AN AUTOMATED ACTIVITY IDENTIFICATION METHOD FOR PASSIVELY COLLECTED GPS DATA

Vetri Venthan Elango
Research Engineer
School of Civil and Environmental Engineering, Georgia Institute of Technology
790 Atlantic Drive
Atlanta GA, 30332
TEL: 404 791 5405
FAX: 404 385 2278
email: vetriventhan.elango@ce.gatech.edu

Dr. Randall Guensler
Professor
School of Civil and Environmental Engineering, Georgia Institute of Technology
790 Atlantic Drive
Atlanta GA, 30332
TEL: 404-894-0405
FAX: 404-385-2278
email: randall.guensler@ce.gatech.edu

Submitted to: The 3rd Conference on Innovations in Travel Modeling

Date of Submittal: November 6, 2009

ABSTRACT
New technology developments have led to increased use of Global Positioning System (GPS) travel survey methods. Travel surveys that passively collect GPS data, obtain travel data accurately for longer durations without survey fatigue errors that arise in traditional travel diaries. The number of trips, duration, distance, start time, and end time of trips are easy to infer from the GPS data. However, identifying the driver, the number of passengers, and the trip activity without the aid of feedback from the participants is difficult. In this research effort, the authors undertake a preliminary evaluation of a proposed methodology to automate the identification of the activity type for passively collected GPS data. Using second-by-second GPS travel data collected by the University of Minnesota in 2008 from 46 commuters, the authors apply the proposed methodology and compare the predicted activity-based trip purpose results for 1730 trips to the data provided by the participants via online electronic travel diaries. The analysis found that the predicted distribution of home, work, and maintenance activities identified were similar. However, discretionary activities and multipurpose activities were not identified accurately. The proposed methodology still needs to incorporate duration of activity, time-of-day and day-of-week variables, and implement learning algorithms from travel diaries to increase the accuracy of activity identification.

INTRODUCTION

Travel survey methods that employ traditional travel diaries have limitations with respect to duration and the accuracy of the survey. Cross-sectional travel data collected by the traditional travel diaries have limitations in the analysis of intra-household variability in travel behavior. These limitations create problems for activity-based modeling techniques which rely upon the capture of travel behavior variability at the micro-aggregate level. Longitudinal travel data are much better at capturing the variations in travel behavior over time *(1)* and GPS devices are excellent instruments for use in longitudinal travel surveys given the spatial and temporal accuracy of the passively collected GPS data. In fact, more and more travel surveys are using GPS devices to collect travel data for at least a portion of their samples. However, passively collected GPS data do not directly capture the human elements of travel diaries such as the purpose of the travel, who was driving and how many people were involved in the activity, without active input from the participants. As such, most activity-based models developed by and for transportation planning organizations do not incorporate GPS data, even though GPS data has the accuracy and higher resolution that is required for those models*(2)*. In part, it is for the agencies or consultants processing the GPS data streams to identify trip activities from the GPS data without extensive data interaction (human resource costs). This research effort proposes a methodology to automate the identification of the activity type for passively collected GPS data, which would increase cost-effectiveness of GPS-based travel data for use in model building. A case study applying this methodology to longitudinal travel data collected by the University of Minnesota in 2008 is used to demonstrate the strengths and limitations of this methodology.

BACKGROUND

Travel behavior studies using cross-sectional data assume that the individual tries to optimize his activities and that a person's activities are habitual. However, many activities occur in cycles of a week, month etc. and are not captured by the cross-sectional data*(1)*. Longitudinal data help in evaluating response lags and leads of behavioral adjustments to an event, habit persistence, and behavioral asymmetry *(1, 3)*. Longitudinal data also help in identifying cause and effect relationships associated with behavioral changes. Longitudinal data collected using passive technologies such as instrumented vehicle studies can collect data for a long time without loss in accuracy or participant fatigue.

Longitudinal travel surveys are done by using panels of traditional travel diaries or using GPS devices to collect data passively.

- Puget Sound Panel Study - The Puget Sound panel study consisted of four waves of traditional travel diaries from 1989 to 1993 *(4)*. In a panel survey, similar measurements are made for the same sample over time. The Puget Sound panel study consisted of two-day travel diaries completed by 1700 to 1800 participants in each wave. There were many changes in the demographic characteristics, home and work location of the participants between the waves. This was the first major study in the United States that captured the demographic and spatial variations over time.
- Commute Atlanta Study - The Commute Atlanta Program was an instrumented vehicle research effort implemented by the Georgia Institute of Technology designed to assess the effects of converting operating costs (gasoline taxes, registration fees, and insurance), into variable per-mile driving costs *(5)*. The research team installed GPS devices in 500

vehicles of participating households to monitor their driving patterns. These volunteer household allowed the research team to professionally install a GT Trip Data Collector in each household vehicle driven more than 3,000 miles per year. Using the equipment, researchers remotely monitored the travel patterns of these vehicles, uploading vehicle, and engine operating data via cell phone. The uploaded data were stored in a server and processed to create trip files. The Commute Atlanta Study has collected around 1.8 million vehicle trips over a three-year period. The study found that the variability in household demographics over time affect the intra-household travel behavior variability significantly. About 70 percent of the households had demographic changes in the baseline and pricing periods *(6)*.

