Transportation Research Part A 44 (2010) 830–840

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study

Cynthia Chen [a,*], Hongmian Gong [b], Catherine Lawson [c], Evan Bialostozky [b]

[a] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, United States
[b] Department of Geography, Hunter College of the City University of New York, New York, NY 10065, United States
[c] Geography & Planning, University at Albany, Albany, NY 12222, United States

ARTICLE INFO

Article history:
Received 3 November 2009
Received in revised form 10 August 2010
Accepted 24 August 2010

Keywords:
Household travel survey
GPS/GIS technology
Travel mode detection
Trip purpose identification

ABSTRACT

The combination of increasing challenges in administering household travel surveys and advances in global positioning systems (GPS)/geographic information systems (GIS) technologies motivated this project. It tests the feasibility of using a passive travel data collection methodology in a complex urban environment, by developing GIS algorithms to automatically detect travel modes and trip purposes. The study was conducted in New York City where the multi-dimensional challenges include urban canyon effects, an extreme dense and diverse set of land use patterns, and a complex transit network. Our study uses a multi-modal transportation network, a set of rules to achieve both complexity and flexibility for travel mode detection, and develops procedures and models for trip end clustering and trip purpose prediction. The study results are promising, reporting success rates ranging from 60% to 95%, suggesting that in the future, conventional self-reported travel surveys may be supplemented, or even replaced, by passive data collection methods.

1. Introduction

As transportation planning models move towards the use of micro-simulations of daily activities and travel patterns (Miller et al., 2004; Pendyala et al., 1997), pressure mounts to increase the quantity and quality of travel survey data. Conventional survey methods (e.g., CATI surveys[1]) require participants to log all their activities and trips, and self-report them as accurately as possible. These self-reported surveys face problems including: declining sample sizes (Stopher and Greaves, 2007); increasing non-response rates (Wilson, 2004); non-representative samples (Murakami, 2008; Stopher and Greaves, 2007); missing activities and trips (Pearson, 2004; Wolf et al., 2003); and imprecise travel time (Stopher et al., 2005).

At the same time, the technological innovations during the last decade have blossomed. GPS/GIS technologies could potentially improve data quality and reduce respondent burden in the collection of household travel survey data. GPS units record paths traversed by users as a series of time and geo-referenced points with associated variables such as latitude, longitude, time, speed, and heading. Together with information on the transportation network and land use patterns, these variables can be used to develop GIS algorithms to detect travel modes and identify trip purposes from the gathered GPS traces. However, for it to be successful, a significant amount of post-processing must be conducted (Du and Aultman-Hall, 2007).

* Corresponding author.
E-mail addresses: qzchen@u.washington.edu (C. Chen), gong@hunter.cuny.edu (H. Gong), lawsonc@albany.edu (C. Lawson), evanb1313@gmail.com (E. Bialostozky).

[1] Computer-aided Telephone Interview.

0965-8564/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tra.2010.08.004

Preliminary studies show varying degrees of success. Wolf et al. (2001) demonstrated the possibility of detecting trip purposes in Atlanta, Georgia, given a detailed GIS database of land use. They found that mixed-use land use parcels posed a major challenge to accurately identify trip purposes. For example, of the 39 trips that were not identified, 26 of them were trips made to mixed-use parcels. In Canada, a study in Toronto suggested that 91.7% of the trip modes can be correctly detected (Chung and Shalaby, 2005), although in that study, trips with origins and destinations in urban canyon areas were dropped. Recent studies show varying success rates in detecting travel modes and identifying trip purposes (Bohte and Maat, 2008; Schuessler and Axhausen, 2009; Zheng et al., 2008).

In this paper, we present the development of a set of GIS-based algorithms to detect travel modes and trip purposes from GPS traces in the complex urban environment of New York City (NYC). Over 8.5 million individuals reside in the five boroughs of NYC, including Manhattan, Queens, Bronx, Brooklyn, and Staten Island. The urban geography of this area presents three major challenges for the implementation of GPS/GIS technologies. The first is the urban canyon effect. In parts of New York City (e.g., Wall Street and Midtown), narrow alleys delineated by tall buildings form urban canyons where GPS signal reception is distorted. The second is the presence of an extremely complex multi-modal transportation network (e.g., auto, subway, bus, rail, and walking), where the same street segments are used by different modes. This creates difficulties for mode detection, since GPS traces recorded on a street segment may be made by any one of the possible modes. The third is the dense and diverse land use patterns in the area. In 2008, Manhattan had an average population density of 71,201 persons per square mile and accommodated 2.3 million jobs. Staten Island, also a borough of New York City, had a population density of 8403 residents per square mile, almost ten times less than that of Manhattan. The diversity of land uses makes a single matching between trip purposes and land use types infeasible for most parts of the city. Buildings with multiple businesses and residential units are commonplace. The combined effect of the availability of multiple modes and the density and the diversity of the land uses in the area is that short trips and short stays occur with great frequency. Short trips can be made by different modes and connected by short stays at various activity nodes, located in mixed land use buildings.

We are interested in determining the feasibility of using a passive GPS travel survey methodology in a real-world implementation. If a passive GPS travel survey is feasible, it will result in objective measurements, instead of subjective ones from self-reported diaries, which have been used as ‘‘behavioral data” for decades. Analyses based on more objective measurements should more accurately reflect the activity and travel behavior in time and space. Equally important, the implementation of a passive GPS travel survey will only require the respondents to self-report socio-demographic information, thus dramatically reducing respondent burden. Consequently, sample size may be significantly enlarged with the same or even a reduced budget and the collection of continuous data streams may be possible. The substantial increase in data volume for longer periods will allow us to answer old questions with a new perspective and explore areas previously constrained by the data. For example, variability in activity and travel patterns in time and space can now be better captured and studied (Stopher et al., 2008a); behavioral changes in time and space may be better linked to changes in the urban environment to draw inferences on causality between the two; and over time, if data can be tracked for a long time (over a number of years), one can spatially and temporally sequence long term events (such as job and home relocations) and short-term events (daily activity and travel patterns) and understand their linkages (Ommeren et al., 1999). These new explorations will clearly have important implications in transportation policy making.

