Methodology for Converting GPS Navigational Streams to the Travel-Diary Data Format

Sivaramnakrishnan Srinivasan(*)
University of Florida

Stacey Bricka
NuStats

Chandra Bhat
University of Texas at Austin

(*) Corresponding author

August 2009

ABSTRACT

Conventional travel-survey methodologies require the collection of detailed activity-travel information, which imposes a significant burden on respondents, thereby adversely impacting the quality and quantity of data obtained. Advances in the global positioning system (GPS) technology have provided transportation planners with an alternative and powerful tool for more accurate travel-data collection with minimal user burden. The data recorded by GPS devices, however, do not directly yield travel information; the navigational streams recorded by GPS devices have to be processed and the travel patterns derived from them.

This paper investigates key issues important in the development of a trip-detection algorithm to automate the processing of raw GPS data and to generate outputs of activity-travel patterns in the conventional travel-diary format and to identify trips and characterize them by several attributes (trip-end locations, trip purpose, time of day, distance, and speed). The information is used to develop an algorithm to identify trips and characterize them by specific attributes, which are then validated.

1. Introduction

Data on household travel patterns constitute a fundamental input to travel demand model development for use in transportation planning and policy analysis. The conventional approach for collecting such data typically employs a Computer Assisted Telephone Interview (CATI) technique to obtain self-reports of activity-travel patterns. Past studies indicate that the potential difficulty of respondents to comprehend survey diary questions, and the recall and reporting limitations of respondents, can critically degrade the quality and quantity of information from such self-reported activity/travel patterns. This is primarily because the respondents need to expend considerable time and effort in recalling and reporting detailed travel information. Significant advances in survey design methods and the effective application of CATI software to minimize data errors have mitigated concerns to some extent, but come at increased cost and do not resolve away issues associated with self-reported accuracy concerns.

In the above context, the Global Positioning Systems (GPS) technology offers a valuable opportunity to use in the place of, or in concert with, conventional data-collection approaches. Devices called the "GPS receivers", positioned anywhere on the earth's surface and in view of the GPS satellites, are capable of self-determining their locations with a time-of-day stamp (Wolf, 2004). Therefore, travel data can be collected by equipping the respondents' (and/or their automobiles) with GPS receivers and recording the position and velocity of the vehicles periodically. However, the data recorded by GPS devices do not directly yield travel information; rather, the outputs from these devices are in the form of navigational streams that have to be processed to derive travel information. Therefore, the success of this new technology as a travel survey instrument depends on the ability of the analyst to derive meaningful trip information from the navigational data streams of GPS devices. The scope of GPS-based travel surveys have been increasing from a few vehicles to thousands of vehicles, from one-day to multiday data collection, and from in-vehicle only to all travel modes (see for example, Stopher *et al.*, 2008), Further, data from GPS surveys are being increasingly used to validate the accuracy of self-reported surveys (see for example, Stopher *et al.* 2007). Overall, there is a growing interest in the field to develop robust and efficient algorithms and software for processing and analyzing GPS data streams. The broad intent of this paper is to contribute towards this end.

The specific focus of this paper is to describe and implement a procedure to automate the processing of raw GPS data to generate the activity-travel patterns in the conventional travel-diary format. The algorithm developed was implemented in prototype software that identified trips and characterized them by several attributes including trip-end locations, trip purpose, time of day, distance, and speed.

The rest of this paper is organized as follows. In Section 2, a synthesis of literature on alternate approaches and the issues involved in the processing of GPS navigational streams is presented and discussed. Section 3 discusses the travel-diary extraction methodology implemented in the GPS-TDG (GPS-based Travel Diary Generator). The paper ends with an overall summary and identifies areas of future research.

2. Processing GPS Navigational streams

This section of the paper begins (Section 2.1) with a description of the structure of the GPS navigational streams. Subsequently, the issues and methods in the processing of the GPS data to identify the travel-elements of interest are discussed. There are two major components in this travel-diary extraction process: (1) Trip (or Stop) Detection and (2) Trip Characterization (i.e., determination of trip timing, trip-end locations, etc.). These two components are inherently

interrelated. The characteristics of the identified trips might provide clues to the possible existence of other trips, which also need to be flagged as part of the trip detection process. Alternatively, attempts to characterize a detected trip may suggest that the trip is infeasible and was falsely detected by the previous trip-identification algorithms. Although the interactive nature of these two steps is recognized, for ease of presentation, the trip detection and trip characterization methods are discussed separately in Sections 2.2 and Section 2.3 respectively.

2.1 Structure of the GPS Navigational Streams

Most GPS receivers' output conforms to the National Marine Electronics Association's "NMEA 0183 GPS" message formats (Wolf, 2004). The outputs are in the form of "sentences", which comprise a number of predefined data fields separated by commas. The NMEA has prescribed the standard specifications for many different sentences types, with each sentence type providing different kinds of data. Of these, the "GPRMC" sentence contains the necessary position, velocity, and time information required for deriving the travel attributes, and hence, is most relevant for our purposes (Wolf, 2004). The position information is recorded in terms of latitude and longitude in fields 3 through 6. Velocity is recorded in fields 7 and 8. Field 7 records the speed in knots (1 knot = 1.5 mph), and the next field contains the direction of movement in degrees. The date and time are recorded as the Coordinated Universal Time (UTC) or the Greenwich Mean Time (GMT) in fields 9 and 1 respectively. Finally, Field 2 of the GPRMC sentence provides this information on the reliability and accuracy of the data recorded. Specifically, this field can either take a value "A" indicating a valid recording or "V" indicating a navigational receiver warning. The warning is recorded when the GPS receiver looses "sight" of the adequate number of satellites required for accurate positioning. This is likely to happen in areas with tall buildings ("urban canyons") or with extensive tree canopies.

