Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 96 (2013) 310 – 319

Procedia
Social and Behavioral Sciences

13th COTA International Conference of Transportation Professionals (CICTP 2013)

Post-processing procedures for passive GPS based travel survey

Jian Liu [a,b], Hongjiang Zheng[a,]*, Tao Feng[b], Shuning Yuan[b], Hongyang Lu[b]

[a]School of Informantion Engineering, WuHan University of Technology, WuHan, 430070, China;
[b]The MOT Transport Satellite Navigation Industrialization Center, China Transport Telecommunications & Information Center, BeiJing, 100011, China

Abstract

A challenge in posteriori data processing for passive GPS based travel survey, which constitute the heart of this paper, is to develop a series of methods to automatically restore the sequences of data points, both in space and time. It means the trips and activities occurred in the survey time should be identifiable chronologically and those identified by the program should respect this definition convention. Reference to the research results of our colleagues, and by combining the experiences of other French travel survey and personal mobility survey at Lille, a series of methods has been developed and put into application. The data outcome is ready for further applications.

Keywords: post-processing procedures; GPS data; travel survey;

1. Introduction

Compared with the GSM and Wi-Fi positioning, GPS positioning is more accurate, and wider coverage, even the most relevant support software is more substantial. So that, GPS based travel survey is the most common research among the others mobile technologies. Mid 1990, it is already used in the vehicle survey (Lee-Gosselin,2006, Murakami,1999, Ueno,1999, Wolf,1999, Flavigny,1998). Recently, an application of large scaled GPS based travel survey is in Netherlands (Bohte,2008) in which, an internet recall survey is used to correct the destination, modal choice, or complete the missing data. They argued that every GPS-based method needs validation of the data by its respondents by showing them the derived results and asking for validation, correction and additions of trips and trip characteristics. A little early, Stopher and Collins (Stopher,2005) have shown the

* Corresponding author. Tel.: +0-86-18606856825.
E-mail address: hongjiangzheng@whut.edu.cn

doi:10.1016/j.sbspro.2013.08.038

value of an internet recall survey. But in our options, if the outcomes of data processing are good enough to compare with the conventional survey, any form the recall survey will be no necessary.

A couple of authors have started to address these problems (De Jong,2003, Chung,2005, Tsui,2006, Flamm,2007). Basically, all approaches contain individual modules accounting for: (1) Data filtering, (2) Detection of trips and activities,(3) Mode stage determination, 4) Mode identification.

Some authors include additional features such as the merging of stages after the mode detection (De Jong,2003) or a feedback between the map-matching and the mode detection (Tsui,2006).

Most recently, a paper of Nadine Schuessler and Kay W. Axhausen (2008) showed us that a GPS data processing is developed to identify trips and activities and their characteristics from GPS raw data without further information. However, without additional information, a lot of post-processing work is required to derive data that can be used for analysis and model estimation. These post-processing procedures are still an ongoing research issue.

All these contributions of our colleagues, gives us the opportunity to stand on the shoulders of giants to complete our own procedures. The remaining space, we will be glade to share our methods about Data filtering, Segmentation, Estimation data missing, Mode determination, Distinction activities and transfer point, liking segments to trip and group stop points to visited location, which have been put on application in the French nation travel survey (Philippe,2008) and personal mobility survey at Lille.

2. Source of information and errors

The GPS receiver comes to record the coordinates of its position every few seconds. Introducing geographic information system, there will be three sources of information that play different roles in our algorithm.

- GPS data: Where it is at what time
- GIS (Points of interest, POI): The environment and surroundings around the point recorded by receiver (place type)
- GIS (Networks): road networks, rail networks (routes chosen)

Between these three sources of information, there is only one piece of information present in the past. That is the discrete GPS data. All the GIS background is seen as static state and they are not changing with the travel behavior of the participants.

This information makes us to restore the past, but they are also the sources of noise and errors in parallel.

- GPS data: Imprecision of position data, data missing at restart or in bad conditions
- GIS: Imprecision of POI and networks, lack of POI, static map not dynamic
- Human: GPS forgotten, No recharged, No turn on

For post-GPS data processing, there are three key issues (Schuessler,2008): (1) How to identify individual trips and activities? (2) How to derive the modes used by the participants? (3) How to get the selected routes on the network? The algorithm should resolve all these three questions, by extract maximum the information from all the three sources and overcoming their errors and noises.