While a number of existing studies report satisfactory success rates in mode detection, most are set in an environment that does not possess the same level of complexity as NYC. Fewer number of studies reported success rates in identifying trip purposes; these rates are often much lower than those for mode detection. In particular, mixed-use land use is a major issue in trip purpose identification (Wolf et al., 2001). The land use pattern in NYC is characterized by the prevalence of mixed-use parcels. Given these challenges, the results of this study can serve as a benchmark in determining the feasibility of a passive travel survey in a complex urban environment.

The rest of the paper is organized as follows. In Section 2, we present a literature review discussing recent developments around the world on the use of GIS/GPS technology in travel surveys. In Section 3, we review the existing developments on travel mode detection and trip purpose identification. Our study effort of using GPS/GIS technologies is described in Section 4. The conclusion follows in Section 5.

2. GPS-based travel surveys

A GPS-based travel survey requires respondents to carry GPS loggers on one or more travel days. GPS loggers receive signals from satellites and record information in multiple dimensions, including location (longitude and latitude), speed, distance, and time. However, essential information in travel surveys, such as travel modes and trip purposes, cannot be directly obtained from GPS loggers. Thus, GPS-based travel surveys typically employ interactive or passive methods to collect these variables (Stopher et al., 2008b,c). In an interactive mode, respondents are prompted to answer questions before, during, or after a trip; in a passive mode, information on travel mode and trip purpose is inferred by the use of GIS algorithms. In recent years, there have been empirical studies pursuing both approaches (Auld et al., 2009; Battelle, 1997; Wolf et al., 2001). The interactive approach still relies on the respondent to some extent to obtain critical information on trip attributes (Auld et al., 2009). The passive approach completely eliminates respondent burden, yet its success rates in detecting certain modes, such as transit and identifying trip purposes are still unsatisfactory (Chung and Shalaby, 2005; Schonfelder and Samaga, 2003; Tsui and Shalaby, 2006).

Compared to conventional travel survey methods, such as CATI surveys that rely on respondents' self-reports, the use of GPS technology has many advantages. Chief among them is the reduction in respondent burden (Murakami et al., 2004; Wolf et al., 1999; Zhou and Golledge, 2007). Because of the much reduced respondent burden, the conventional 1-day or two-day travel survey can be extended to weeks or even longer, providing sufficient data to examine variability in travel patterns and potentially reduce overall sample size (Stopher et al., 2008a, 2007). The use of GPS technology also increases data accuracy. Wolf et al. (1999) reported that trip start and end times, as well as trip distances, in self-reported surveys tend to be rounded; short trips, especially those on foot, are easily omitted. These errors and omissions can be corrected with the use of GPS technology (Forrest and Pearson, 2005; Murakami et al., 2004; Stopher et al., 2007). In addition to travel surveys, GPS technology has also been applied to identify travelers' route choices (Duncan et al., 2007; Krizek et al., 2007) and to evaluate the variation in vehicle speeds (Murakami et al., 2004).

The existing literature has reported relatively few disadvantages regarding the deployment of GPS technology in travel surveys. Respondents' privacy concerns have been minimal (Swann and Stopher, 2008; Wolf et al., 2006). The main difficulty has been signal loss or degradation. The urban canyon effect, for example, occurs when the satellite signals received by a GPS logger do not emanate directly from the satellite source, but bounces off tall buildings surrounding the GPS logger. The urban canyon effect is particularly noticeable in densely built central business districts (CBDs), where the greatest travel data accuracy is often needed (Stopher and Greaves, 2007). Another issue is cold (or warm) start, in which the GPS logger, after being off (or underground) for an extended period of time, requires from 5 to 30 s more (depending on the type of the urban environment) (Lawson et al., 2007, 2008) to find enough satellite signals to accurately locate itself. In a complex urban canyon environment, the time required for a GPS logger to accurately locate itself may be even longer. Because a person's daily trips are sequenced, or the previous trip's destination is often the origin of the next trip, the information lost due to these issues may be recovered by examining the trips taken before and after (Wolf et al., 2003).

3. GPS/GIS technology on mode detection and trip purpose identification

The ability of a GPS logger to passively, continuously, and accurately record location and time information makes it uniquely capable of enhancing and/or replacing conventional self-reporting methods in travel surveys. The proof-of-concept experiment by Murakami et al. (1997) first established the feasibility of adopting the technology for travel survey purposes. Since 2000, substantial advances have been made to GPS/GIS technologies. GPS units are now lightweight and user-friendly (Stopher et al., 2005; Stopher and Greaves, 2007; Wolf et al., 2006).

The integration of GIS algorithms with GPS loggers has the potential of significantly reducing respondents' burden by passively collecting data and detecting travel modes and inferring trip purposes automatically. Existing studies have demonstrated some satisfactory success rates of achieving these objectives. Though they did not use GIS as an analysis tool, Schuessler and Axhausen demonstrated that reasonable determinations can be made for trip distance, duration, and mode simply based on the most basic GPS data, such as position coordinates and time, without any additional respondent-provided data or data about the validity of the GPS points (Schuessler and Axhausen, 2009).

Among the studies that employed GIS algorithms, an analysis of 60 trip records in Toronto correctly matched 78.5% of all traveled links to the road network and accurately determined the auto mode 91.7% of the time. Most of the missed links were the result of the scarcity of GPS data due to signal loss or degradation (Chung and Shalaby, 2005). Another study in Toronto used fuzzy logic-based mode identification algorithms both with and without additional GIS-based analysis (Tsui and Shalaby, 2006). The application of the GIS-based analysis improved the bus detection rate from 76% to 80%. In Netherlands, Bohte and Maat were able to correctly detect 70% of the trips using auto, walking, and bicycling, but their bus detection rate is substantially lower (Bohte and Maat, 2008).

Chung and Shalaby (2005), while finding their own results "satisfactory", identified a few areas in need of improvement in order to arrive at a point where travel surveys can be entirely GPS/ GIS-based. Specifically, GPS signal loss and degradation, mainly caused by urban canyon effects and cold starts, are the chief reasons for the inaccuracy of the GIS algorithms. They suggested testing the use of shortest path algorithms to fill in those segments with no GPS data. The other primary area of improvement is the development of additional and more complex GIS algorithms.

Stopher et al. conceptualized ways to "clean" GPS traces and to infer important trip related information (Stopher et al., 2005). To detect travel modes, they suggested the use of an elimination method, through which walking segments are first identified, followed by the identification of public transit and private auto. The rules developed are based on distinct attributes associated with unique modes; for example, public transit must travel on designated lines and there are periodic stops. Examples were shown to demonstrate the applicability of these rules.