In the application of GPS for travel surveys, the output from the receiver as described above is recorded or logged periodically. Thus, the GPS navigational stream can be defined in general as a sequence of periodically-recorded sentences. The structure of the overall GPS navigational stream is therefore dependent on the data-logging mechanism, which in turn, depends on three important aspects: (1) driver involvement, (2) data logging rules, and (3) the power-systems of the equipped vehicles.

Based on driver involvement, the data logging mechanism can be classified as either "user-flagged" or "purely passive". In "user-flagged" systems, the driver explicitly "flags" the start and end of each trip by turning the recording device on and off, respectively. In such systems, the GPS sentences are logged only during the trip and not when the vehicle is at a stop. In "purely-passive" systems, the data are recorded as long as the GPS receiver/antenna is powered on (the recording device in this case is always powered on via an internal battery). Hence, in contrast to user-flagged systems, purely-passive systems could also be recording points when the vehicle is at a stop.

There are two types of data logging rules. In the "frequency-based" logging approach, all valid data (GPS sentences) are recorded at the preset frequency (e.g., every 1 second or 5 seconds) and as long as the GPS receiver/antenna is powered on, irrespective of whether the vehicle is moving or not. In the "speed-checked" logging approach, data are recorded only when movement is detected (for example, if the speed is greater than 1 mph). Recording data only when motion is detected increases the storage capacity.

The data logging mechanism can also be affected by the power-system characteristics of the equipped vehicles. The GPS receiver/antenna unit is typically powered by the vehicle's

power system using the cigarette lighter adapter in the vehicles. Wolf, 2000 and Bachu *et al.*, 2001, have found that, in some vehicles, the power to the cigarette lighter remains on even if the vehicle is powered off. Correspondingly, there can be two data logging systems: (1) the "continuous-power" system, in which the cigarette lighter is always powered on and hence the GPS receiver/antenna unit is also continuously powered on, and (2) the "switched- power" system, in which the GPS receiver/antenna is powered on and off by powering the ignition on and off, respectively.

In summary, the navigational streams obtained from GPS-based travel surveys are a sequence of periodically-recorded sentences. While the structure of this sentence is fixed, the sequence of sentences recorded may or may not include the times when the vehicle is not traveling. This has strong implications for the trip detection procedure discussed next.

2.2 Trip (Stop) Detection

The first component in the travel-diary extraction process is trip detection (or equivalently, stop detection). The central idea to the identification of trips (stops) from GPS navigational streams is the detection of the nonmovement of the vehicle. If the duration of nonmovement exceeds a certain threshold, called the "dwell-time threshold", the presence of a stop and a corresponding trip is inferred.

2.2.1 Dwell-Time Threshold

The dwell-time threshold should be chosen to identify even short-duration stops (for example, stops for pick-up or drop-off), while at the same time guarding against detection of false stops (*e.g.*, waiting at stoplights or congestion delays) (Wolf, 2000). For most urban areas, the use of 120 seconds as the dwell-time threshold is a reasonable rule for signaling a (potential) stop (Stopher, 2004). However, dwell times of less than the threshold duration of 120 seconds could be quick stops for purposes such as pick-up or drop-off of passengers, which would be missed with a strict dwell time threshold for trip detection. To address these issues, the Trip Identification and Analysis System (TIAS), propriety software developed by GeoStats (see Axhausen *et al.*, 2004) uses three thresholds in its preliminary trip detection procedure. Specifically, the trips are classified as "confident" if the dwell times exceed 5 minutes, "probable" if the dwell time is between 2 and 5 minutes, and "suspicious delays" if the dwell time is between 20 seconds and 2 minutes. The "probable" and "suspicious delay" trip ends are subject to subsequent scrutiny based on the trip characteristics before being ultimately classified as a trip or not. The trip detection procedure developed by Stopher and colleagues (see Stopher *et al.*, 2002) uses two thresholds; dwell times of 30 to 120 seconds due to engine turn-off are classified as "potential trip ends" and dwell times of greater than 120 seconds are designated as "trip ends." Again, as in the case of the TIAS approach, the "potential trip ends" are subject to further scrutiny.

2.2.2 Identification of Nonmovement

As described in section 2.1, the data-logging mechanism determines whether GPS sentences are recorded even during periods of nonmovement of the vehicle. We define the "power-off" case as corresponding to situations in which the GPS sentences are not recorded during periods of nonmovememt and the "power-on" case as corresponding to situations in which the GPS sentences are recorded even during periods of nonmovememt. In the "power-off" case (switched-power system, or when the driver involvement is user-flagged, or when speed check rules are used for data logging), the data recording stops when the vehicle is not moving. In these

situations, extended periods of nonmovement are necessarily represented by time breaks in the record streams. Therefore, nonmovement for long periods of time can be determined by simply looking for gaps in the time stamps between successive records (Wolf, 2000).

