assumption. Our study proposes the methodology how to achieve this task.

The test of discrete choice modeling for identifying activities was evaluated in the New York's study[12]. In the New York's study, a detailed tour composition was collected from a total of forty-nine respondents. Overall, two multinomial logit models were calibrated. The first model was developed for the home-based trips and the second model for the non-home-based trips. Several findings are consistent with those found in other studies, i.e. traveler's behavior for identifying activities is explained by land use, activity start time and durations, and historical dependence variables. The results of the study lacked a discussion of the choice set formation used to calibrate both logit models. The predictive power of the non-home based model is reported to be around 60% which means that there is a room for improvement.

The use of machine learning for identifying activities was tested in several studies[13, 14, 15]. McGowen et al.[13, 14] used a household travel survey to train machine learning model for a total of twenty-six trip purposes. The training of the model required having over a hundred various attributes including age, gender, income, activity start time and durations. The limitation of this model is a requirement for the machine learning model to be trained on a very comprehensive household travel survey. Frequently GPS data sets come without any information about the household composition of the surveyed participants. This problem prevents the application of this model to real-life projects.

3. Methodology

The objective of our study is to develop a methodology for identifying activities from the GPS data streams. This objective is achieved through an implementation of the following steps:

1) Data collection;
2) GPS data cleaning;
3) Identification of trip ends;
4) Developing a rule-based model for identifying home activities;
5) Developing a discrete choice model for identifying non-home activities;

3.1. Data collection

We used three data sets to conduct our study: Transportation Tomorrow Survey (TTS), land use and GPS. TTS is the most comprehensive travel survey conducted in Ontario, Canada which collects a wealth of information on how residents of the southern Ontario travel. We used the 2006 version of the survey because it was the only version available to us. Our study area is located in Waterloo, and TTS collected data from 8,732 households with 23,527 persons, making 60,121 trips in there. Each person in the survey is described by age, gender, driver license possession, employment, occupation and student statuses. In addition, each person provides a snapshot of a typical daily travel pattern. These travel patterns include a trip purpose, the start time of a trip, travel mode, origin and destination zones for trips, and a number of persons in a vehicle.

In regard to the land use data set, it included detailed information of the current land uses in Waterloo. The land use data set contained ten different categories and they are commercial, open area, parks and recreations, postsecondary institutions, public schools, residential, resource and industrial, unknown and water body.

In regard to the GPS data set, it was collected in 2011 from 108 cyclists in Waterloo. This data set contained data representative of 541 survey days or 5 survey days per cyclist. Cyclists were given a low-cost GPS tracking devices to record their daily activity patterns for two weeks. These GPS devices were recording positional data (e.g. latitude, longitude, altitude, instantaneous speed) every five seconds and allowed to gather more than two million raw data points. The GPS data set was collected in the passive form, meaning that it had to be mined in order to obtain the start and end of the trips and activity purposes. In addition to the GPS data set, cyclists were surveyed for a range of the socio-economic and cycling related characteristics, such as household income, possession of a driver license,

age, cycling skill levels, type of safety hazards, among others. A more detailed description of the online survey results can be found in the reference material[3].

3.2. GPS data cleaning

The GPS data was processed the first. The cleaning of GPS data seeks procedures to remove suspicious points from the raw data without the loss of travel information. We are not the first ones to clean GPS data. Hence the procedures evaluated in the similar studies suggested that simple rules are effective to accomplish this task. First, filter out points with the instantaneous speed of zero km/h. This rule is implemented in these studies[3, 6, 7] in order to set the lower boundary for data points. Second, filter out points with the instantaneous speed greater than 75 km/h. This rule is implemented in the study[3] upon which our research is based, and suggested that 75 km/h is the upper boundary for cyclists. Third, filter out points if their altitude value does not correspond to the regional elevations. The acceptable regional elevation for Waterloo is ranged from 200 to 400 meters above the sea level. This rule is implemented in this study[4] and recommended to be used in the absence of the number of satellites used to obtain the data. The data cleaning procedures are automated in Python and processed over two million raw data points. Less than 10% of the original raw data points are removed using these rules. The clean data contained just above 1.8 million points.