3. Source of information and errors

3.1. Data cleaning

In the survey period, sometimes, GPS is used in poor conditions, such as cold/hot restart, or receiver on a position where evaluated HDOP is high, or the accuracy is reduced by multi-path effect or atmospheric layers and so on (Wilson,2010, Wikipedia,2009, Sharif,2004). Under the conditions mentioned above, the points recorded will be very absurd, which will make the next steps output the wrong results. Thereby, filtering these points is necessary.

3.1.1. Method: number of GPS satellites and DOP

As we know, 4 satellites are needed to solve a mathematical equation in 4 unknowns in 3 dimensions plus the offset of the receiver clock time and satellite(Wikipedia,2009, Sharif,2004). In order to maintain maximum data that is useful, we leave aside altitude, which is not very usable in our case. So, only 3 satellites will be enough. We keep the altitude unclear, but which is quite accurate on the plan. This is not to say that we lost altitude. This means that there will be points where altitude is unfaithful in data collected.

And distance error of a single point recorded by GPS respect Rayleigh distribution, and relationship between HDOP (horizontal dilution of precision) and distance error on 2D horizontal is “DRMS (HDOP) = HDOP × UERE” (Wilson,2010, Sharif,2004) (DRMS is square root of the average of the square errors, and UERE is user equivalent range error).

When a point recorded with HDOP equal 5 and the normal precision is about 5 meters, so the precision of this point is about 25.

Based on our analysis, in the condition UERE equals 5, and receiver record every 1 seconds, we adopt the criteria require at least 3 satellites to be available and HDOP (horizontal dilution of precision) less 5.

If $\left(HDOP_i \leq 5\right) \& \& \left(nSate_i \geq 3\right)$, then P_i valid;

If not, P_i invalid; wherein $i \in [1, N]$

3.1.2. Estimation of real speed

In our hands, there are two kinds of speed, first, speed recoded by receiver, another, speed calculated by the coordinates of two points successive. Who is more approach the real speed? We know that speed recoded by receiver is an instant speed who can only present the speed at the moment that the point is recorded. That means he is the best candidature when points are recorded every second without any gap.

On the contrary, speed calculated by the coordinates of two points successive is the average speed between the two points registered. But due to the imprecision of the coordinates, the precision of the speed declines with reduce of the distance between two points successive. Given a speed constant, the precision increases with augment of the interval between two points successive. That means if receiver lost the signals for a quite long time, that speed will be more approach the real speed.

Since interval between two points successive is not always equal to the Interval settled, the gap could be several seconds or some days, and the interval settled could change for different receiver the formula have to be a little “universal”.

In application, we use the weighted mean of two speeds to present the real speed. And the weight for each is the function of the interval between two points successive and the interval settled.

$$v_{es} = EXP(0.09531 \frac{In_{cg}\sqrt{In_i}}{In_i - 1})v_i + (1 - EXP(0.09531 \frac{In_{cg}\sqrt{In_i}}{In_i - 1}))\bar{v}_{(i-1,i)} \tag{1}$$

where In_i and In_{cg} are interval between two points successive and interval settled, respectively. v_i, $\bar{v}_{(i-1,i)}$ and v_{es} stand for speed recorded by GPS receiver, average speed between two points successive, and estimated speed, respectively.

3.2. Data smoothing

As we discussed above, Data cleaning methods “AS” and “SH”, they both can not remove some kinds of aberrant points, especially small error. So, in order to minimize the influence of the errors and smoothing the small errors, this step is necessary.

The smoothing method is simple: if the estimated speed of a point is over than the average speed plus 2 times the standard deviation of the "n" points before and also above the average speeds plus 2 times the standard deviation of "n" points after; or that is well below the average speed minus 2 times the standard deviation of the "n" points before and also below the average speeds minus 2 times the standard deviation of "n" point after; then we correct that speed as the average speed of these 2 "n" point. The mathematical expression will be as follows.

Note $\bar{v}_{es_{n-}} = \frac{1}{n}\sum_{k=i-n}^{i-1} v_{es_k}$, $\bar{v}_{es_{n+}} = \frac{1}{n}\sum_{k=i+1}^{i+n} v_{es_k}$

$$\sigma_{n-} = \frac{1}{n}\sqrt{\sum_{k=i-n}^{i-1}(v_{es_k} - \bar{v}_{es_{n-}})^2} \quad \sigma_{n+} = \frac{1}{n}\sqrt{\sum_{k=i+1}^{i+n}(v_{es_k} - \bar{v}_{es_{n+}})^2} \quad ,$$

If $(v_{es_i} > \bar{v}_{es_{n-}} + 2\sigma_{n-}) \&\& (v_{es_i} > \bar{v}_{es_{n+}} + 2\sigma_{n+})$

then

$$v_{es_i} = \frac{1}{2}(\bar{v}_{es_{n-}} + \bar{v}_{es_{n+}}) \tag{2}$$

And we define "*n*" as following: $n = \left\lceil \frac{9}{In_{cg}} + 1 \right\rceil$

Where *n* is number of upstream or number of downstream, and -, +, and $\lceil \ \rceil$ stand for upstream, downstream and ceiling function, respectively.