Longitudinal data do have significant limitations. Passive collection of longitudinal data requires state-of-the-art technology and hence highly skilled labor. The turnover of equipment during the course of the data collection can affect efficient data collection. Data collection depends on external services such as wireless and GPS services that may affect the study. The cost of longitudinal data collection is large compared to that of cross-sectional data. Longitudinal data collection in both panel surveys and passive instrumented surveys also face issues associated with sample attrition and major demographic changes in participant households over time. If the data collection method is instrumented vehicle, the identification of the driver, missing knowledge about activity type, and the omission of other travel modes are other issues that exist *(1,7)*.

Travel Surveys using GPS

Current GPS-based travel surveys are of three general types; handheld diaries, in-vehicle diaries, and Internet-based solutions.

- The first GPS method uses handheld devices to imitate the traditional travel diary. The participant carries a personal GPS device, or his/her vehicle is instrumented with a device. The participant is also provided with a handheld computer in which he enters the trip characteristics at the end of the trip *(7)*. The GPS data stream and the supplemental data entered by the participant form the completed data. In this method, the spatio-temporal accuracy of the GPS device and the human elements from the participant make the dataset fairly comprehensive. However, this kind of study generally cannot be done for a long period due to survey fatigue for the participants.
- The second method is to passively collect GPS data by installing a GPS device on the participant's vehicle *(7)*. Wired devices generally do not suffer from the same level of data losses and omissions noted in handheld diary studies. This method is very useful for safety studies where the primary interest is the vehicle parameters. However, because there is no human input (unless there is an in-vehicle data terminal provided), identifying the activity type, driver, and passengers is difficult. Using this method, data can be collected for long time periods because the participants have no equipment management/maintenance responsibilities. The Commute Atlanta Study is a good example of a vehicle-based travel survey.
- The third method is a hybrid of the passive data collection with interim travel diary surveys *(7, 8)*. In this method, the participants visit a website to review their GPS-based travel traces which help them recall their trips. The information about the activity types is entered by the participant for each trip end in the survey. The revealed location data

from the surveys can then be used by researchers to help identify the activity that occurs on other days at the same location.

Identifying Activity type from GPS Data

Wolf et al. undertook a proof-of-concept study with 30 participants on the possibility of using data from GPS data-loggers to identify all parameters including trip purpose*(9)*. The study overlaid the trip ends on a geographic referenced land use database. The land use attributes of the parcel was assigned to the trip end and trip purpose was identified. The study found that for a number of trip ends, the trip purpose was not automatically assigned. For these trip ends, an investigator manually compared it with other roadway and aerial image layers to identify suitable trip purpose. The study found that only 22% of the trip ends needed follow-up questions to identify the trip purpose. It should be noted that the study had a small number of subjects and trips and hence, it was possible to manually assign trip purpose for trip ends that were not automatically assigned to a land use parcel.

Schönfelder et al. explored the potential of using automatic GPS for travel behavior analysis in 2002 using the data collected from Borlänge between 1999 and 2001 *(7)*. The Schönfelder study processed raw GPS data to first identify trips and trip ends. To identify trip purpose, the study used the underlying land use parcel data, survey information about occupation and habitual patterns to travel. The Schönfelder study noted that for different land use blocks, the radii to search for the trip ends are different.

METHODOLOGY

The objective of this research is to explore a new methodology for automatic identification of trip activity using passively-collected instrumented-vehicle GPS data. As noted in previous studies, to identify the trip purpose based on the trip ends, an underlying layer of the land use type is necessary. However, it is difficult to find high quality geographically referenced land use data, as mentioned by both Wolf and Schönfelder in their studies *(7, 9)*. The availability of accurate land-use parcel data also limits the boundary of space within which the trip purpose can be identified.

Commercial mapping software, such as Microsoft's MapPoint, include geo-coded business locations and points of interest. Because these commercial software applications cover the entire United States, the differences in the land use data formats between cities and regions are controlled to a reasonable extent. Hence, a single format of data for the entire US from the commercial software helps in the automation of land use search procedures, irrespective of the city. This research explores the use of a standalone version of the MapPoint software, coupled with Perl scripts, to identify potential trip purpose and activity based upon proximal land use characteristics at the trip end.

Assumptions

A series of Perl scripts are used to process trip end coordinate data and to identify potential trip purposes as a function of the land uses near the trip end. The following assumptions were included in the script-based methodology.

- The radius within which people tend to park their vehicles and walk to a destination is 0.2 miles. Most locations in the US are accessible by vehicles and people tend to park as close to

their destination as possible. The radius of search is assumed to be 0.2 miles to accommodate locations with large parking lots.