3.3. Identification of the trip ends

After the cleaning procedures are applied, the GPS data of each cyclist remained to be a continuous stream of points. The continuous stream of points is necessary to split into discrete events, such as the start and end of a trip (i.e. trip ends), for which the activity purposes could be identified. Based on this study[3], the following procedure is recommended to split the continuous GPS data streams: "A trip is defined a continuous movement from an origin to a destination with no stops longer than 10 minutes". Thus if no movement is detected for longer than 10 minutes, then a trip end is identified in the data stream. Using this procedure, we automated the process in Python. Approximately 1.8 million of continuous GPS data streams transformed to 4,200 trip ends.

3.4. Developing a rule-based model for identifying home activities

To overcome the lack of cyclists' home addresses in our study, we developed a model for identifying home activities from the GPS trip ends. Since there is no universally accepted way to do it, we base our model on the property of any tour: a tour is defined as a sequence of trips starting and ending at the same location[2]. All tours in our study are assumed to be home-based tours, meaning that they start and end at home. This assumption is validated in our model using the feedback loop, as explained in details below.

The assumption that all tours in our study are home-based, finds the support in the instructions given to cyclists. Specifically, cyclists were instructed to start data collection at the beginning of the day and turn the receiver off at the end of the same day[3]. For every cyclist, the very first and the very last trip ends of any daily record are assumed to be their home location. From this assumption, we developed our model, shown in Figure 1 below.

Our model consisted of three sections: 1) select the potential home candidates; 2) estimate a home location from the selected home candidates; 3) validate the initially selected home candidates and find home activities that are missed from the identification. We started by selecting the very first and the very last trip ends of each daily record. These trip ends are then compared with each other, in order to find if they represent the same location. We used 150 meters as the acceptable threshold. In the similar studies[4, 5, 6], the threshold of 200 meters is used. If the distance between the trip ends is less than150 meters, then these trip ends are saved to the *list of potential home candidates*. Otherwise, the next set of records is evaluated.

When the *list of potential home candidates* was populated, we computed the average home location from them. On average, our GPS data set collected data representative of 5 survey days per cyclist or 10 home observations per cyclists. A sample of such size is not ideal but it is large enough to estimate the likely home location of every cyclist.

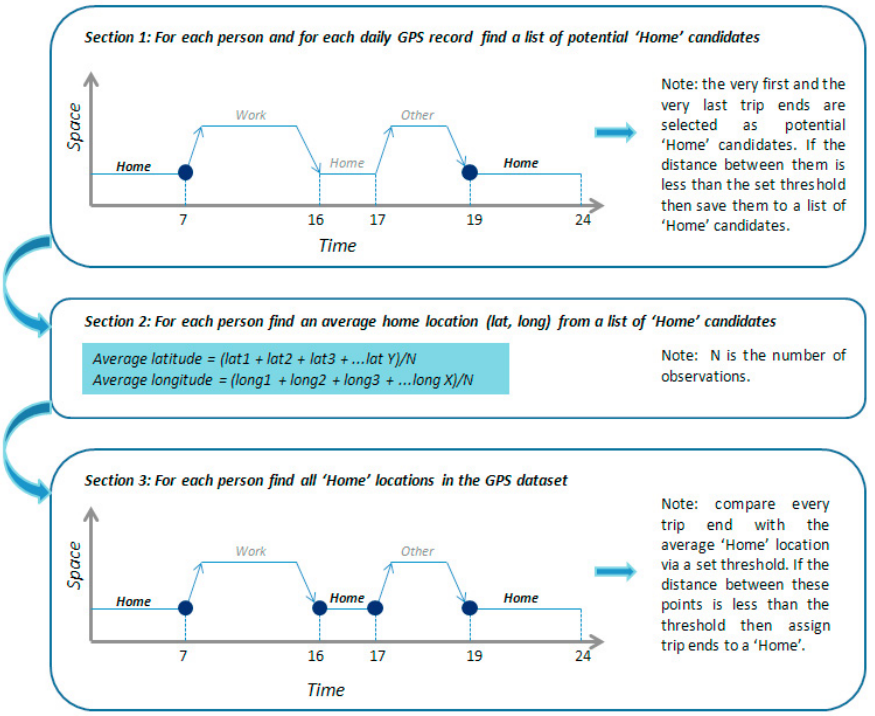

Figure1. Rule-based model for identifying home activities

The last section of our model is designed to validate the initially selected home candidates and to find home activities that are missed from the identification. To accomplish this task we compared the average home location of every cyclist with the full list of trip ends that corresponded to the cyclist. By feeding back the average home location value to the full list of trip ends of a cyclist, we validated those home candidates that were initially selected to the list of home candidates. In addition, this feedback allowed us to found home activities that are missed from the identification. We used 150 meters as the acceptable threshold. The entire process is automated in Python and results are presented in Section 4.