In contrast, in the "power-on" case (a purely-passive data logger without any speed check rules is used in a continuously powered system), the data points are being continuously logged, even when the vehicle is at a stop and is powered off. Therefore, nonmovement cannot be detected based on time breaks in the stream. Similarly, the logic of looking for gaps in the time stamps of the successive recordings cannot be applied in switched-power data collection protocols with frequency-based logging rules to identify stops when the engine is not powered off (these could be short-duration stops at drive-through or for pick-up/drop off). In these cases, nonmovements have to be detected by explicitly examining the recorded position and speed data. Specifically, the detection of stops/trip-ends involves identifying a sequence of data records over a certain period of time during which there is little change in the position of the vehicle and the speed is almost zero. The following approach suggested by Stopher *et al.* (2002, 2008) can be used as the implementation logic: If the difference in successive latitude and longitude values is less than 0.000051 degrees (about 7.4 meters), the heading is unchanged or zero, and the speed is zero for 2 minutes or more, then nonmovement is inferred.

2.2.3 Use of Supplemental Data

The above discussions have focused on using solely the GPS data for trip detection. In this context, prior research has been largely successful in developing algorithms to identify stops of durations greater then a certain minimum dwell-time threshold (often 2 minutes). Stops of very short durations, however, are more difficult to identify, particularly when the vehicle is not powered off at the stop. Further, using only the GPS streams, it is not possible to guarantee that all trip ends identified are true stops (rather than congestion delays or wait times at the traffic signals). Supplemental data on transportation network characteristics can be used to alleviate these concerns and enhance trip detection by minimizing the number of missed trips and false trips. The TIAS software uses a GIS road network layer for trip-detection enhancement in two ways (Axhausen *et al.*, 2004): First, "probable trip ends" and "suspicious delay" points identified from the preliminary trip-detection procedures are overlaid on the GIS road network, and those that fall within the last 1/3 of a road segment upstream of an intersection are classified as congestion delay and not a trip end. Second, the software examines the travel paths for overlaps (*i.e.,* loops in the travel path) and "circuity" (extent of directional change). Points classified as "suspicious delay" from the preliminary analysis are reclassified as "trip ends" if they fall strategically on a path with high circuity or overlaps.

2.2.4 Trip Detection During Periods of Signal Loss

Finally, the methods described above cannot be applied to scenarios in which stops occur during a period of signal loss (for example, if the stop is the downtown area). The following methodology (Stopher *et al.* 2002) deals with such situations:

1. The average speeds immediately before and after the period of signal loss is determined using the last 10 track points before and the first 10 track points after the period of signal loss.

2. The estimated speed during the period of signal loss is determined using the straight-line distance between the locations of signal loss and signal reacquisition and the time period of signal loss.
3. If the estimated speed is considerably lower than the average speeds before and after the signal loss period, a potential stop is inferred.
4. If a potential stop is inferred, the expected time to traverse the signal-loss distance at the average speeds prior to the period of signal loss is computed. This is subtracted from the time period of signal loss to obtain an estimate of the stop duration. If this stop duration is greater than 120 seconds, a stop is inferred; otherwise no stop is assumed to have occurred.

More recently (Stopher *et al*. 2008) algorithms that take advantage of a GIS roadway network layer to detect stops during periods of signal loss have also been developed.

2.2.4 Refining Trip Detection Based on Trip Characteristics

As already indicated, the characteristics of the identified trips might provide clues to the possible existence of other trips missed by the trip detection process. For example, if the origin and destination locations of a trip are the same, this suggests the possibility of a missed stop (although this could also be indicative of a pure-recreation or an abandoned trip, *i.e.*, a round trip with no apparent purpose; see Axhausen *et al.*, 2004). In such a scenario, one could examine the specific trip further to determine if there was a missed stop. Another possible approach would be to examine if there is a reversal in direction of the vehicle along this trip. Stopher *et al.* (2002) provide an implementation definition of reversal as a change in heading between 178 and 182 degrees within 30 seconds. It is also possible that attempts to characterize a trip may suggest an infeasible trip that has been falsely detected by the previous trip identification algorithms. Axhausen *et al.* (2004) reclassify a trip end as erroneous if the trip duration is less than 30 seconds, the average trip speed is greater than 50 kmph (31 mph), or the trip distance is greater than 25 kilometers (15.5 miles). Similarly, the SCAG vehicle activity study (Stiefer *et al.*, 2003) required further examination of trips of duration less than 1 minute or greater than 1 hour and with average speeds less than 5 mph. or greater than 60 mph.

2.3 Trip Characterization

The second component in the overall trip diary generation procedure involves the characterization of the identified trips and stops. The trip attributes that may be derived from the GPS navigational streams and other supplemental data are the geographic location of the trip ends, trip timing, trip distance and speed, activity/trip purpose, and route. (Since the focus of this paper is on in-vehicle GPS surveys, the mode of the trip is known).

2.3.1 Trip-end Locations

Origin and destination trip-end locations are determined by reading the location information (latitude and longitude) from the first and last records of the GPS navigational stream corresponding to the trip. However, when switched-power systems are used, the first valid point recorded may not be the starting point of the trip due to the time required by the GPS device to acquire a signal. It has been found that, in situations in which the vehicle is driven almost immediately after ignition-on, it may take anywhere between 15 seconds to 4–5 minutes for signal acquisition, depending on speed of movement, the duration for which the GPS receiver was powered off before reactivation (warm versus cold starts) and other extraneous factors, such

as the presence of tree canopies and tall buildings (Stopher, 2004). A signal-reception distance analysis undertaken by Bachu *et al.*, (2001) indicates that the average distance traveled by a vehicle before the signal is first acquired is about 0.166 miles (the median value is 0.11 miles). However, this problem can be remedied by assuming that the origin location of a trip is the same as the destination location of the previous trip (Schonfelder *et al.*, 2002).