The extant research on trip purpose identification is carried out via two approaches: location-only and location-based. The former uses the location reference data for visual inspection to infer trip purposes (Wolf et al., 2001) while the latter incorporates other supplementary trip related information collected through additional or existing travel surveys (Schonfelder and Samaga, 2003; Stopher et al., 2008a,b).

Wolf et al. initiated an effort on trip purpose detection. Their study used three major datasets: land use, road network, and aerial photo (Wolf et al., 2001). The land use data is the primary dataset and the other two are used for visual inspection. Given the data, the derivation process uses the following major steps: (i) identify trip ends, (ii) establish relationships between trip purpose and land use, (iii) assign a trip purpose to each trip end using the land use information

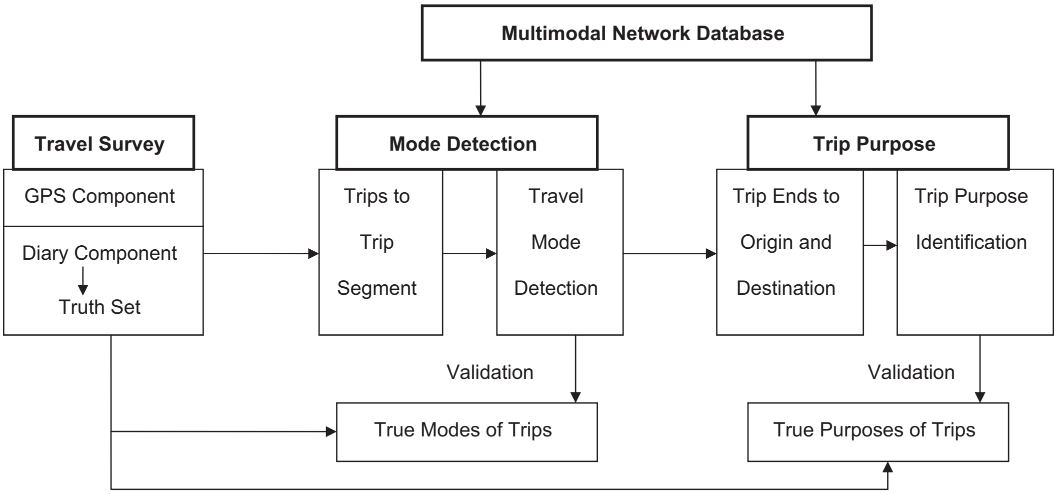


Fig. 1. Project components (travel survey, multi-modal network database, travel mode detection, and trip purpose identification).

of the destination, and (iv) estimate the most probable trip purpose for the rest of trip ends and clusters by visually inspecting the aerial photo and locating the closest intersection of the road network. Among the 39 trips that were miscoded, 26 were trips made to mixed-use parcels, suggesting that mixed-use land use is a major challenge in trip purpose identification.

Schonnefelder and Samaga used the same trip end identification process, with additional criteria[2] (Schonfelder and Samaga, 2003). To infer trip purposes, they used a multi-stage hierarchical matching procedure, calculated a cluster center of stop ends by combining trip ends, identified trips with obvious purposes,[3] and established relationships between trip purposes and the socio-demographics of the respondents as well as the temporal information of the activity. Since they did not have the information on the true trip purposes, the distribution of the inferred purposes was compared to that from a regional household travel survey. The results indicated differences in a number of purposes including private business, work and work related, shopping, and leisure activities.

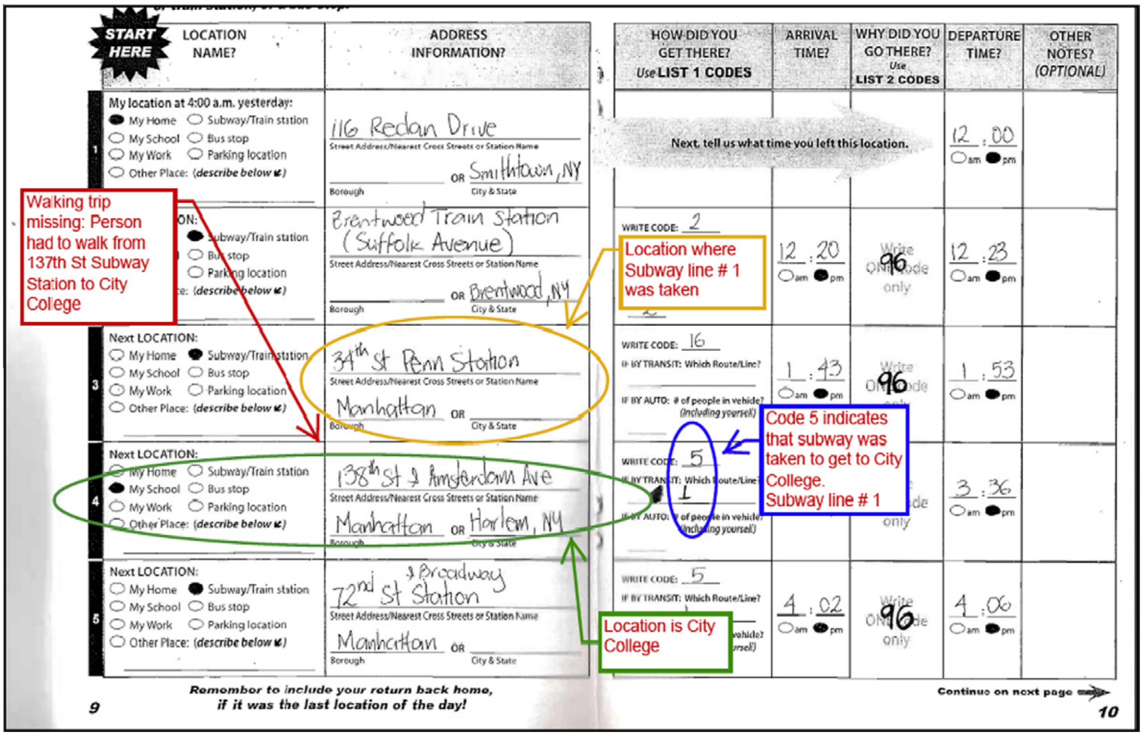
In a series of studies, Stopher et al. acquired additional information from survey respondents, most notably, a list of frequently visited locations (Stopher et al., 2008b,d). Using this information along with respondents' socio-demographics and land use information for the destination, they determined trip purposes by overlaying the trip layer (obtained from GPS traces) with the land use layer in GIS. To identify trip purpose, Stopher et al. divided trips into private-level destinations and public-level destinations. In addition, their study utilized two types of important information, which were collected during the recruitment stage: major activity location (MAL) such as home, workplace, school, and most frequently visited grocery stores; and occupation information. The trip purpose is identified within 200 m to the MAL. The MAL alone explains 75% of trip ends and over 60% of the trip purposes. To identify the rest of the trips, both trips and land use types are categorized into several groups by the desired level of details. Then, the trip purpose is deduced by heuristic rules given respondents' occupation information.