As for those interval between two points successive is too great but drop in the set of upstream or downstream, the values of speed before or after that interval will not count in the set of upstream or downstream.

3.3. Cutting the chain GPS data into segment in moving or stopping

Since all points recorded in one memory, GPS raw data those we gotten is a chain. In order to derive the trips and activities, first of all we have to cut the chain into segments. Then, we distinguish the moving part and the stopping part for each segment.

3.3.1. Chain cutting

Criteria 1:

In 2003, Zmud, Johanna and Jean Wolf (Zmud,2003,) argued in their paper, that 120 seconds stopping is the right stopping time to cut the chain off, then this threshold have been confirmed by other researchers. According to the statistic of short time activities in French nation travel survey 1994, and the experiences of the data processing, and of the daily life, the threshold, 120 seconds, seems like a good value for us either.

When the speed recorded is low, we believe that the status of respondent is stopping. And if the speed of successive records is low, the time between two records is added, is called "stopping time" until a high speed for an upcoming recording (That recording is considered as the first point of next moving period). So, if "stopping time" is greater than 120 seconds, the records are cut there. Then we get two different segments. And after thousands of testing, we take 1.1km/h as the boundary of moving and stopping.

To avoid the exception high speed or the exception of boundary speed of moving and stopping, that makes the wrong cut, we add a modal of exception control. "Moving factor" is given as

$$n = \begin{cases} 1 & In_i \le In_{cg} \\ \sqrt{\dfrac{In_i}{In_{cg}}}\dfrac{v_{es_i}}{1,4} & In_{cg} < In_i \le \dfrac{S_{Ta}}{2} \end{cases}, \quad plafond = \begin{cases} \dfrac{8}{\sqrt{In_{cg}}} & In_{cg} < 16 \\ 2 & 17 < In_{cg} \end{cases} \tag{3}$$

Where S_{Ta} is threshold of stopping time 120 seconds.

Once we meet a height speed point after some stopping points, if “stopping time” is greater than 120 seconds, then we will calculate the continues “moving factor” until the added “moving factor” exceed the “plafond”, we cut the chain at the end of “stopping time”, if the “plafond” is not exceeded, these part of moving will be ignored, and “stopping time” will continue to be added.

Criteria 2:

If there is a period that receiver lost the signals of GPS, interval between two points before and after this gap will be great. The method “estimation real speed” can approach a speed for this gap. But this speed can’t tell us how much time the respondent was in moving and how much time he stopped during the gap. If he is moving during entire gap, then speed should be high. In our daily life, normally, the longer trip, the higher the speed. So we created an empirical threshold function in gap time as follows.

$$S_v = 0.00344 In_i + 1.58 \tag{4}$$

If the speed is higher than the result of the formula then we considered that respondent is in moving, if not, he is in stopping.

Start point and end point of a gap in moving, could be the point of changing transport mode, especially for public transport. So, we also cut the chain if the gap is considered in moving and gap time is more than 300 seconds.

Criteria 3:

Imagining what has happed, an 8 hour’s gap with average speed 4k/m. It’s probably one or several trip missing. So in this case, Start point and end point of a gap don’t belong to the same trip, and stopping and moving is not continue. Therefore, we cut the chain if the gap is considered in stopping and gap time is more than 600 seconds.

3.3.2. Identify the moving part and the stopping part for each segment

After that, we distinguish the moving part and the stopping part for each segment to get pure moving trace and stopping point.

3.4. Assuming the mode choice

Each mode has its own characteristics (see Table 1.) These elements will be included in the GPS data collected. The different characteristics of various modes of transport offer us the references to develop a method to infer the mode for each segment.

To make development and evaluation of the method easier, we limited types of modes in 7, which covered most of the trips. They are: walking, bicycle, car / motorcycle, bus, subway, train or plane.

Moreover, the method has another role, to distinguish fault segment that should be a part of the stopping.