When multiday travel data are analyzed, it is possible that the recoded coordinates of the trip ends are found to be different even if the actual trip destinations are the same. This could be due to different parking spots and/or inherent randomness in the GPS position determination. Schonfelder and Samaga (2003) have developed an algorithm to identify the main destination locations from a clustered set of trip-end recordings. In this procedure, for each trip-end location, the distance to all other trip-end locations within a radius of 200 meters was computed. Those trip ends that have the most neighbors and the smallest average distance to these neighbors (*i.e.*, the cluster centers) are classified as unique destination locations. For the remaining, non-central, trip ends (*i.e.*, those trip ends that are not classified as a unique destination location), the nearest cluster center is assigned as the destination location.

Subsequent to the determination of the trip-end locations in terms of the latitude and longitude, the likely land use parcel associated with the trip end and, hence, the address of the trip end can be determined using suitable GIS data and spatial overlay procedures. With the improvements in GIS files and software, this approach has become more commonplace (NuStats/GeoStats 2008).

2.3.2 Trip Timing

The trip start time is primarily determined based on when the GPS device acquires its first fix (*i.e.*, the time stamp on the first valid record for the trip). Similarly, the trip-end time is the time stamp on the last valid position assumed to be the end of the trip. Consequently, the determination of the correct trip start times can be impacted by the signal acquisition time, if switched power systems are used. Further, if there is a loss of fix at the end of the trip (e.g., driving into a parking garage or parking off-site), the recorded trip end may not be the true trip end. As a result of these issues, the recorded vehicle trip time can be expected to be systematically less than the actual trip time (and the reported person trip time), with the discrepancy being between several seconds to several minutes (Murakami and Wagner, 1999, Stopher *et al.* (2002)).

2.3.3 Trip Distance

There are two main approaches to determining trip distances from the GPS data (Battelle, 1997). These are: (1) the point-to-point sum of distances (PP) over the entire trip and (2) the link-to-link sum of distances (LL) over the entire trip after matching the GPS points to network links.

The first method, *i.e.*, the point-to-point sum of distances, involves the computation of the distance between successive pairs of recorded locations. These pair-wise distances are then summed over the entire trip to determine the trip length. The computation of the distance between successive points may be accomplished using either the latitude and longitude information for the two points (the formula to calculate this distance is provided by Wolf *et al.*, 2003) or as a product of the recorded instantaneous speed and the time gap between the successive data recordings (Wolf, 2000). The primary advantage of the PP approach is that the trip distance is determined without the use of any secondary data (as would be necessary in the LL approach). However, it has been found that the PP approach could result in the

overestimation of trip distances, especially when the position information is used for computing the distance between successive points in a trip. Specifically, the positional errors associated with each data record could add up, leading to overestimation of the trip length. The magnitude of this error can be particularly large for trip segments through urban canyons, where multipath errors and satellite line-of-sight issues can significantly deteriorate the positional accuracy of the GPS points (Wagner *et al.*, 1996). In this context, it has been suggested (see TRB NCHRP Synthesis, 2001) that the use of positional data recorded every 10 seconds instead of using the data recorded every second (which is the typical recording frequency) can help reduce the overestimation error by almost 50 percent.

The second method for trip distance computation, *i.e.*, the link-to-link sum over the entire trip (the LL approach), requires that the GPS traces be matched to an underlying road network to identify the actual links traveled by the vehicle. The trip length is determined as the sum of the length of all the roadway links traveled. The advantage of this approach lies in its ability to accommodate signal loss. The accuracy of such an approach depends on the quality and quantity of valid GPS points available for identifying the network links used (Murakami and Wagner, 1999).

2.3.4 Trip Purpose

The identification of activity/trip purpose is perhaps the most challenging of all GPS data processing tasks. The first step in this direction appears to have been taken by Wolf (2000) in her dissertation research. In this work, she proposed to use land use information at the trip end as the primary means to identify trip purpose. Specifically, this approach involves a "point-in-polygon" analysis to first match the trip end location (a point) to a polygon-based land use inventory to determine the land use type at the trip end. Further, each land use type was associated with a primary trip purpose and, whenever possible, secondary and tertiary trip purposes were also identified. The study employed 25 land use type categories and 11 trip-purpose categories. The land use at the trip-end location along with the time-of-day of travel and activity duration at the stop was used to *manually* assign trip purposes. The major problem encountered during this step was that it was not possible to associate certain land use categories (such as mixed-use land parcels and vacant lots) with a specific trip purpose. Further, the success of this methodology requires a very detailed land use GIS database at a fine spatial resolution.

The Swiss researchers (see Axhausen *et al.*, 2004; Schonfelder and Samaga, 2003) have developed a comprehensive approach for trip purpose identification in the context of multiday travel data collection. These researchers used data on the demographic characteristics of the survey respondents, facility-location data, land use patterns, and national travel patterns to develop a probabilistic approach to trip purpose determination. The overall methodology is:

1. For trip end destinations that are within 200 meters of the household location, the trip purpose is "home".
2. For full-time workers, the trip purpose is "work" if (a) the destination location is the second most frequented of all, (b) the structural and temporal characteristics of the stop are consistent with those determined from the national travel surveys for the work purpose, and (c) the record is for a weekday.
3. For the trip destinations not classified as either home or work, "most probable" trip purposes are determined in three different ways:

a. For each trip destination, all points of interest within a catchment area of 300 meters are identified. Each POI is assigned an *apriori* probability of being associated with a trip purpose. The probability of each trip purpose is determined as the weighted sum of the individual trip-purpose probabilities associated with each of the POIs within the catchment area. (POIs closer to the trip destination have a higher weight.) The most probable trip purpose is determined.
b. The land use patterns within 200 meters of each trip destination are examined. Each land use class is assigned an *apriori* probability of being associated with a trip purpose. The trip-purpose probabilities of all the distinct land use classes found within the buffer zone are examined to identify the most probable trip purpose.
c. A third "most probable" trip purpose is determined using driver characteristics (gender, automobile availability, and employment status) and temporal characteristics of the stop (e.g., day of the week, activity start time, and activity duration). The national travel characteristics are used to develop rules of association between the driver characteristics, the temporal characteristics of the stop, and the trip purpose.