4. Study setup

This study demonstrates GIS-based algorithms used to automatically detect travel modes and infer trip purposes in the complex urban environment of New York City to reduce respondent burden in future household travel surveys. Fig. 1 illustrates the methodological flow process used in this study. There are four major components: the travel survey, a multi-modal network database, the detection of mode of travel, and trip purpose identification. The travel survey component produces the GPS data for the algorithms and the "truth set" designed to evaluate the performance of the developed algorithms. The multimodal network database contains the data necessary to represent the New York City environment and facilitates travel mode and trip purpose detection. The travel mode detection component is carried out prior to the trip purpose component. We describe each component in more detail in the rest of this section.

[2] Wolf et al. (2001) defined non-movement using the 120-s threshold. Schonfelder and Samaga, 2003 used several parameters, including distance, speed, duration, and activity duration.

[3] For example, a trip with an end in the parking lot of a shopping mall is likely a shopping trip. A trip with an end near a home location is likely a home-based trip. A work trip needs to meet the following conditions: (a) second most frequently visited after home; (b) trip occurring during a weekday; and (c) matching the structural and temporal characteristics of a work trip. Business/leisure trips can be matched to the nearest landmarks.



	START HERE LOCATION NAME?	ADDRESS INFORMATION?	HOW DID YOU GET THERE? Use LIST 1 CODES	ARRIVAL TIME?	WHY DID YOU GO THERE? Use LIST 2 CODES	DEPARTURE TIME?	OTHER NOTES? (OPTIONAL)
1	My location at 4:00 a.m. yesterday: ● My Home ○ Subway/Train station ○ My School ○ Bus stop ○ My Work ○ Parking location ○ Other Place: (describe below)	116 Redan Drive Street Address/Nearest Cross Streets or Station Name Borough OR Smithtown, NY City & State	Next, tell us what time you left this location.			12 : 00 ○ am ● pm	
	...ON: ● Subway/Train station ○ Bus stop ○ Parking location ...e: (describe below)	Brentwood Train Station (Suffolk Avenue) Street Address/Nearest Cross Streets or Station Name Borough OR Brentwood, NY City & State	WRITE CODE: 2	12 : 20 ○ am ● pm	Write ONE code only 96	12 : 23 ○ am ● pm	
3	Next LOCATION: ○ My Home ● Subway/Train station ○ My School ○ Bus stop ○ My Work ○ Parking location ○ Other Place: (describe below)	34th St Penn Station Street Address/Nearest Cross Streets or Station Name Manhattan Borough OR City & State	WRITE CODE: 16 IF BY TRANSIT: Which Route/Line? IF BY AUTO: # of people in vehicle? (Including yourself)	1 : 43 ○ am ● pm	Write ONE code only 96	1 : 53 ○ am ● pm	
4	Next LOCATION: ○ My Home ○ Subway/Train station ● My School ○ Bus stop ○ My Work ○ Parking location ○ Other Place: (describe below)	138th St & Amsterdam Ave Street Address/Nearest Cross Streets or Station Name Manhattan Borough OR Harlem, NY City & State	WRITE CODE: 5 IF BY TRANSIT: Which Route/Line? 1 IF BY AUTO: # of people in vehicle? (Including yourself)		...de only	3 : 36 ○ am ● pm	
5	Next LOCATION: ○ My Home ● Subway/Train station ○ My School ○ Bus stop ○ My Work ○ Parking location ○ Other Place: (describe below)	72nd St Station & Broadway Street Address/Nearest Cross Streets or Station Name Manhattan Borough OR City & State	WRITE CODE: 5 IF BY TRANSIT: Which Route/Line? ...vehicle? ...rself)	4 : 02 ○ am ● pm	Write ONE code only 96	4 : 06 ○ am ● pm	

9 Remember to include your return back home, if it was the last location of the day!

Continue on next page 10

Walking trip missing: Person had to walk from 137th St Subway Station to City College

Location where Subway line # 1 was taken

Code 5 indicates that subway was taken to get to City College. Subway line # 1

Location is City College

Fig. 2. A sample travel log.

4.1. Travel survey

For this project, we collected two datasets in New York City. In one dataset, 25 employees at the New York Metropolitan Transportation Council were recruited (NYMTC dataset) and asked to carry GPS units for one weekday; in the other dataset, 24 students and staff at the City University of New York were recruited (CUNY dataset) and asked to carry GPS units for five weekdays. Each GPS unit was configured to automatically log the person's position every 5 s, providing information on date and time, latitude, longitude, speed, etc. Each respondent was asked to turn on the GPS unit at the beginning of each day and carry the unit with them at all times. The GPS unit was only turned off at the end of the day after the person came home and didn't plan to go out again. In both datasets, respondents were also asked to fill out a paper travel diary for one designated travel survey day. Data collected in the travel log included each person's contact information, auto ownership, frequently visited locations, and activities and trips made during that day. For each activity and trip, detailed information was solicited, such as the name of the location (e.g., Broadway movie theatre), the exact address or intersection, the mode taken to access the activity site, departure and arrival times, and the purpose of the trip.