We take 95th% maximum speed, average speed without stop, median speed, 95th% maximum acceleration, data quality, lasting of segment, maximum signals lost, and the ratio of maximum signals lost as index to identify the mode choice. Data quality is defined like number of points recorded divide number of points should be recorded. A maximum signal lost is the biggest gap in the segment. Ratio of signals lost defined as maximum signal lost divide lasting of segment.

Table 1. Comparison of characteristics of different modes

Index Mode	average speed without stop (Km/h)	median speed (Km/h)	95th% maximum speed (Km/h)	95th% maximum acceleration (m/s^2)	data quality	lasting of segment (s)	maximum signals lost (s)	Ratio of signals lost
walking	2-7	1.5-7	3-12	<4	>0.15	>30	<300	<0.5
bicycle	7-20	6-20	12-24	<4	>0.3	>30	<300	<0.5
car/ motorcycle	9-120	1.5-120	24-200	2-15	>0.3	>60	/	/
Public Transport 1 (bus)	8-80	6-80	15-100	<15	<0.3	>120	10-120	/
Public Transport 2 (metro)	7-50	/	7-80	/	<0.05	>240	>180	/
Train	50-330	/	200-1500	/	<0.01	>1800	>1200	/
plane	200-800	/	50-400	/	<0.1	>1800	>600	/
Stopping 1	<7	<1.5	<12	<4	>0.1	/	<300	/
Stopping 2	<7	/	<12	/	/	30-180	/	>0.5
Stopping 3	<20	/	<20	/	/	<30	/	/
Stopping 4	<3	<1.5	<3	<4	<0.1	/	/	/
Stopping 5	<6	/	<6	/	/	240	/	>0.97
No definition	Other values	Other values	Other values	Other values	Other values	Other values	Other values	Other values

In determining the values of all indices, we need only one set of data to fix a single mode or to stopping or not definition. So that one segment has only one mode.

More than 300 samples have tested to evaluate the method. With five sets of indices values, all stopping points is correctly detected, and no segment in the movement are seen as stopping points

All car modes have been detected, about seven in ten bus and metro mode are recognized as public transport 1 or 2, and three in ten are recognized as car. 5% walking is found as no definition, others done well. Two trips (segment) has correctly detected on train. There are no plane and bicycle mode in the samples, and no segment identified as on plane or by bicycle.

In the next step, bus stop and metro station will be offered by GIS POI to help us further distinguish bus and metro.

3.5. Group stopping point into visited location

The mobility of people during a week or longer, shows some regularity. In particular, some stopping point will be repeated at high frequency, such as home, schools, offices, supermarkets and even restaurants, stadiums, subway station and so on. By comparison of all the stopping points during the survey period, we could group the stopping points who present the same locations.

3.5.1. Geometry method

Above all, we arrange all of the stopping point in order of chronology. The first stopping point which is the start of first segment is usually the place of home. Starting with this point like a "seed", taking this point as the center, then draw a circle with a certain rays distance. If there are other(s) point (s) included in the circle, calculate the gravity center of the points in the circle. And then taking the center of gravity as the new center, draw another circle with the same ray length. If the points in the new circle are not the same as in the last circle, recalculates the center of gravity, and again ... until the points in the new circle do not change any more, no more enter, no more exit. As so, we come to find in a circle contains the maximum points. And we adopted the center of the circle as the last position held, and the points in the last circle are regarded as one point. After that, go down to find another "seed", the nearest points does not include in any circle. Then restart the cycle…

Converge to the place where the density of points is the highest, even the "seed" doesn't in the last circle (Fig.1.).

One point cloud belongs to different circle (Fig.2.). To resolve this problem, we make the points belong to the circle whose centre that the point is closest to.

The circle can't too big to identify two location as one, can't too small to separate one location into two. In application, we use 45 meters as the circle's ray.

3.5.2. Logic method and identify location type

If the survey period is long enough, among the visited location, we can find the most visited (Table 2.).Among them, the most unique is home. It is usually a place where we pass the night, and also the place the most often visited. With these two features, we can probably determine the location "home". Followed by the office, school, they are often visited at high frequency, and second longer stay. Particularly, service station or transfer point, such as bus stop, metro station, may have a high frequency, but stay time on the point is quite short.