3.5. Using discrete choice modeling for identifying non-home activities

Whereas a rule-based model is suitable for identifying home activities, we used discrete choice modeling for identifying non-home activities. Discrete choice modeling offers a flexible framework to model the variability of activity start time and duration intervals which are common to non-home activities. Following the works of Erath et. al[11] and Chen et.al[12], we used a multinomial logit model of the following form:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \quad (1)$$

The interpretation of this equation is following: the likelihood of selecting alternative *i* amongst all alternatives *n* is a function of the utility derived (V_i) from the attributes of choice *i*, relative to the utility obtained from the attributes of all alternatives[1]. The estimation of parameters in a utility function is based on the maximization of the likelihood function that the chosen alternative has the highest probability of being selected. Following this direction, we calibrated for the β values in equation 2 by maximizing the number of times the chosen activity has the highest probability of being selected.

$$V_{alt1} = \beta_1(Trip\ start\ time_1) + \beta_2(Activity\ duration_1) + \beta_3(Land\ use_1) \quad (2)$$

To calibrate for the β values of the multinomial model, we generated a choice set which combined the chosen and the non-chosen alternatives. The chosen alternatives are the actual choices that are obtained from TTS. The non-chosen alternatives are simulated because these alternatives cannot be measured by other methods, such as *skimming* the network. Based on the literature review[4-15], the following factors are selected significant for identifying non-home activities: activity start time and duration intervals, and a land use type. To generate the non-chosen alternatives, we used these steps: 1) estimate statistical profiles for each activity attribute (i.e. activity start time and duration); 2) simulate the non-chosen alternatives using the estimated statistical profiles for every chosen alternative. The estimation of the statistical profiles and the simulation of the non-chosen alternatives is obtained using a program called *R*.

3.6. Estimation of the model for identifying non-home activities

To estimate the multinomial logit model we used Transportation Tomorrow Survey (TTS). Although each person in TTS is categorized by age, student and employment statuses, only the full-time workers are included in the training sample. This group represents the largest segment of TTS. To estimate the model, we randomly selected 80% of the non-home activities or 13,098 records. The commercial software EasyLogit Modeler is used to estimate the model parameters and the overall fit. The final model specification is shown in Table 2, alongside with t-test results that evaluate the statistical significance of parameters. The absolute ratio is calculated separately for each activity and is a metric of the relative importance of each parameter.

Table 2. Multinomial logit model for identifying non-home activities

Estimated parameters	Activity	t-value	t-statistic	Abs. Ratio
Constant	Other	-0.68	-13.37	8
Start time				
time < 7AM	Work	0.49	13.63	4
7 AM ≤ time <8AM	Work	0.60	16.16	4
8 AM ≤ time <9AM	Work	0.69	17.27	5
5 PM ≤ time <6PM	Other	0.32	5.72	4
6 PM ≤ time <7PM	Other	0.54	10.05	7
7 PM ≤ time <9PM	Other	0.26	5.15	3
Activity duration, hours				
time < 1	Other	0.25	5.09	3
1 ≤ time <2	Other	0.21	4.61	3
2 ≤ time <3	Other	0.08	1.45	1
3 ≤ time <7	Work	1.23	25.77	9
11 ≤ time <14	Work	0.14	2.67	1
Land use type				
Commercial	Other	6.39	22.96	80
Commercial	Work	1.50	4.73	11
Residential	Other	1.17	16.97	14
Resource and Industrial	Other	2.08	21.40	26
Resource and Industrial	Work	2.20	27.51	16

The signs of the parameters are consistent with our expectations. The work activity is represented by activities that start from 7 to 9 a.m., last between 3 to 14 hours, and occur at the commercial, resource and industrial land uses. The absolute ratio shows that both land use parameters have the strongest effect in explaining the work activity.