The primary purpose of requiring respondents to fill out a one-day travel diary is to generate a "truth set", containing the true travel mode taken for each trip, as well the trip purpose. If unprocessed, the paper diaries can contain errors. Fig. 2 illustrates the case where a walk trip from the 137th subway station to City College is missing. This required a close visual inspection of all completed diaries to identify any missed trips and illogical mode sequence and complete and correct them as needed. For each respondent in the two datasets, a true profile of his/her daily activity and travel pattern on the travel diary day was generated. Aggregating all profiles results in a truth dataset.[4]

4.2. Multi-modal network database

As noted earlier, creating a multi-modal network database for the study is critical because many trips made in New York City are multi-modal and the creation of a single multi-modal geodatabase facilitates faster processing on travel mode and trip purpose detection. We first obtain various geo-spatial layers representing different modes, including roadway networks, bus routes and stops, subway routes, subway entrances and exits, and commuter rail routes and stops. The use of subway entrances and exits instead of subway stops is particularly useful because a subway stop can have multiple entrance/exit stations that may be located several blocks apart; accurate subway entrance/exit station information helps significantly re-

[4] These manual efforts are to generate a truth set to test the performance of the developed algorithms. In future deployment of the developed and tested algorithms, such an effort is not necessary.

duce the urban canyon effect and sequence consecutive trip ends. Then, these various layers representing different modes must be spatially joined to accurately reflect the real-world situation. This process requires several edits. The "spatially cross but do not connect" problem refers to the case where segments that cross spatially, but do not connect (bridges or overpasses) are not automatically recognized by overlying multiple layers representing different modes together. An attribute table must be edited to differentiate these "spatially cross but do not connect" cases from crossings that actually connect. Sections of underground tracks on certain parts of rail and subway lines must be identified and defined in attribute tables. In merging, there are instances of overlapping segments and stops. For example, streets designated as bus routes will have the same segments on the roadway network file and on the bus route file. This requires a spatial join that would denote street segments that are also part of bus routes. These various edits produced a single multi-modal network database file that permits a quick determination of street connectivity during the GPS data analysis for travel mode and trip purpose detection purposes.

4.3. Travel mode detection

Shown in Fig. 1, travel mode detection consists of two critical procedures: one divides trips into trip segments and the other detects travel modes. The developed algorithm for travel mode detection identifies five modes: walk, car, subway, rail, and bus. We first define a "trip segment". A trip segment is a portion of a trip that is made by a single mode of transportation; a trip may consist of one or more trip segments.

4.3.1. Identifying trip segments

Prior to mode detection, we divide trips into trip segments by identifying potential mode change points. This involves two sequential steps. In the first step, we check whether a trip has any gap. Gaps imply underground trip segments. Wherever the time interval or the distance between consecutive points is greater than 120 s (seconds) or 250 m (meters), it is considered a gap. The time interval parameter is based on the work of Schuessler and Axhausen (2009), while the distance parameter used is increased from Chung's 150 m (Chung, 2003) to 250 m, to allow for greater urban canyon distortion in New York City. In the second step, we identify potential mode change points. We first distinguish between underground modes and on-street modes. The point before and after the marked underground gaps is a potential mode change location. This rule does not always apply as one may drive through a tunnel or ride in a subway line comprising both underground and above-ground tracks. In those cases, the modes before and after the marked underground gap are the same. This rule must be re-checked when the mode information is obtained. We then distinguish between on-street modes. We determine that any modal transfer must be marked by at least a short walk segment.[5] In the context of New York City, we define all walk segments to be longer than 60 s in time. Not all identified potential mode change points are true. After modes are detected, these rules are revisited and some potential mode change points will be dropped, if necessary.

4.3.2. Detecting travel modes

Once trip segments are identified, the developed algorithm attempts to detect the mode used for each segment, according to a set of criteria developed based on the characteristics of each mode. The algorithm first attempts to identify walk segments; walk segment identification is also essential in trip segmentation. Detection of the walking mode relies on three primary characteristics: (a) a pedestrian must walk on a pedestrian-accessible link and this requires map-matching between the walking path and the pedestrian-accessible street segment; (b) the travel time must be longer than 60 s (Schuessler and Axhausen, 2009); (c) the walking speed cannot exceed 10 km/h (Stopher et al., 2005).

After walk segments are identified, the algorithm proceeds to determine if any of the undetermined segments are made by subway or rail. This is relatively easy, as trips made by subway or rail exhibit a distinctive pattern compared to on-street modes, such as car and bus. Subway or rail trips must travel on designated links; a mode change must occur near a subway or a rail stop; and some parts of the trips may be underground and thus receive no GPS signals. There are gaps that do not follow the above-mentioned rules. Examples include the mode transfers between rail and subway at the Grand Central Terminal, where both rail and subway are underground. In these cases, it is hard to differentiate these two modes as no signals are received.

The primary criterion to distinguish between car and bus is that buses travel on bus routes and stop at bus stops. However, the latter criterion can be complicated by congestion and traffic signals, during which the bus will stop. In fact, these slow-moving and stopping bus segments are likely to be classified as walk segments in the earlier step. The algorithm addresses this problem by searching any three consecutive segments within one trip with the pattern of street-walk-street where the middle walk segment is less than 5 min[6] long and then combining them into one street mode (bus or car). This criterion is applied based on the fact that a trip made of car–short walk–car, bus–short walk–car, car–short walk–bus is rare.[7] Trips made of bus-short walk-bus are possible in New York City, indicating a transfer from one bus to another one.

[5] Similar rules were also used by Schuessler and Axhausen (2009) and Zheng et al. (2008).

[6] The 5-min threshold value was determined by visually examining GPS datasets from earlier field tests and by testing various values in the algorithm. The 5-min value was found to capture most car or bus slowdowns marked as walk while also not incorrectly eliminating actual walk segments.

[7] There are of course exceptions, where a person might drop off a car and have a short walk to a bus stop. Within the city limits, such trips are rare. In suburbs, such trips clearly exist and usually the bus ride is much longer than the car ride and the car is usually parked at a park and ride lot.

Table 1
Mode detection accuracy rates (sample size = 49 subjects).

	Identified as						Total	Success rate (%)
	Walk	Subway	Rail	Car	Bus	Und. T		
Walk	161	0	0	14	1	0	176	91.5
Subway	1	45	0	5	0	15	65	68.2
Rail	0	1	4	9	0	0	14	28.6
Car	0	0	0	23	1	0	24	95.8
Bus	4	1	0	9	16	0	26	53.3
Und. T[1]	0	0	0	1	0	5	6	83.3
								79.1 (mean)

4.3.3. Results

Table 1 reports the accuracy rates after applying the above-described algorithm to the 49 subjects in the sample. Among the five modes, car achieved the highest success rate, 95.8%. Trip segments made by walking also have a high success rate, 91.5%. Subway and bus, serving similar trips in terms of trip distance, receive similar success rates, 68.2% and 53.3% respectively. Rail achieved the lowest success rate, only 28.6%.[8] When an underground subway trip segment immediately follows or precedes an underground rail segment (e.g., at the Grand Central Station), no signal will be received during the entire course even though a mode transfer is made. These segments are designated as “Und. T”, indicating “underground transfer”. Using GPS traces collected near the subway and rail stations when a subject enters or exits from a station, the developed GPS algorithm is able to identify 83.3% of these underground transfer segments.