4. The final trip-purpose assignment is accomplished using the three probable trip purposes:
   a. If all the three approaches yield the same result for the most probable trip purpose, then the agreed purpose is assigned.
   b. In case of any mismatch, the POI/land use categorization is preferred, except when the trip purpose determined from the third method (*i.e.*, using demographic characteristics of the driver and the structural characteristics of the stop) is "pick-up and drop-off", in which case, this is the assigned trip purpose.
   c. If there is no clear POI/land use assignment possible, the categorization from the third method is used to determine the trip purpose.

Stopher *et al.* (2008) also describe additional heuristic rules for determining the purpose of trip-ends which are neither home nor work. Again these methods rely on trip characteristics and known land use patterns.

2.3.5 Trip Route

The trip detection algorithms discussed above identify stops. The stream of GPS data records between successive stops describes the path of movement during the trip. Hence, the trip route can be identified using map-matching procedures *i.e.,* matching the GPS data points to appropriate links on an underlying GIS roadway network map. It is important to note that this matching is not trivial, as both the GPS data and the digital roadway-network data have different levels of spatial accuracy and inherent errors. Consequently, the development of map-matching algorithms is in itself a very vast and complex field of study. Researchers have developed a wide array of methods using deterministic, probabilistic, and fuzzy-logic-based approaches (see for example, TRB NCHRP Synthesis, 2001) for matching GPS traces to GIS maps. The reader is referred to the following for some recent contributions in this area and further references: Chung and Shalaby (2004), Greenfeld (2002), TRB NCHRP Synthesis (2001), and Doherty *et al.* (1999).

3. The GPS-TDG Travel Diary Extraction Algorithm

This section describes an algorithm (called GPS-TDG for GPS-based Travel-Diary Generator) for converting navigational data streams collected passively from in-vehicle GPS devices into an

electronic travel diary. This derived travel diary comprises a sequence of vehicle trips identified from the GPS streams, with each trip characterized in terms of attributes such as trip-end location, trip purpose (or activity type at destination), time of day, duration, distance, and speed. In addition accuracy measures are also generated to capture the impacts of signal loss or equipment malfunction on the identification and characterization of trips. The determination of the trip route (i.e., the specific network links traveled) is not within the scope of this work.

The rest of this section is organized as follows. Section 3.1 describes the data required as inputs. Section 3.2 discusses the overall travel-diary extraction algorithm. Section 3.3 presents some validation results.

3.1 Data Inputs

There are four categories of data required as inputs to the GPS-TDG. The first input is the preprocessed (obtained from the raw GPS output streams by primarily removing the invalid records) GPS streams. Each record in this stream contains the household and vehicle identifiers, local date and time, latitude, longitude, speed, heading, and the number of invalid records immediately prior to this record in the raw file. The second set of inputs is the characteristics of the primary driver of each GPS-equipped vehicle including gender, employment status, number of children, and the home- and work locations (latitudes and longitudes). This information is used in the determination of trip-purpose. The third input is a GIS layer of the parcel-level land-use characteristics of the region. This information is also used in the determination of trip-purpose. The fourth input is the set of algorithm parameters. A description of all the user-defined parameters is presented in Table 1. The values of these parameters can be varied to suit the requirements of different study areas and for various analysis objectives.

3.2 Algorithm

The GPS-TDG algorithm is designed to extract the trip-diary information for one vehicle at a time. Correspondingly, in Step 1, all data for a single vehicle are read in. In the next step, the pre-processed navigational stream is scanned until a potential trip-end is detected. In Step 3, the characteristics of the detected trip are determined. In Step 4, the reasonableness of the detected trip is examined. If the potential trip passes the validation checks, it is recorded as a trip. If not, the potential trip is not recorded as an actual trip. If the end of the GPS navigational stream has not been reached, the algorithm reverts to Step 2 to continue examining the rest of the stream. If not, the algorithm reverts to step 1 and proceeds with the processing of the next vehicle. Steps 2, 3, and 4 form the core of the travel-diary extraction algorithm and hence these are discussed further. All user-specified algorithm parameters referenced henceforth are underlined.

3.2.1 Trip Detection

The detection of a potential trip involves identifying a pair of GPS records within the overall navigational stream, with one record corresponding to the end of a trip and the other corresponding to the start of the next trip. As already discussed, the non-movement of the vehicle can be represented in the navigational stream in two ways. The detection procedure for each of these cases is discussed below.