The other activity is represented by activities that start from 7 to 9 p.m., last between 1 to 3 hours, and occur at the commercial, residential and industrial land uses. The absolute ratio shows that the commercial land use parameter has the strongest effect on the model. The magnitude of the alternative specific constants is low, meaning that most of the effects are captured through the estimated parameters.

The most appropriate evaluation of the model's forecasting ability is to use these utility parameters for identifying activities not used in the model formulation. We used 20% of data from TTS, or 3,274 records, to validate the model. The validation results are summarized in Table 3 as a prediction-success matrix, where diagonal elements represent a correct match between the model and observations. The overall success rate is 86% (2,809/3,274 = 86% correct).

Table 3. Prediction-success matrix for the model

Obs./Model	Work	Other
Work	1985	120
Other	345	824
Type 1	15%	13%

From Table 3, work and other activities are forecasted correctly 85% and 87% of the time. In a similar study conducted by Chen[12], the predictive ability of the model for identifying non-home activities is reported to be 60%. Also, we have computed Type 1 error which corresponds to activities that are identified wrongly by the model. From Table 3, work is mistaken by other activity in 15% of cases, and other is mistaken with work in 13% of cases. In comparison to a similar study conducted by Chen[12], we improved the forecasting ability of the non-home activities from 60% to 86%.

4. APPLICATION OF THE MODELS AND RESULTS

The proposed rule-based and multinomial logit models are evaluated by applying them to the GPS data set of cyclists. As it is covered in Sections 3.2 and 3.3, our GPS data set collected over two millions of raw point from 108 cyclists in Waterloo, Canada. We then clean the raw data points from the outliers and transform the continuous streams of points to discrete events, such as the trip ends are. A total of 4,200 trip ends is identified using these procedures.

The rule-based model is applied the first to the trip ends, for identifying home activities from them. Using the rule-based model we initially identified 1,623 home activities. However, we were forced to remove records that have a missing home location at the beginning or end of a tour. As a result of this action, our sample reduced to 2,328 trip ends which contained home-based tours only. From 2,328 trip ends 1,032 of them are identified as home. Although a portion of the trip ends is removed, the remaining sample is still large to provide meaningful results. The multinomial logit model was applied next, for identifying work and other activities from the remaining 1,296 trip ends. Using the multinomial logit model we identified 602 work and 694 other activities.

The evaluation of the modeling results is discussed next. Due to the absence of the actual responses from the cyclists we used Transportation Tomorrow Survey (TTS) to contrast the share of each activity with the results obtained using our models. A similar process was used by Wolf[5], Bohte[8], Stopher[10], and Erath[11] and proved to be applicable. The results of our comparison are presented in Table 4.

Table 4. Comparison of activities from TTS and GPS data set

Activity type	2006 TTS	2012 GPS data set	Relative accuracy (in %)
Home	40% (10,676/27,049)	44% (1,032/2,328)	90
Work	27% (7,404/27,049)	26% (602/2,328)	96
Other	33% (8,969/27,049)	30% (694/2,328)	91

From Table 4, it is shown that TTS contained 27,049 trip ends and 40% of them are identified as home, 27% as work and 33% as other. Using our GPS data set, a total of 2,328 trip ends is derived and 44% of them are identified as home, 26% as work and 30% as other. The comparison of activity frequencies revealed the following observations. First, the order of activity frequencies is the same between the TTS and the GPS data set. For instance, home activities are the most frequent, other is the second most frequent, followed by the work. Second, the rule-based model overestimated home activities by 4% or 93 records. A slight overestimation of home activities is explained by the presence of the non-home-based tours in the GPS data set. Although it is assumed that the starting and ending point of every cyclist is home, there is a chance that some of the non-home-based tours are registered as home-based and thus inflate the presence of home activities in the GPS data. In comparison to TTS, 1% of tours are registered as non-home-based. Assuming that such an adjustment is valid to our GPS data set, the share of home activities can be reduced from 44% to 43% and bring the results of home shares between TTS and the models closer.