4.4. Trip purpose identification

Our approach to identify trip purposes from the GPS traces combines those of Schonfelder and Samaga (2003) and Stopher et al. (2008b,d). The approach has two steps: (1) clustering trip ends into origins and destinations and (2) identifying trip purposes. Before any purpose can be inferred, various trip ends must be clustered into distinct locations representing origins and destinations. This is necessary because trips with the same location may have different trip ends due to signal degradation or behavioral variations; for example, a building where an activity takes place can have more than one entrances and exits. Existing studies (Stopher et al., 2008b) have used a simple deterministic rule to cluster trip ends, for example, trip ends within a 200 m buffering zone are considered to have the same location. While this deterministic criterion is appropriate for some parts of the study area, other areas require shorter distance thresholds. Following Schonfelder and Samaga (2003), we apply the hierarchical clustering method to cluster trip ends into activity locations. Hierarchical clustering method differs from the popular K-means clustering method in the sense that it does not require the number of clusters as an input, but needs a termination condition. For our study, we use distance between consecutive trip ends as input.

Once the various trip ends are clustered into origins and destinations, we proceed to identify trip purposes, which can then be divided into two sequential sub-steps. In the first, GIS operations are carried out to overlay two GIS layers onto the point file representing trip origins and destinations. One GIS layer is the point-based file representing business listings and frequently visited locations (information supplied by the respondents). According to InfoUSA, there are more than 140,000 businesses in New York City and Manhattan alone has approximately 47,000 businesses. Another is the polygon file representing land use parcels in New York City. What is generated from this process is a polygon file containing information on land use, business listings, and frequently visited locations (if any) within an immediate area of each origin/destination of a trip. This first sub-step serves two purposes: one, we create a linkage between each origin/destination and the immediate built environment; and second, for a small number of trip ends located in low-density areas, a single deterministic matching between a trip origin/destination and a particular land use type may be possible. In some studies (e.g., Schonfelder and Samaga, 2003) conducted in an area where the density is relatively low, this deterministic direct matching was used.

In the second sub-step, for those trips whose purposes cannot be deterministically decided, we apply a probabilistic model to evaluate how various factors determine the purpose of a trip probabilistically. Three types of factors are considered: time of day, history dependence, and land use characteristics.[9] The time of day relevance in travel behavior was noted in several studies (Kitamura et al., 1997). History dependence refers to the notion that current behavior is dependent upon past behavior. Thus, if an activity has been performed previously, there may be a higher or a lower chance of performing it again. The use of land use characteristics recognizes the spatial clustering of certain types of activities: shopping is likely done in an area with many shopping opportunities and work is likely to occur where the employment density is high.

We estimated two multinomial logit (MNL) models to calculate the probability that a trip is serving a particular purpose for home-based trips and non-home-based trips. For both types, four trip purposes were considered: work/school related, personal business, shopping, and social recreation. In both models, it is assumed that if a specific trip purpose is selected,

[8] The reasons for the low success rate of rail segments are explained in Section 5.

[9] We cannot include variables representing frequently visited locations because their inclusion leads to the loss of about one third of the observations. However, in a larger sample, we expect these variables to exert powerful influences in predicting trip purposes, as shown in Stopher et al.’s studies.

Table 2
Multinomial logit model results for home-based trips.

Variables	Estimates	*t*-Ratios
Constant (social–recreational)	0.83	0.42
Constant (shopping)	2.85	1.50
Constant (personal business)	0.71	0.45
Time of day		
Dummy variable (9–10 am) (work)	1.02	0.84
Dummy variable (19–21 pm) (work)	3.59	2.3
Dummy variable (17–19 pm) (shopping)	1.05	1.03
Dummy variable (20–22 pm) (shopping)	4.39	2.18
History dependence		
# of trips ends within 250 m prior to the current trip on the survey day (work)	−0.02	−0.04
# of trips ends within 250 m prior to the current trip on the survey day (personal business)	0.33	0.71
# of trips ends within 250 m prior to the current trip on the survey day (social recreation)	0.36	1.46
Travel time from the previous location (min) (personal business)	−0.04	−2.05
Travel time from the previous location (min) (shopping)	−0.04	−2.25
Land use characteristics		
Percentage of business listings classified as work related within 250 m buffer (work)	7.51	1.98
Percentage of business listings classified as personal business related within 250 m buffer (personal business)	8.64	2.17
Log-likelihood (0): −72.08		
Log-likelihood (β): −34.83		

Table 3
Multinomial logit model results for non-home-based trips.

Variables	Estimates	*t*-Ratios
Constant (social–recreational)	2.54	2.11
Constant (shopping)	2.74	2.33
Constant (personal business)	1.13	1.60
Time of day and duration		
Dummy variable for (6–9 am) (work)	2.28	1.67
Dummy variable (9–11 am) (work)	1.66	1.39
Dummy variable (13–14 pm) (work)	0.98	1.16
Dummy variable (11 am–13 pm) (personal business)	0.56	1.05
Dummy variable (14–15 pm) (social recreation)	4.19	2.67
Dummy variable (21–22 pm) (social recreation)	3.18	2.06
Dummy variable (15–16 pm) (shopping)	1.99	1.52
# of minutes at the present location (work)	0.003	1.67
# of minutes at the present location (shopping)	−0.02	−2.37
History dependence		
# of trips ends within 50 m prior to the current trip on the survey day (work)	0.49	1.72
# of trips ends within 50 m prior to the current trip on the survey day (social recreation)	−0.33	−0.51
# of trips ends within 50 m prior to the current trip on the survey day (shopping)	0.09	0.18
Travel time from the previous location (min) (social recreation)	−0.007	−0.81
Travel time from the previous location (min) (shopping)	−0.008	−0.90
Land use characteristics		
Percentage of total square footage of office space within 250 m buffer (work)	15.63	4.59
Percentage of business listings classified as personal business related within 250 m buffer (social recreation)	71.15	2.49
Percentage of business listings classified as work related within 250 m buffer (personal business)	15.69	4.65
Percentage of business listings classified as personal business related within 250 m buffer (shopping)	14.74	3.34
Log-likelihood (0): −184.37		
Log-likelihood (β): −106.44		

it means the respondent derives a higher utility from conducting the trip with this purpose than with any of the other three purposes. Mathematically, it can be expressed as: $\Pr[i] = \Pr(U_i > U_j,\ i,j \in C,\ i \neq j]$, where U_i and U_j represent the utilities derived from conducting a trip with purpose *i* and *j*, respectively.