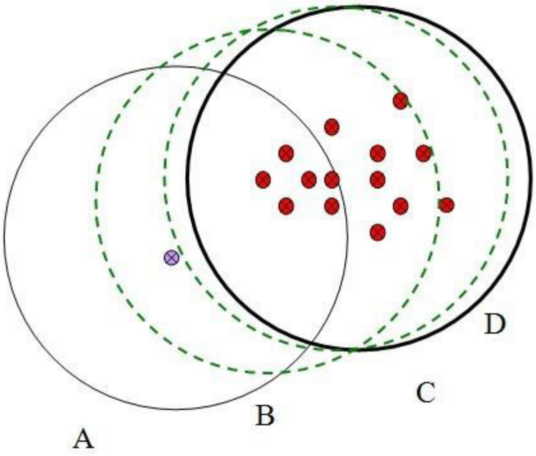

Fig. 1. The basic principle illustration of geometry method

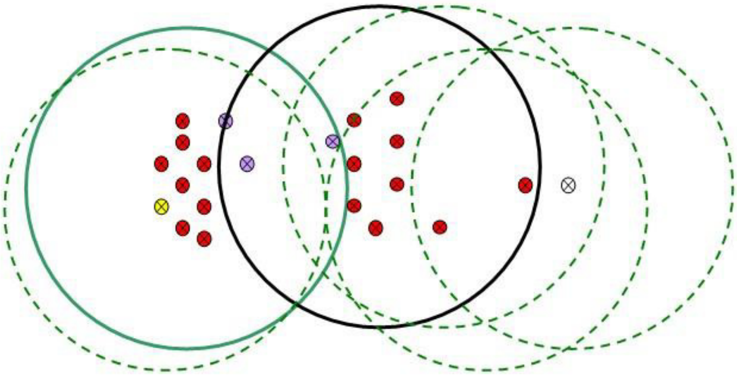

Fig. 2. Another situation to group stopping points with geometry method

Therefore, with these characteristics were mentioned, we can obtain some type of a certain visited location. First, hierarchize the waiting time on the stopping point according to the normal travel behaviours (Table 2).

Table 2. Types of visited location

Stop time (s)	Pass night	>10800	3600-10800	1200-3600	400-1200	120-400	<120	autre
category	1	2	3	4	5	6	7	0

A visited location, grouped by geometry method, containing most the stopping point of category 1, is the place "home". When there are others visited location contain the stopping point marked category 1, and if the distance between them is smaller than 150 meters, on think that they are the same place, "home".

If the number of stopping points in category 2 or 3 included in a visited location more than half of survey days, we mark the "VRL" (visited regularly over a long period). If there are several "VRL", they will be numbered, and the most frequent is "1".

Same for "VRS" (visited regularly over a short period), when the number of stopping points in category 5,6 and 7 is more than half of survey days, we mark as "VRS". If there are several "VRS", they will be numbered, and the most frequent is "1".

Obviously, "VRL" is rather an activity, and "VRS" could be a transfer point like bus stop or subway station.

3.5.3. Evaluation of the methods

A sample of six respondent product 415 stopping points, in which 105 is identified as "home", and 185 visited locations. About 9% most visited location group 45% of the stopping point. 31% visited location contain 70% of stopping point. 68% of visited locations contain only one stopping point.

Is that reasonable if there are 68% locations that we visited only once during 10 days?

3.6. Connect segments to trips

A trip is the part of the connection of two activities. It can be combined in variety modes of transport.

After having distinguished activities and transfer point, the solution is simple. We start with an activity, and then connect the successive segment, until the end of a segment is activity.

However, because of lack of the good method to choose the right POI, and poor found ratio, and without the feedback from map-matching, determination of the activity, point transfer, is not so favorable than we imagine. Despite all, thanks to the mode assuming and logic location type identification, we still have developed some rules to convert the segments to trips, and distinguish temporary stopping, transfer point and activity.

(1) If the stopping time is shorter than 120 seconds, this segment is linked with the next.

(2) If the mode of a segment is public transport 1, 2 or already concreted to metro, and the mode of previous segment is walking, and stopping time between these two segments less than 900 seconds, these two segments is connected. In addition, if the transfer point is marked as "VRS", the threshold value time increase to 2100 seconds.

(3) Same for the public transport + walking, and stopping time less than 300 seconds, then connects the two segments.

4. Conclusion and Perspectives

Post-processing to derive trips and activities from GPS data is a complicated and trivial thing. You have to dig out lots of hidden details, pay attention to every detail and re-rang them. No outcome is independent. They can modify each other. So feedback becomes the key to the quality of the results. Length of staying time in a place can help to find the right POI, the correct POI help us to distinguish the activities and transfer point. Transfer point can correct or confirm the mode of transport. But, at first, mode of transport derived from the characteristics of segment can help to find the right POI. But they do have some order in priorities. Data clearing must be done at the first place and liking segments to trip is certainly the last step.