Next, we evaluated the accuracy of our results with respect to TTS by computing the relative accuracy, as shown in Table 4. The proportions of activities obtained from TTS are considered as the base case, and the results obtained

from the modeling are computed relative to TTS. The accuracy of the rule-based model for identifying home activities is computed as 90%. The accuracy of the multinomial logit model for identifying non-home activities is 96% for work and 91% for other activities. In the similar studies evaluated by Chen[12], the accuracy of the model is 90% for home activities and 60% for non-home activities. The accuracy of the models reported by Bohte[8] is 80% for home activities, 88% for work and 60% for other activities.

Overall the proposed rule-based model proved to be a powerful tool for identifying home activities without relying on the prompted recall surveys. The application of the multinomial logit model is an appropriate way to model the complexity of travel behavior common to non-home activities. The results from both models are validated with TTS, then compared to the similar studies and found positive.

5. LIMITATIONS AND FUTURE WORK

This paper describes a methodology for identifying home, work, and other activities from the GPS data set. This study demonstrated a significant improvement in the accuracy of the results, in comparison to the similar studies. The novelty of our methodology is in proposing and implementing a rule-based model for identifying home activities. Using the property of home-based tours, we are able to identify home activities from the GPS data using very limited information. To identify non-home activities, such as work and other, we calibrated and validated a multinomial logit model using TTS.

There are several limitations to our study which presents an opportunity for future work. In our study, we removed tours that have missing home activities either at the beginning or at the end of tours. A different approach is to add the missing home activities to make the tours complete and evaluate the results again. Another limitation of our study is the absence of the actual responses from cyclists to make a direct comparison of results. Despite these limitations, we find this study positive as our knowledge of the travel behavior and specifically about the choice in activity selection has advanced.

Acknowledgements

The author is thankful to Professor Jeff Casello of University of Waterloo for providing GPS data set, Jeffrey Newman of Northwestern University for providing EasyLogit and to Karl A. Peuser for improving the manuscript.

References

1. Ben-Akiva, M., Lerman, S. (1985), Discrete Choice Analysis: Theory and Applications to Travel Demand. MIT Press Series in Transportation Studies
2. Ortuzar, J., Willumsen, L. (2011), Modelling Transport, 4th Edition. Published by John Wiley & Sons, Ltd
3. Rewa, K., (2012), An analysis of stated and revealed preference cycling behavior: a case study of the regional municipality of Waterloo, (MSc Thesis), http://hdl.handle.net/10012/6910
4. Axhausen, K. W, Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M. (2004), 80 weeks of GPS-traces: approaches to enriching the trip information. Transportation Research Record 1870, 46–54.
5. Wolf, J., Guensler, R., Bachman, W. (2001), Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: Paper Presented at 80th Annual Meeting of the Transportation Research Board, Washington, DC.
6. Wolf, J. (2001), Using GPS data loggers to replace travel diaries in the collection of travel data. Ph.D. dissertation, Georgia Institute of Technology.
7. Stopher, P., FitzGerald, R., Zhang C., Clifford, E. (2008), Deducing mode and trip purpose from GPS data. Institute of transport and logistics studies. The Australian Key Centre in Transport and Logistics Management. Paper number ITLS-WP-08-06
8. Bohte, W., Maat, K. (2009), Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large scale application in the Netherlands. Transportation Research Board C 17
9. Krygsman, S., Nel, J. (2008), Deriving transport data with cell-phones: methodological lessons from South Africa. 8th international conference on survey methods in transport: harmonization and data comparability.
10. Stopher, P., Li, S. (2013), A process for trip purpose imputation from Global Positioning System data. Transportation Research Part C 36.
11. Erath, A., Charikov, A. (2012), Activity identification and primary location modeling based on smart-card payment data for public transport
12. Chen, S., Gong, H., Lawson, C., Bialostozky, E. (2010), Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A 44, 830-840
13. McGowen, P. (2006), Predicting activity types GPS and GIS data. Ph.D. dissertation, the University of California at Irvine.
14. McNally, M., McGowen, P. (2006), Predicting Activity Types from GPS and GIS Data. Presented at the 86th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.
15. Montini, L., Rieser-Schussler, N., Horni, A., Axhausen, K. (2014), Trip purpose identification from GPS tracks. Transportation Research Record: Journal of Transportation Research Board, Volume 2405, pp. 16-23