- The locations that are closest to the trip end are the most likely locations visited by the individual. It is likely that a series of exceptions should apply to this assumption (depending upon land use configuration, mixed use composition, etc.). But for the purposes of this analysis, the closest location was examined first. For example, three locations are returned by the search that are within 0.2 miles of the trip end and two of them are located 0.05 miles and the third is located at 0.07 miles . The methodology will only consider the two locations that are 0.05 miles away and not the third.
- The search radius for home location is 500 feet from the trip end. Vehicles are parked as close to the home location as possible, hence the radius of search is tighter. Preliminary analysis indicates that larger radii may be required for apartment dwellings.
- The search radius for work and school locations are 1000 feet from the trip end. Parking lots at work and school are frequently much further from the office or school location.
- If no businesses or points of interest within 0.2 mile of the trip end, it is possible that the individual stopped at an unlisted business or at a residential neighborhood. Since there is no way of finding the purpose, these locations will be classified as 'Unknown' purpose.
- The activity types that will be used to classify the trip ends include Home, Work, Maintenance (shopping, services, schools, and dining), Discretionary (social visit, recreation, sports, landmarks, etc.), and MultiPurpose.
- 'Potential MultiPurpose' activity type implies that there are more than one of the activity types available at the trip-end. When there are multiple activity types close to the trip end (e.g. a trip to a regional or strip shopping mall), it is not possible to conclude whether the trip was made for a single purpose or multiple purposes. For this preliminary research effort, such trips are coded as MultiPurpose, even though the trip may actually be for a single purpose.

Process

The first step is to process the raw GPS points to identify trips, trip ends, trip duration, trip distance, start timestamp, end timestamp, and eliminate bad GPS points. The detailed methods for processing raw GPS points into trips are complex and require the application of multiple quality assurance procedures. These procedures are important, but are beyond the scope of this paper.

The next step is to geo-code the home, work and school locations. Standard household demographic data and address information for home, school, and work locations are usually collected during participant recruitment. It has been observed that during longitudinal surveys, participants change household and work locations *(6)*, meaning that follow-up surveys in longitudinal efforts are required. The geo-coded work location may not be where the participant is parking their vehicles. To ensure spatial accuracy, the work locations and the home locations need to be verified using all of the longitudinal travel data that are collected. The home location can typically be identified as the most frequent trip end of all trips that occur between 6:00 PM and 6:00 AM. The work location(s) can typically be identified as the most frequent trip end of all trips that occur between 6:00AM and 10:00AM. Frequently households have multiple work locations and the vehicle can travel to either location. Based on heuristic results in the case

study, the second location is also identified as a work location if the frequency of that location is at-least 10 over a four month period.

The first step in activity identification for a trip end is to find its distance from the home, and work locations for that household. If the distance falls within the search radius, the trip purpose is assigned to Home, or Work.

If the trip end is not Home, or Work, then all businesses within 0.2 miles of the trip end are identified. The algorithms consider only the businesses/places of interest from the search results that are closest to the trip end and find the place type classification of MapPoint for these locations. Using the cross table shown in Table 1, an activity type is then assigned to the location. If all the places under consideration are of the same activity type, then that activity type is assigned to the trip end. If there is more than one activity type, the 'Potential MultiPurpose' activity type is assigned to the trip end. If there are no businesses/places of interest within 0.2 miles of the trip end, assign 'Unknown' activity type to the trip end. A flow chart illustrating the script logic for the algorithms described above is provided in Figure 1.

TABLE 1 Cross Table between MapPoint Place Type and Activity Type

MapPoint Place Type	**Activity Type**	**MapPoint Place Type**	**Activity Type**
Airports – Major	Maintenance	Restaurants - Greek	Maintenance
Airports – Minor	Maintenance	Restaurants - Indian	Maintenance
ATMs	Maintenance	Restaurants - Italian	Maintenance
Auto Services	Maintenance	Restaurants - Japanese	Maintenance
Bus Stations	Maintenance	Restaurants - Mexican	Maintenance
Campgrounds	Discretionary	Restaurants - Other	Maintenance
Cinemas	Discretionary	Restaurants - Pizza	Maintenance
Convention Centers	Discretionary	Restaurants - Seafood	Maintenance
Galleries	Discretionary	Restaurants - Steak	Maintenance
Gas Stations	Maintenance	Restaurants - Thai	Maintenance
Hospitals	Maintenance	Schools	Maintenance
Hotels and Motels	Discretionary	Shopping	Maintenance
Landmarks	Discretionary	Casinos	Discretionary
Libraries	Maintenance	Stadiums and Arenas	Discretionary
Marinas	Discretionary	Subway Stations	Maintenance
Museums	Discretionary	Theaters	Discretionary
Nightclubs and Taverns	Discretionary	Train Stations	Maintenance
Park and Rides	Discretionary	Banks	Maintenance
Police Stations	Maintenance	Grocery Stores	Maintenance
Rental Car Agencies	Maintenance	Ski Resorts	Discretionary
Rest Areas	Discretionary	Golf Courses	Discretionary
Restaurants - Asian	Maintenance	Wineries	Maintenance
Restaurants - BBQ	Maintenance	Amusement Parks	Discretionary
Restaurants – Chinese	Maintenance	Parking	Maintenance
Restaurants - Delis	Maintenance	City/Town Halls	Maintenance
Restaurants – French	Maintenance		