In the “power-off” case, a pair of successive GPS records, one corresponding to the end of a trip and the next corresponding to the start of a subsequent trip, are detected by examining the dwell times between successive valid records in the pre-processed GPS stream. The procedure is as follows:

1. Compute the total time difference (*TotTimeDiff*) between successive GPS records as the difference in the time stamps of the two records.
2. Compute the signal loss time (*SigLossTime*) between the successive GPS records as the product of the number of invalid records removed (during preprocessing) immediately prior to the second GPS record and the Average Update Rate.
3. Compute the Dwell Time (*DwellTime*) as the difference between *TotTimeDiff* and *SigLossTime*.
4. If the *DwellTime* between a pair of successive GPS records exceeds the Power-off Dwell-Time Threshold, a potential trip end is flagged. The first of these two successive GPS records corresponds to the end of a trip and the second corresponds to the start of the next trip.

The dwell time is computed by subtracting the time resulting from signal loss from the total time between successive valid GPS records. Thus the procedure guards against classifying time gaps caused by extended signal loss periods as indications of trip ends.

Stops under the "power-on" scenario cannot be detected using the dwell time concept, because the GPS points are continually recorded, even during the period when the vehicle is stopped. Hence, the speed data is examined to identify nonmovement and a non–engine power-off trip end. Specifically, if the instantaneous speed recorded in the GPS stream is less than the Speed Threshold for Power-on Stop Determination continuously for a period greater than the Non–Power-on Dwell-Time Threshold, a potential trip end is detected. The GPS record from which the speed continually remains below the threshold value is taken to represent the end of a trip. The first subsequent GPS record with a speed above the threshold value represents the start of the next trip.

In addition, a potential trip end is also flagged at the end of the GPS stream for the vehicle and the last GPS record represents the trip end of this last trip. The first valid GPS record in the stream corresponds to the start of the first trip.

3.2.2 Trip Characterization

Once a potential trip end is detected using the procedures described above, several trip attributes are computed.

The position information (i.e., latitude and longitude) on the first and last GPS records of a trip determines the most detailed trip-end locations. In addition, the trip-end locations are also specified in terms of the traffic analysis zone (TAZ) at which the trip is terminated. Specifically, the latitude and longitude of the trip destination-end is overlaid on the "TAZ boundaries" GIS layer to determine the TAZ and a spatial "join" procedure is invoked. The trip origin TAZ is simply determined as the TAZ of the destination of the previous trip. The only exception to this is in the case of the first trip, where the trip origin TAZ is determined as the TAZ of the home location because the first trip is assumed to start from home.

The trip start time is computed as the time stamp on the first GPS navigational stream record corresponding to a trip and the trip end time is computed as the time stamp on the last GPS navigational stream record corresponding to the same trip. The trip duration can then be computed as the difference between the start and end times of the trip. The duration of activity at a trip end can be computed as the difference between the end time of a trip and the start time of the subsequent trip. The activity duration at the origin of the first trip and at the destination of the last trip cannot be determined.

The trip length (or distance) is determined using the point-to-point sum of distances approach (the PP approach). Broadly, this method involves computing the distance between successive pairs of recorded locations, the two points in each pair spaced apart by at least the Time Threshold for Trip Distance Computation. As already discussed, choosing a higher value (say 5 or 10 seconds) for the above threshold can help minimize the overestimation of trip distance. These distances are then summed over the entire trip to obtain the trip length. The summing of the distances is not performed over segments of the trip where the speed is less than the Speed Threshold for Trip Distance Computation (i.e., short stretches when the vehicle is not moving).

The instantaneous speeds are averaged over all the GPS records corresponding to a trip to compute the average trip speed. The standard deviation of the instantaneous speed measurements is also computed to provide a measure of variation in speed along the trip length.

The activity type undertaken by the driver of the vehicle at the trip-end location is determined next. The trip-end activities are classified into one of three aggregate types: home, work, and other. As already discussed, the locations of home and work in terms of latitude and longitude are provided as inputs to the algorithm. A trip-end activity is classified as "home" if the distance of the trip-end location from home is less than the Home Location Distance Threshold. (This is necessary to account for the difference between where the vehicle is parked and where the home is located.) A trip-end activity is classified as "work" if (1) the distance of the trip-end location from work is less than the Work Location Distance Threshold and (2) the activity duration at the trip end is greater than the Work Duration Threshold. (Again, it is necessary to consider vehicles being parked off-site from the main work location.) The activity type at the trip-ends which are neither home nor work is determined using a multinomial-logit model (See Srinivasan *et al*. 2005 for details).

Finally, an accuracy measure is computed for each trip detected. This measure, called *NRecRatio,* is computed as follows: $NRecRatio = \frac{N_{Valid}}{N_{Valid} + N_{Invalid}}$

where N_{Valid} is the number of valid GPS points for the trip and $N_{Invalid}$ is the number of invalid GPS points for the trip. The algorithm is developed such that the preprocessor removes the invalid records from the raw GPS stream and records the number of such invalid records removed immediately prior to each valid GPS record. Thus *NRecRatio* is a measure of the extent of missing or invalid GPS data for a trip. Smaller values of this measure indicate that a significant fraction of the complete GPS records corresponding to this trip were invalid and hence the trip attributes computed are less accurate than those records with higher values.

3.2.3 Reasonableness Checks

Reasonableness checks on each attribute, as well as combinations of attributes (e.g., trip timing and purpose), can be undertaken to ensure that the predicted trip characteristics are reasonable. Two specific checks were identified as part of the current effort. The first ensures that the trip duration is of at least a certain minimum value by comparing the computed duration against the Minimum Trip Duration Threshold. The second ensures that the trip length (distance) is of at least a certain minimum value by comparing the computed trip length against the Minimum Trip Length Threshold. Potential trips are classified as false trips when they have trip durations lower than the Minimum Trip Duration Threshold or trip lengths lower than the Minimum Trip Length Threshold.