Tables 2 and 3 show the MNL results. In both cases, the estimated model significantly improved the baseline log-likelihood. The influences of the included variables are consistent with our expectations. Among the three categories of independent variables, time of day related variables and land use characteristics appear to exert a stronger influence in determining the purpose of the trip than the history dependence variables. The importance of time of day variables is consistent with findings from the literature (Kitamura et al., 1997), pointing to the regularity of human behavior patterns. The land use characteristics in the surrounding area of an identified trip location are also found to be prominent especially in the case of non-home-based trips. In model estimation, we experimented with land use variables using three buffers: 50, 150, and 250 m from the identified trip origins and destinations. We found that a radius of 250 m is the most significant. Variables related to history dependence are found to be mostly insignificant in both models, probably because the word "history" is currently calculated within a one-day time frame.[10] We expect their significance to increase when calculated in a longer-time frame.

5. Conclusion

Our study builds upon and expands the existing literature on the use of GPS/GIS technology in household travel surveys. In this study, we tested the feasibility of using a passive travel data collection methodology in a complex urban environment, using survey respondents wearing GPS loggers, and developing GIS-based algorithms to automatically detect travel modes and identify trip purposes.

The study is set in New York City. As mentioned earlier, the New York City environment presents several major challenges, including urban canyon effect, vast variability in density and diversity of the land use pattern in the area, and a complex, extensive transit network. The complexity of the transit network is not only attributed to the many modes in the region but also that a substantial portion of the city's subway and rail lines are directly above streets and thus GPS points obtained from those trip segments are spatially no different from the points that would have obtained if a trip is made by on-street modes such as walk, bus, or car.

We employed a number of methods to meet the challenges mentioned above. First, a spatial database representing the multi-modal transportation network of New York City was created. This geodatabase not only includes the line files of various modes in the city but also point files representing rail stations, subway entrances and exits, and bus stops. Second, we adopted an iterative approach incorporating multiple criteria in the algorithm development to accommodate and reduce urban canyon effects, delineate signal gaps and stops from trip segments, identify potential mode change points, and distinguish those trip segments that are spatially similar but represent different modes. We discuss several examples here. To identify gaps that potentially indicate an underground trip segment, we used a time interval threshold of 120 s and a distance limit of 250 m initially, followed by revisiting those identified trip segments later by checking the modes preceding and following the gap. In identifying walk segments, a similar iterative approach was used, where a simple speed-based rule was adopted first, followed by a revisit of these segments through linking GPS traces, mapping the links to the street network using the similarity index method (Chung, 2003), and developing a set of heuristic rules to identify walking segments. To differentiate those segments that are spatially clustered but represent different modes, multi-dimensional checks were applied, including checking the distances from trip ends to subway/rail stations and to subway/rail links, as well as examining the distance from every point of a trip segment to an above-ground subway/rail link. These methods have been proven successful. For instance, as shown in Table 1, the success rate for walking is 91.5% and the rates for subway, bus, and car range from 53% to 95%. The low success rate of rail is due to the fact that most of the trips by rail start or end outside of New York City and rail lines and stations outside of the city are clipped in order to define New York City as our study area. This problem can be corrected later by expanding the multi-modal network database to include surrounding areas. Third, we adopted the hierarchical matching clustering method, used by Schonfelder and Samaga (2003), to cluster trip ends into origins and destinations. And last, the density and diversity of the land use patterns in New York City deems the use of single and deterministic matching between trip end and land use type inappropriate in many parts of the area. Instead, probabilistic multinomial logit models were developed for home-based trips and non-home-trips to account for time of day effects (Kitamura et al., 1997), history dependence (Kitamura et al., 1997), and spatial land use effects. The results are encouraging, reporting 67% and 78% prediction rates for home-based trips and non-home-based trips, respectively.

Even though the algorithms and procedures developed in this study have achieved a reasonable level of success in detecting travel modes and identifying trip purposes in a complex urban environment, challenges remain. While some may be overcome by technological advances in the near future, others require more research identifying a promising set of procedures that can be applied in a real-world situation. In this study, we cleaned the raw GPS traces to remove and correct those distorted by the urban canyon effect. We expect that the urban canyon effect will be reduced in the future as better GPS chips are developed. The same applies to those signal gaps where GPS traces are completely missing (e.g., underground tunnel), as the dead-reckoning technology will eventually make its way to personal GPS loggers (Lawson et al., 2007). The main challenge in mode detection lies in the capability of correctly identifying transit modes. The current difficulty is attributed to, to some extent, limitations in transit data. The map-matching technology used in the existing studies (including this study) is limited to matching GPS traces against the static components of the transportation infrastructure such as bus stops. If matching can be done against real-time bus traveling routes, the success rate will likely significantly improve. Trip purpose identification will remain a challenge in a complex urban environment like NYC, where mixed-use parcels dominate. We show that probabilistic matching is a useful approach and a number of factors including time of day and land use play an important role. Although obstacles remain for a real-world application of a GPS-only travel survey in a complex urban environment, we remain hopeful that an implementation in the future is feasible.

Lastly, most of the existing studies, including this study, are empirical case studies. There is a need to conduct some uncertainty analyses to test the sensitivities of a set of parameters associated with the algorithm developed and the urban

[10] Even though the CCNY dataset has five-day data, the NYMTC has only one-day data. Thus, history related variables can only be calculated within a one-day frame.

environment. For example, Du and Aultman-Hall (2007), tested how dwell time, heading change, and distance between GPS points and the road network affected the accuracy of trip end identification. Referring to the previous discussion on the difficulties in identifying trip purposes, more work needs to be conducted in understanding how various parameters may affect the accuracy of trip purpose identification. These parameters relate to the categorization of the trip purposes to be identified, the complexity of the land use in terms of density and diversity, the size of the buffer used to capture the land use around identified origins and destinations, and the number of days with GPS traces available, etc. It is also possible that different methodologies (e.g., probabilistic models vs learning algorithms) used in trip purpose identification may perform differently under different scenarios. These issues will be the subject of our future work in this area.