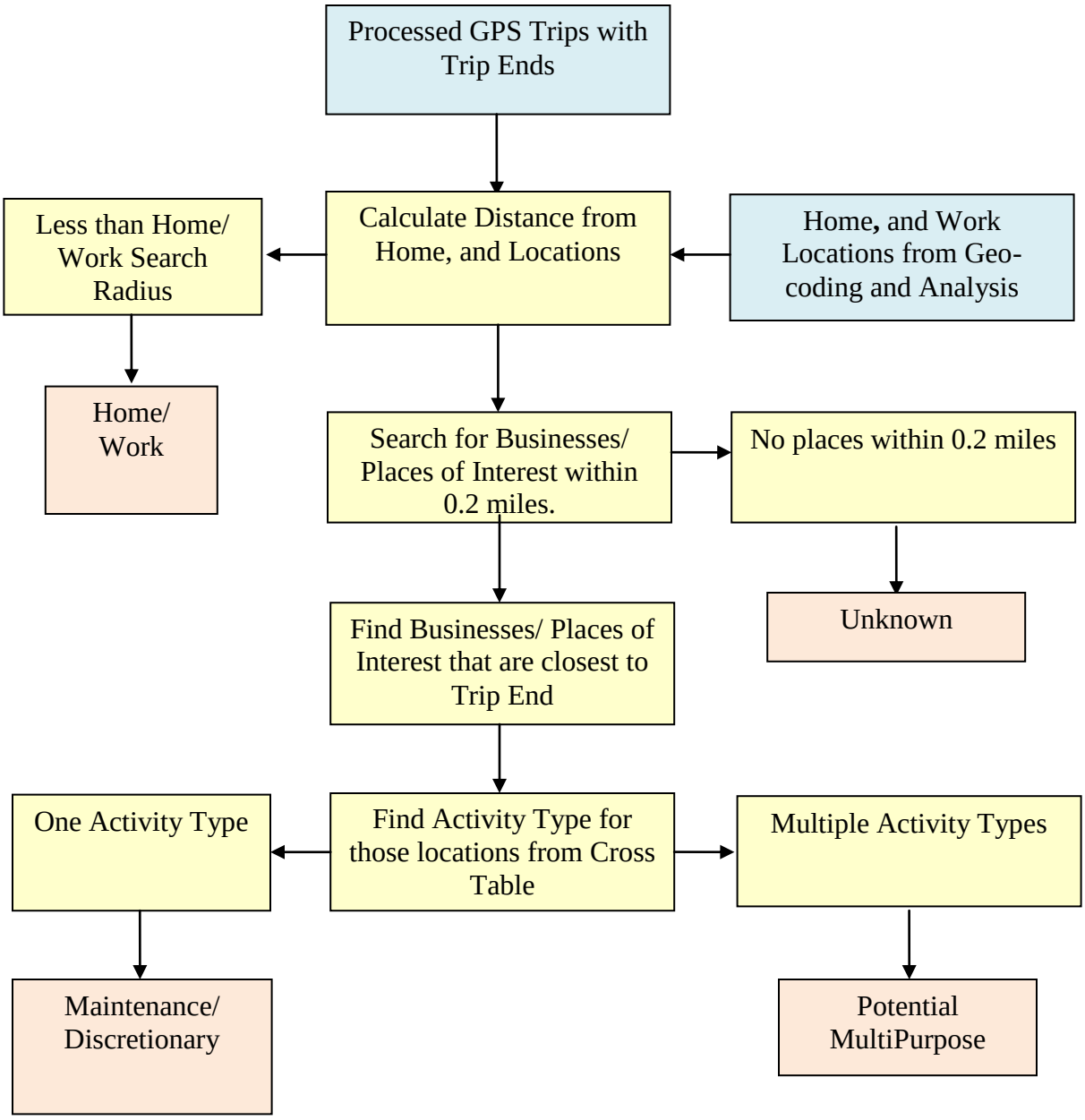

FIGURE 1 Flow Chart of Activity Identification.

CASE STUDY - DATA

The University of Minnesota conducted a travel behavior study on the use of the I-35W Bridge, which reopened in September 2008 after its fatal collapse in 2007. The study included 46 participants who commute across the I-35W Bridge. Each participant's vehicle was outfitted with a vehicle-based GPS system provided by Vehicle Monitoring Technologies, Inc. that transmitted second-by-second vehicle position data to a central server in real-time using GPRS/GSM communications. Data were collected from September 2008 through December 2008. The GPS device also transmitted engine on/off reports to the server.

The demographic data of the entire household were not collected during recruitment and only the individual participant's data are available. Data of other drivers in the household, work location of other family members and school locations of children are not available in this dataset.

The raw GPS data were processed to trips and maps containing trip traces were automatically created by the server. The participants could log in to a website to see their travel journal and complete an online travel diary to provide trip purpose and other trip-related details. The travel survey was a hybrid of passive data collection with interim requests for travel diaries. Each participant was requested to fill 6 to 14 days of travel diary through the study period.