Fuzzy mathematics seems will be another key to the algorithm, which will be very useful in mode determination and POI finding. Since the nearest POI is always not the best, and the speed rang of each mode cover each other.

Because every GPS receiver has his characteristics, the transport in each city is different, unique. So every parameter motioned in the article can't adapt automatically to other projects. They need to be readjusted.

As the technology matures, the standard or universal processing software will be ready in the next few years to allow devices to be used more widely. In this way, there will be less expenditure on survey, lighter burden on respondents, and faster results.

Acknowledgements

This work was supported in part by the National High Technology Research and Development Program of China under Grant 2013AA122403, the National Natural Science Foundation of China under Grant 51268051.

References

Bohte, W. and Maat, K. (2008). Deriving and Validating Trip Destinations and Modes for Multi-day GPS based Travel Surveys: an Application in the Netherlands. *presentation and publication at the 87th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2008

Chung, E.-H. and A. Shalaby (2005). A trip bases reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28 (5), 381-401.

De Jong, R. and Mensonides, W. (2003). Wearable GPS device as a data collection method for travel research . Working Paper, ITS-WP-03-02, University of Sydney, Institute of Transport Studies, Sydney.

Flamm, M. and Kaufmann, V. (2007). Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behavior adaptation processes during life course transitions. *presented at the 11th World Conference on Transportation Research*, Berkeley, June 2007.

Flavigny, P.-O., Hubert, J.-P., Madre, J.-L. (1998). Analyse de trafic routier observé par GPS et comparaison avec d'autres sources statistiques. *Rapport pour* l'ADEME.

Lee-Gosselin M., Doherty S.T., Shalaby A., (2006). Personal Data Collection Using Mobile ICTs: Old Wine in New Bottles?. *Second International Specialist Meeting on ICT*, The Netherlands, 10-11 November 2006.

Murakami, E. and D. P. Wagner (1999). Can using Global Positioning System (GPS) improve trip reporting?. *Transportation Research*, Vol. 7C, No. 2/3, 1999, pp. 149-165.

Philippe, M., Roux, S., Yuan, S., ect.(2008), A study of non-response in the GPS sub-sample of French National Travel Survey 2007-08, *8th international conference on survey methods in transport:* Annecy, France, may 25-31, 2008

Schuessler, N. and Axhausen, K.W. (2008), Identifying trips and activities and their characteristics from GPS raw data without futher information, *8th international conference on survey methods in transport*, Annecy, France, may 25-31, 2008

Sharif, M., Stein, A., Schetselaar E.M.(2004). Integrated approach to predict confidence of GPS measurement, *International Institute for Geo-Information Science and Earth Observation (ITC)*, http://www.itc.nl/library/Papers_2004/peer_conf/sharif.pdf.

Stopher, P.R. and Collins, A. (2005). Conducting a GPS Prompted Recall Survey over the Internet. *Presented at 84th Annual Meeting of the Transportation Research Board*, Washington D.C., 2005.

Tsui, S. Y. A. and A. Shalaby (2006). An enhanced system for link and mode identification for GPS-based personal travel surveys. *Transportation Research Record*, 1972, 38-45.

Ueno, M., Noël, N., Doherty, S.T., Lee-Gosselin, M.E.H, Théberge, F. & Sirois, C. (1999). Extending the scope of travel surveys using differential GPS. *Proceedings of ION-GPS 1999*, Nashville, Tennessee (CDROM)

Wikipedia (2009),Global Positioning System, http://en.wikipedia.org/wiki/GPS, Accessed July.25, 2009

Wilson, D.L.(2010) .HDOP and GPS horizontal position errors, http://www.erols.com/dlwilson/gps.htm, Accessed July. 30, 2010

Wolf, J., Guensler, R., Washington, S., Frank, L. (1999). The Use of Electronic Travel Diaries and Vehicle Instrumentation Packages in the Year 2000 Atlanta Regional Household Travel Survey. *Proceedings of the TBR Conference on Personal Travel* : The Long and Short of It, Washington DC, June 1999.

Zmud, J. and Wolf, J.(2003), Identifying the Correlates of Trip Misreporting – Results from the California Statewide Household Travel Survey GPS Study. *presented at the 10th International Conference on Travel Behavior Research,* Lucerne, August 2003