3.3 Implementation and Validation

The algorithm presented above has been implanted using the Object Oriented Programming methodology in the Java programming language. ArcGIS 9.0 is employed as the platform for GIS processing. The software implementation details are available in Srinivasan *et al*. 2006.

Data from household travel surveys conducted in Laredo and Tyler-Longview (TX) were used for the development, testing, and refinement of the algorithm. These self-reported surveys were administered using the traditional CATI procedures. In each case, a subset of households was recruited for the GPS-based component. Thus these surveys provide both passively recorded and self-reported travel data for several households. The GPS and reported travel data used for testing were drawn from these households. Specifically, we identified 45 vehicles (38 households) from Laredo and 92 vehicles (86 households) from Tyler-Longview, which provided all required information for our analysis. All these vehicles were switched-powered systems.

In addition to the GPS and self-reported travel data, we also had access to files (one for each for each of the two surveys) containing trip-ends identified from the GPS streams, using methods developed by the GPS contractor on the survey team. These files provided only the start and end times of trips and no other trip-related attributes (such as purpose, distance, and trip-end location). The procedures used to identify the trips by the GPS contractor are proprietary and hence were not available. The evaluation included a comparison of the number of trips and trip-timings fromour algorithm to those from the contractor's procedures as another means of validation.

When applied to the Laredo GPS data, our algorithm identified 262 trips for the 45 vehicles from that location. The GPS contractor identified 305 trips for the same set of vehicles, and the primary drivers of these 45 vehicles reported undertaking 215 vehicle trips in the survey. For the Tyler-Longview data, our algorithm identified 545 trips for the 92 vehicles examined. The GPS contractor identified 582 trips for the same set of vehicles. The primary drivers of these 92 vehicles reported undertaking 534 vehicle trips in the survey. In general, these findings suggest that the travel-diary extraction methods presented here are effective. However, it is useful to note that neither the self-reported trips nor the number of trips determined by the GPS contractor can be completely construed as the "true" values. Therefore, the validity of trips detected from the GPS streams but not self-reported in the surveys could not be inferred.

4. CONCLUSIONS

GPS technologies are being increasingly used to improve the accuracy and completeness of travel behavior data. The passive nature of GPS data collection is very beneficial in reducing respondent burden and enhancing the quality of data. However, the data is collected in the form of a navigational stream, which has to be processed to derive travel patterns. Consequently, the use of passive GPS technology in travel surveys shifts considerable burden from the respondent to the analyst. Therefore the success of GPS technology as a survey instrument depends on the ability of the analyst to derive the activity-travel information from the GPS streams.

This paper summarizes an effort to develop and implement an algorithm to automate the process of converting navigational data streams collected passively from in-vehicle GPS devices into an electronic travel diary. This derived travel diary data comprises a sequence of vehicle trips identified from the GPS streams, with each trip characterized in terms of attributes such as trip-end location, trip purpose (or activity type at destination), time of day, duration, distance, and speed. The proposed algorithm was implemented in a flexible, user-friendly prototype

software. As the algorithm is controlled by a set of user-defined parameters, the software can be calibrated to any region.

Calibration and validation are the most important and immediate next steps toward enhancing and refining the algorithm. Such efforts require data on the "true" travel of the equipped vehicles. Unfortunately, obtaining "true" travel data is difficult unless respondents of equipped vehicles are individually tracked using means that are known to be close to 100% accurate. As a fall back, one may use CATI data where the extent of under-reporting has been documented to be minimal (assuming that information on the vehicle used for each person-trip is recorded in the survey, and this information has also been documented to be very accurate). In addition, the validation exercise can also benefit from well-designed test runs aimed at (1) fine-tuning procedures for handling signal-loss situations because of travel through urban canyons, (2) identifying GPS stream patterns that may help distinguish between short duration stops without engine off and signal delay, (3) developing algorithms for determining the trip timing more accurately (accounting for signal acquisition times), and (4) evaluating trip distance and trip speed computation procedures using odometer readings and self-recorded times (note that trip distances and speeds are not collected in travel surveys).

ACKNOWLEDGMENTS

The authors express appreciation to Michael Chamberlain, TxDOT, and the rest of the TxDOT project monitoring committee for their valuable input throughout the course of the project. This research was first documented in a series of reports for TxDOT as Project 0-5176, "Conversion of Volunteer-Collected GPS Diary Data into Travel Time Performance Measures."

REFERENCES

Axhausen, K. W., Schonfelder, S., Wolf, J., Oliveria, M., Samaga, U. (2004) "Eighty Weeks of GPS Traces, Approaches to Enriching Trip Information", Transportation Research Board 83rd Annual Meeting Pre-print CD-ROM.

Bachu, P. K., Dudala, T., and Kothuri, S. M. (2001) "Prompted Recall in Global Positioning System Survey: Proof-of-Concept Study", *Transportation Research Record 1768*, pp 106–113.

Battelle (1997) "Global Positioning Systems for Personal Travel Surveys Lexington Area Travel Data Collection Test-Final Report" prepared for the FHWA, USDOT. www.fhwa.dot.gov/ohim/lextrav.pdf, accessed on March 12, 2004.

Chung, E. and Shalaby, A. (2004) "Development of a Trip Reconstruction Tool to Identify Traveled Links and Used Modes for GPS-Based Personal Travel Surveys", Transportation Research Board 83rd Annual Meeting Pre-print CD-ROM.