Figure 2 shows the screen snapshot of Trip purpose recording page. One participant did not receive the travel diary requests because of an error in the email. The rest of the participants completed 94% of the travel diary requests. Some of the participants were apparently intrigued enough by the new travel diary system that they voluntarily completed travel diaries for additional days, without being asked to do so. This led to an unexpected data provision rate of 200%. That is, participants reported trip purpose details for twice as many trips as they were asked to provide data for. Participants recorded the trip purpose data for more than 4300 trips. However, we cannot be sure about the participants seriousness when completing the survey for non-requested days. Hence, for this study, only the 2185 trips for which the purpose was requested are considered.

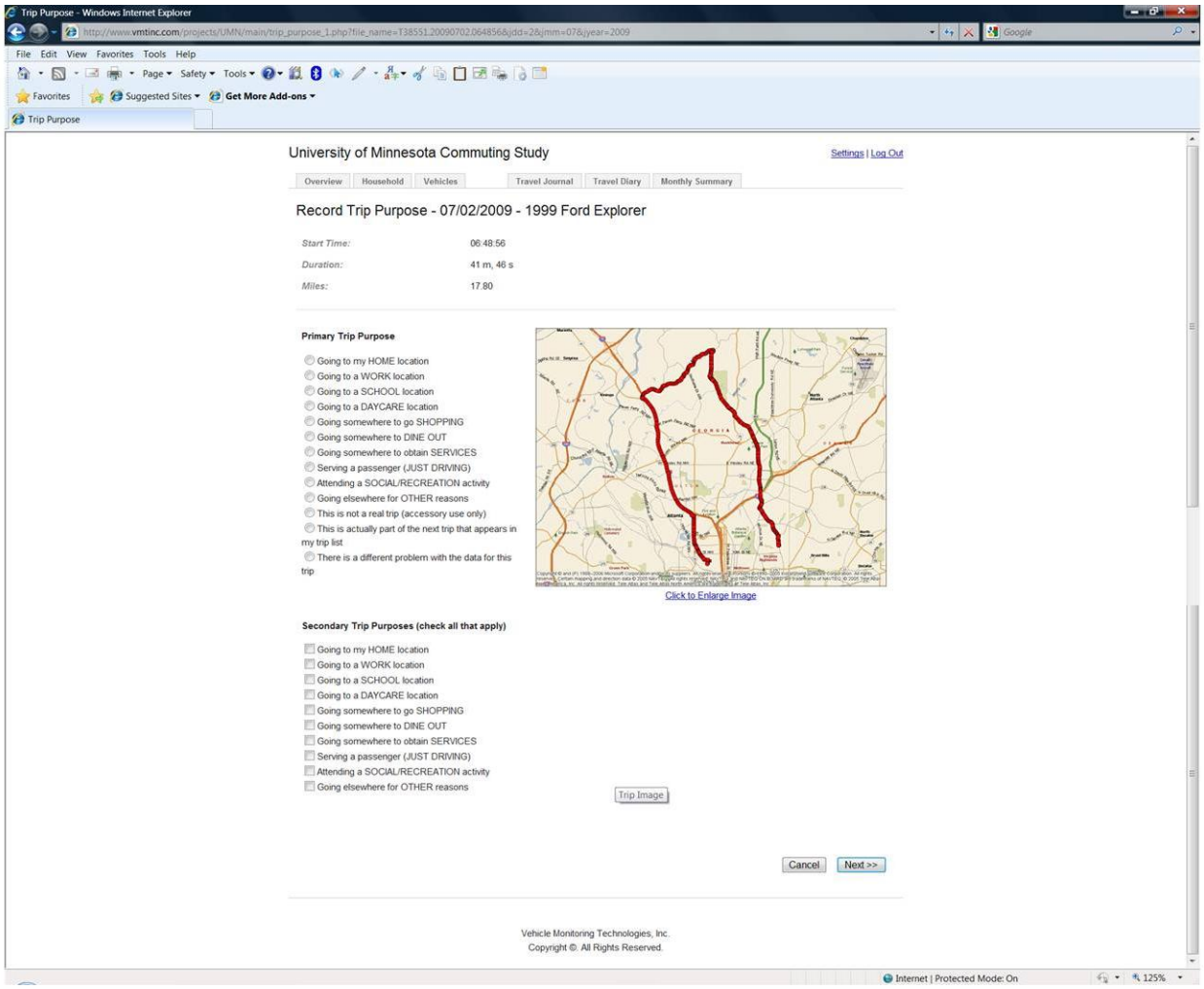

FIGURE 2 Screenshot of the Primary and Secondary Trip Purpose Recording Page

CASE STUDY

The supplemental research effort included a case study using the data from the University of Minnesota Travel Survey. As part of this analysis, scripts were developed in Perl that would implement the activity identification methodology on the GPS data. The paper compares the results of the methodology with the revealed trip purpose from the travel diaries.