Acknowledgement

The authors wish to thank Haiyun Lin, Jonathan Muckell, and Monica Rodriguez for their assistance in the project.

References

Auld, J., Williams, C., Mohammadian, K., Nelson, P., 2009. An automated GPS-based prompted recall survey with learning algorithms. Transportation Letters: The International Journal of Transportation Research 1, 59–79.

Battelle, 1997. Global Positioning Systems for Personal Travel Surveys: Lexington Area Travel Data Collection Test. A Report to Office of Highway Information Management (HPM-40), Columbus, Ohio.

Bohte, W., Maat, K., 2008. Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. In: 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability, Annecy, France.

Chung, E., 2003. Development of a Trip Reconstruction Tool for GPS-based Personal Travel Surveys. Master's. University of Toronto.

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

Du, J., Aultman-Hall, L., 2007. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues. Transportation Research Part A 41, 220–232.

Duncan, M.J., Mummery, W.K., Dascombe, B.J., 2007. Utility of global positioning system to measure active transport in urban areas. Medince and Science in Sports and Exercise 39 (10), 1851–1857.

Forrest, T.L., Pearson, D.F., 2005. Comparison of trip determination methods in household travel surveys enhanced by a Global Positioning System. Transportation Research Record 1917, 63–71.

Kitamura, R., Chen, C., Pendyala, R., 1997. Generation of synthetic activity-travel patterns. Transportation Research Record 1607, 154–162.

Krizek, K., El-Geneidy, A., Thompson, A., 2007. A detailed analysis of how an urban trail system affects cyclists' travel. Transportation 34, 611–624.

Lawson, C., Chen, C., Gong, H., Karthikeyan, S., Kornhauser, A., 2007. GPS Pilot Project. Phase One: Literature and Product Review. New York Metropolitan Transportation Council, New York City.

Lawson, C., Chen, C., Gong, H., Karthikeyan, S., Kornhauser, A., 2008. GPS Pilot Project. Phase Two: GPS Unit Comparison, Field Tests, and Market Analysis. New York Metropolitan Transportation Council.

Miller, E., Hunt, J.D., Abraham, J.E., Salvini, P.A., 2004. Microsimulating urban systems. Computers, Environment and Urban Systems 28 (1–2), 9–44.

Murakami, E., 2008. Hard to Reach Populations (Presentation at the NYMTC Survey Workshop). New York.

Murakami, E., Wagner, D.P., Neumeister, D.M., 1997. Using global positioning systems and personal digital assistants for personal travel surveys in the United States. In: Transport Surveys, Raising the Standard, Grainau, Germany.

Murakami, E., Taylor, S., Wolf, J., Slavin, H., Winick, B., 2004. GPS applications in transportation planning and modeling. The Travel Model Improvement Program Connection Newsletter, 1–3.

Ommeren, J.v., Rietveld, P., Nijkamp, P., 1999. Job moving, residential moving, and commuting: a search perspective. Journal of Urban Economics 46, 230–253.

Pearson, D., 2004. A comparison of trip determination methods in GPS-enhanced household travel surveys. In: 84th Annual Meeting of the Transportation Research Board, Washington, DC.

Pendyala, R., Kitamura, R., Chen, C., Pas, E., 1997. An activity-based microsimulation analysis of transportation control measures. Transport Policy 4 (3), 183–192.

Schonfelder, S., Samaga, U., 2003. Where do you want to go today? More observations on daily mobility. In: 3rd Swiss Transport Research Conference, Monte Verita/Ascona.

Schuessler, N., Axhausen, K.W., 2009. Processing raw data from Global Positioning Systems without additional information. Transportation Research Record 2105, 28–36.

Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: where are we going? Transportation Research Part A 41, 367–381.

Stopher, P., Jiang, Q., FitzGerald, C., 2005. Processing GPS data from travel surveys. In: 28th Australasian Transport Research Forum, Sydney, Australia.

Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the Sydney household travel survey with GPS. Transportation 34, 723–741.

Stopher, P., Clifford, E., Montes, M., 2008a. Variability of travel over multiple days: analysis of three panel waves. In: 87th Annual Meeting of the Transportation Research Board, Washington, DC.

Stopher, P., Clifford, E., Zhang, J., FitzGerald, C., 2008b. Deducing Mode and Purpose from GPS data. Working Paper of the Austrian Key Centre in Transport and Logistics. University of Sydney, Sydney, Australia.

Stopher, P., Kockelman, K., Greaves, S.P., Clifford, E., 2008c. Reducing burden and sample sizes in multi-day household travel surveys. In: 87th Annual Meeting of the Transportation Research Board, Washington, DC.

Stopher, P., FitzGerald, C., Zhang, J., 2008d. Search for a Global Positioning System device to measure personal travel. Transportation Research Part C 16 (3), 350–369.

Swann, N., Stopher, P., 2008. Evaluation of a GPS survey by means of focus groups. In: 87th Annual Meeting of the Transportation Research Board, Washington, DC.

Tsui, S.Y.A., Shalaby, A.S., 2006. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transportation Research Record 1972, 38–45.

Wilson, J., 2004. Measuring Personal Travel and Goods Movement (Transportation Research Board Special Report 277). Transportation Research Board, Washington, DC.

Wolf, J., Hallmark, S., Oliveira, M., Guensler, R., Sarasua, W., 1999. Accuracy issues with route choice data collection by using Global Positioning System. Transportation Research Record 1660, 66–74.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: 80th Annual Meeting of the Transportation Research Board, Washington, DC, p. 24.

Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of underreporting on mileage and travel time estimates – results from Global Positioning System-enhanced household travel survey. Transportation Research Record 1854, 189–198.

Wolf, J., Bonsall, P., Oliveira, M., Leary, L., Lee, M., 2006. Review of the Potential Role of ''New Technologies'' in the National Travel Survey. Department of Transport, London.

Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.-Y., 2008. Understanding Mobility Based on GPS Data, Ubicomp, Seoul, Korea.

Zhou, J., Golledge, R., 2007. Real-time tracking of activity scheduling/schedule execution within a unified data collection framework. Transportation Research Part A 41, 444–463.