Doherty, S. T., Noel, N., Lee-Gosselin, M. L., Sirios, C., and Ueno, M. (1999) "Moving Beyond Observed Outcomes: Integrating Global Positioning Systems and Interactive Computer-Based Travel Surveys", In the Proceedings of the *Transportation Research Board Conference on "Personal Travel: The Long and Short of It"*, Washington D.C.

Greenfeld, J. S. (2002) "Matching GPS Observations to Locations on a Digital Map" Transportation Research Board 81st Annual Meeting Pre-print CD-ROM.

Murakami, E. and Wagner, D. P. (1999) "Can Global Positioning System (GPS) Improve Trip Reporting?" *Transportation Research Part C*, No 7, pp 149–165.

Schonfelder, S., Axhausen, K. W., Antille, N., Bierlaire, M. (2002) "Exploring the potentials of automatically collected GPS data for travel behavior analysis – A Swedish data source", http://roso.epfl.ch/mbi/papers/schoenfelder.pdf, accessed November 29 2004.

Schonfelder, S., and Samaga, U. (2003) "Where do you want to go today? – More observations on daily mobility", presented at the 3rd Swiss Transport Research Conference (STRC), March, 2003. http://www.strc.ch/Paper/Schoen.pdf, accessed November 28, 2004.

Srinivasan, S. P. Ghosh, A. Sivakumar, A. Kapur, C.R. Bhat, and S. Bricka "Conversion of Volunteer-Collected GPS Diary Data into Travel Time Performance Measures: Final Report," Report 5176-3, prepared for the Texas Department of Transportation, February 2006.

Srinivasan, S., P. Ghosh, S. Bricka, and C.R. Bhat, "Conversion of Volunteer-collected GPS Diary Data into Travel Time Performance Measures: Algorithm for Extracting Travel Diary Data from GPS Streams and GPS-TDG Software Design," Report 5176-2, prepared for the Texas Department of Transportation, August 2005.

Stiefer, P., Coe, D., Wolf, J., and Oliveria, M. (2003) "Investigating the Impact of Driving Activity on Weekend Ozone Levels using GIS/GPS Technology", http://www.geostats.com/papers/wolf0396.pdf, accessed on March 12, 2004.

Stopher, P., FitzGerald, C., and Zhang, J (2008) Search for a global positioning system device to measure personal travel, Transportation Research Part C, pp. 350 – 369.

Stopher, P., FitzGerald, C., and Xu, M (2007) Assessing the accuracy of the Sydney Houehold travel survey with GPS, Transportation, Vol 34, pp. 723-741.

Stopher, P. R. (2004) "GPS, Location, and Household Travel", in Handbook of Transport Geography and Spatial Systems, Edited by Hensher, D. *et al.*, Elsevier Ltd., pp 432–449.

Stopher, P. R., Bullock, P., and Jiang, Q (2002) "GPS, GIS and Personal Travel Surveys: An Exercise in Visualization", presented at the 25th Australian Transport Research Forum, Canberra, Australia, http://www.btre.gov.au/docs/atrf_02/papers/57Stopher%20GPS.pdf, accessed March 12, 2004.

TRB NCHRP Synthesis (2001) "Collecting, Processing, and Integrating GPS data into GIS: A Synthesis of Highway Practice", NCHRP Synthesis 301, Transportation Research Board National Research Council.

Wagner, D. P., Neumeister, D. M., and Murakami, E., (1996) "Global Positioning Systems for Personal Travel Surveys", paper presented at the *National Traffic Data Acquisition Conference (NATDAC)*, May 1996, Albuquerque, New Mexico.

Wolf, J. (2004a) "Defining GPS and its Capabilities", in Handbook of Transport Geography and Spatial Systems, Edited by Hensher, D. *et al.*, Elsevier Ltd., pp 411–431.

Wolf, J. (2003) "Tracing People and Cars with GPS Diaries: Current Experience and Tools" presentation at ETH, Zurich, http://www.ivt.baum.ethz.ch/allgemein/wolf_030228.pdf, accessed on March 12, 2004.

Wolf, J. (2000) *Using GPS Data Loggers to Replace Travel Diaries In the Collection of Travel Data*, Dissertation, Georgia Institute of Technology, Atlanta.

Table 1 List of user-specified algorithm parameters

Parameter Name	Description
Power-off Dwell Time Threshold	Minimum dwell-time gap between successive valid GPS records for signaling an engine power-off stop (seconds)
Power-on Dwell-Time Threshold	Minimum duration for which the speed should be less than the Speed Threshold for signaling a non–engine power-off stop
Speed Threshold for Engine Power-on Stop Determination	Value of instantaneous speed below which the vehicle is assumed to be at rest (used in determination of non–engine power-off stops)
Speed Threshold for Trip Distance Computation	If the value of instantaneous speed is below this threshold, this GPS point is not used in distance computation
Time Threshold for Trip Distance Computation	Minimum time between pairs of GPS points for which the distances are computed and summed to determine the trip length
Home Location Distance Threshold	Maximum distance between a trip-end location and home location for classifying the trip-end activity purpose as "Home"
Work Location Distance Threshold	Maximum distance between a trip-end location and work location of primary driver for classifying the trip-end activity purpose as "Work"
Work Duration Threshold	Minimum activity duration at a trip end for classifying a trip-end activity as "Work"
Average Update Rate	Average duration between successive navigational points recorded by the GPS instrument used in the survey
Minimum Trip Duration Threshold	Minimum trip duration for a potential trip to be classified as a real trip
Minimum Trip Length Threshold	Minimum trip length (distance) for a potential trip to be classified as a real trip