The trip purpose data provided by the participants was at a disaggregate level. For example, participants reported fast-food dining as an individual category. For the purposes of the automated trip purpose comparisons, these data were first aggregated into the general trip purpose categories of Home, Work, Maintenance, and Discretionary activity types (Table 1). If the participant reported multiple activity types, then the activity is assigned 'MultiPurpose' as against "Potential MultiPurpose" in the calculated activity. Of the 2185 trips for which the participants recorded trip purpose, about 150 of them had problem in their GPS data and were eliminated from the analysis.

The research team also found that about 10% of the trips had purpose coded as “Other”. After examining a random subset of these trips, the research team believes that when a participant could not recall their trip purpose they coded it as “Other”. For the purposes of this comparative analysis, the approximately 250 trips recorded by participants as ‘Other’ were eliminated from the analysis. One household was also using their vehicle for commercial purpose and was excluded from the study (cite). Upon detailed analysis of the data stream, it also appears that one household may not have taken trip purpose reporting seriously, as evidenced by random assignment of trip purposes to known home and work locations. This household had completed the travel diary for almost every day the vehicle was instrumented and the recorded purposes were random. Hence that household was eliminated from the analysis. The final dataset has 1730 trips.

Figure 3 illustrates the distribution of the revealed activity by the participants and Figure 4 provides the distribution of the calculated activity by the automated MapPoint methodology. From Figure 3 and 4 we can see that there are more maintenance activities in the calculated activity distribution compared to the revealed activity distribution. About 67 trips fall under the ‘Unknown’ category and they are 4% of all trips in the analysis. Those trips may have ended in residential neighborhoods for social visits.

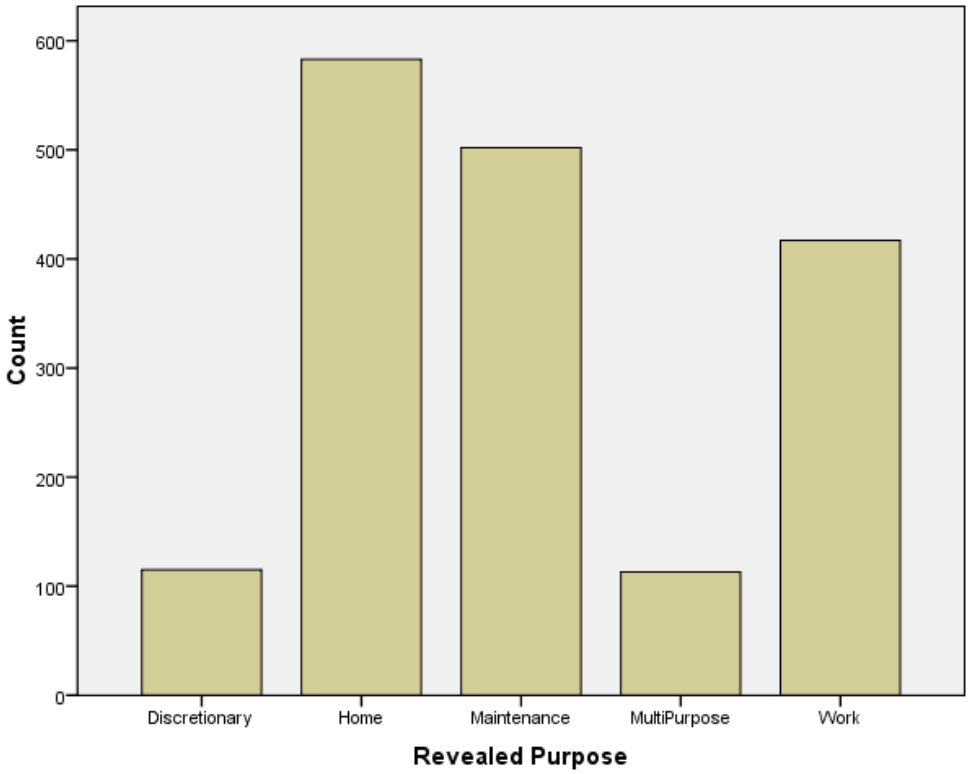

FIGURE 3 Distribution of Revealed Activity (n=1730).

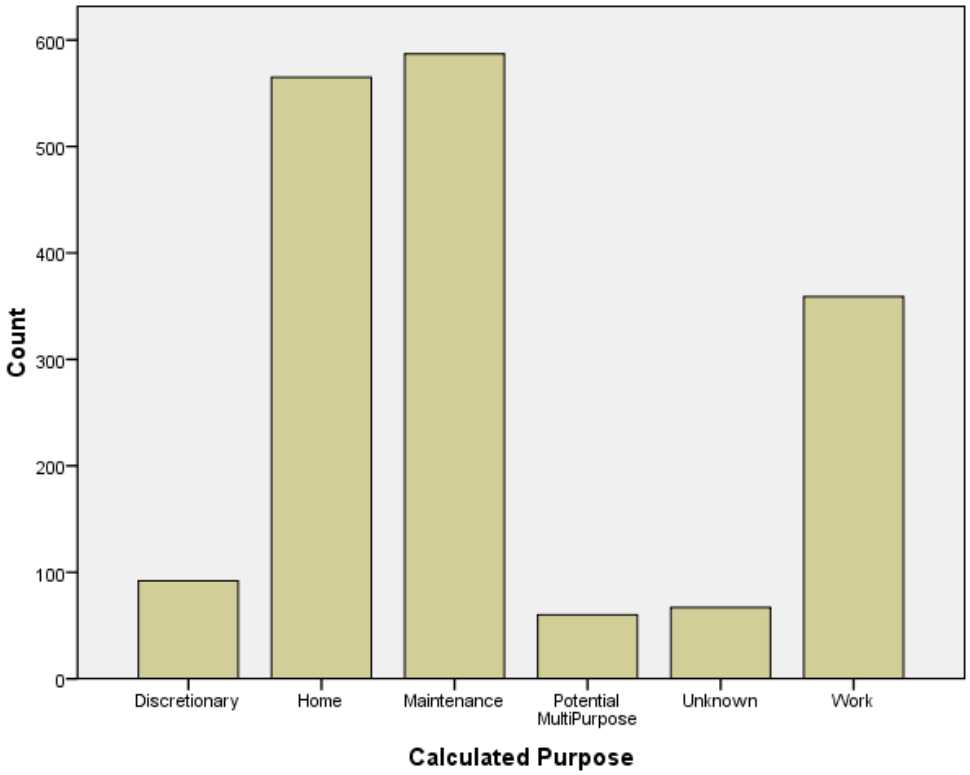

FIGURE 4 Distribution of Calculated Activity (n=1730).

Figure 5 shows the bar chart for calculated activity clustered by reported activity. Table 2 shows the numerical counts of the Cross-tabulation between reported activity and calculated activity. Home activities are predicted accurately in 84% of the cases, maintenance activities are identified with 66% accuracy, and work activities are identified with 71% accuracy. Discretionary activities and Multi-Purpose activities are poorly predicted. Overall 65.4% of the trips have been identified accurately.

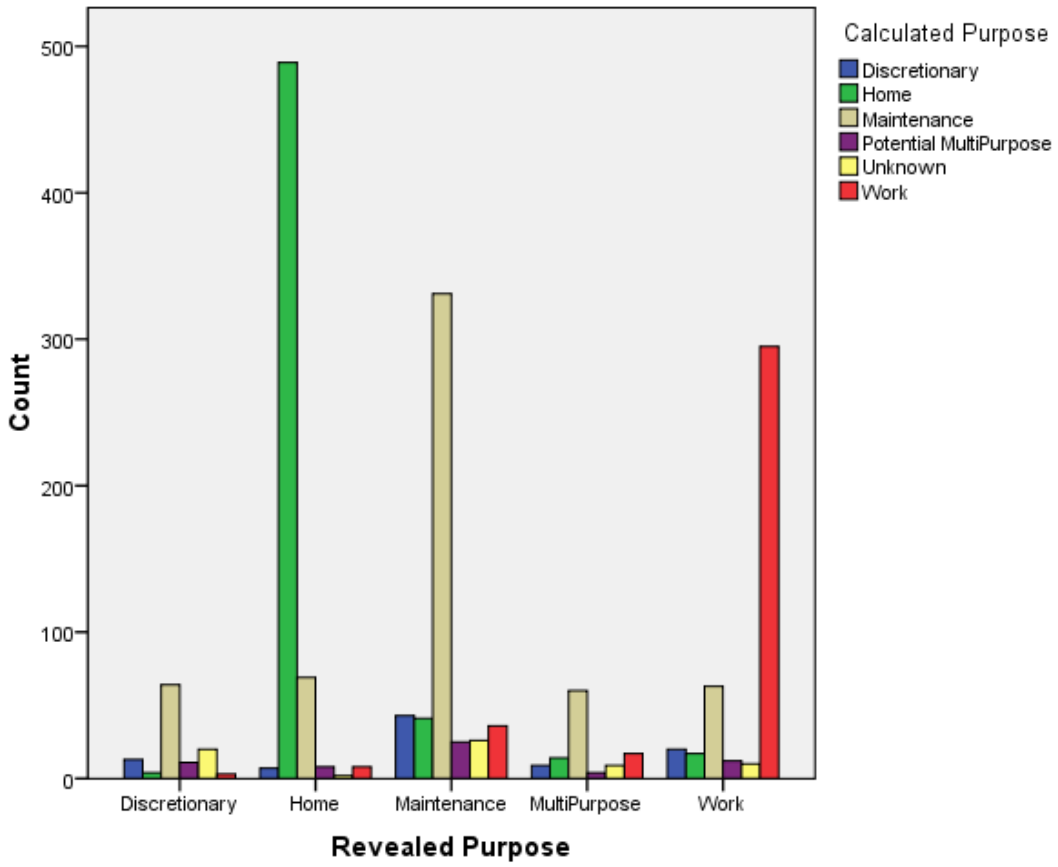

FIGURE 5 Revealed Activity vs Calculated Activity (n=1730).

TABLE 2 Cross-tabulation of Revealed Activity vs Calculated Activity

Revealed Purpose	Calculated Purpose						
	Discre-tionary	Home	Main-tenance	Potential Multi-Purpose	Unknown	Work	Total
Discretionary	13	4	64	11	20	3	115
Home	7	489	69	8	2	8	583
Maintenance	43	41	331	25	26	36	502
Multi-Purpose	9	14	60	4	9	17	113
Work	20	17	63	12	10	295	417
Total	92	565	587	60	67	359	1730

The assumption in this case study is that the reported activity is the ground truth. However, on close examination of the revealed trip purpose along with the GPS traces and time of day we find that not all revealed-purposes are accurately coded. For example, one participant coded three consecutive trips starting at 16:04, 16:18 and 16:50 as trips to home. The first two trips ended at-least a linear mile away from the home location and the last trip was the one that ended at home. The participant has obviously coded the first two trips incorrectly.

There are also limitations in the commercial software being used. For this research, MapPoint 2006 was employed. The 2006 version of this software did not include information about dentists, opticians etc. About 4% of all trips did not have any businesses in the neighborhood. To improve the quality of the results, MapPoint 2009 and other software with more updated business information will be tested.

From the above case study we find that the methodology needs further improvement to make the process more accurate for Discretionary and Multi-Purpose trips. Knowledge of other home (parent home or significant other's home where one might stay), work and school locations of all members of the household will help in accurately identifying those activities.

CONCLUSIONS AND FUTURE WORK

A new methodology that automatically identifies the activities for passively collected GPS data has been proposed. The proposed methodology does not require human investigation of the GPS data to identify the activity type. This methodology uses commercial mapping software, such as MapPoint, in the place of geographically referenced land use data. This helps make the methodology applicable anywhere in the United Sates and eliminates the variability in the data formats of the land use data by different organizations. The various assumptions that go into the methodology are based on passively collected data from instrumented vehicles (and the high-levels of contiguous data and spatial accuracy associated with vehicle-based data stream). Hence, these assumptions should be re-evaluated if data are collected from hand-held GPS loggers or by other means.

A case study compared the activity types predicted from this methodology with the revealed activities from travel diaries. The data collected by the University of Minnesota in 2008 was used for this study. The analysis showed that this methodology can accurately predict Home, Work and Maintenance activities. Using the automated tool to identify discretionary and multi-purpose activities will require significant improvements. Overall the methodology identified 65.4% of the trips accurately. However, the authors also found that the Revealed Purpose is not always the ground truth.

The authors are currently taking the next steps to improve the methodology by incorporating duration of activity, time-of-day, and day-of-week into the algorithms being used to identify trip activity. Many travel-behavior activities are habitual, occurring at the same locations and at the same times. The methodology will also incorporate learning algorithms that will use two-day travel diary data to automatically predict the activities that occur on other days. With these future improvements, this methodology may be useful in predicting activity types for hybrid travel surveys that employ passive GPS data collection with interim travel diary surveys.

ACKNOWLEDGEMENTS
Vehicle Monitoring Technologies Inc., a Georgia Tech Venture Lab Company, collected the vehicle activity and travel diary data used in this analysis for the University of Minnesota. VMT, Inc. can be reached at info@vmtinc.com.

REFERENCES

1. Elango, V. V., R. Guensler and J. H. Ogle *Day-to-Day Travel Variability in the Commute Atlanta, Georgia, Study*. Transportation Research Record, 2007.
2. Yanzhi Xu and Randall L. Guensler, *Advantages of Long-Term Continuous GPS-Based Survey Data For Activity-Based Travel Demand Modeling*, Manuscript submitted to the 89th Transportation Research Board Annual Meeting, 2010
3. Pendyala, R. M. and E. I. PAS *Multi-Day and Multi-Period Data for Travel Demand Analysis and Modeling*. Transportation Research Board, 2000.
4. Thomas F. Golob, R. K., Lyn Long. Chapter 6 - Puget Sound Transportation Panel. In *Panels for Transportation Planning*, Kluwer Academic Publishers, 1997.
5. Ogle, J., R. Guensler and V. Elango *Georgiaâ€™S Commute Atlanta Value Pricing Program: Recruitment Methods and Travel Diary Response Rates*. Transportation Research Board, 2005.
6. Xu, Y., L. I. Zuyeva, D. Kall, V. V. Elango and R. Guensler *Mileage-Based Value Pricing: Phase II Case Study Implications of Commute Atlanta Project*. Transportation Research Board, 2009.
7. Schönfelder Stefan, A., Kay W. Antille Nicolas, Bierlaire Michel. *Exploring the Potentials of Automatically Collected Gps Data for Travel Behaviour Analysis*. ETH, E. T. H. Z., Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau IVT, 2002.
8. Doherty ST, Noel N, Gosselin M-L, Sirois C, Ueno M, *Moving Beyond Observed Outcomes, Integrating GPS and Interactive Comuter Based Travey Behavior Surveys*, Personal Travel: The Long and Short of It, Transportation Research Board, 2001.
9. Wolf, J., R. Guensler and W. Bachman *Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data*. Transportation Research Board, 2001.