REGION II
UNIVERSITY TRANSPORTATION RESEARCH CENTER

Final Report

Advanced Applications of Person-based GPS in an Urban Environment

Prepared by

Catherine T. Lawson
Associate Professor
Geography and Planning
University at Albany - SUNY
518-422-4775
lawsonc@albany.edu

Cynthia Chen
Assistant Professor
Civil Engineering
City College - CUNY
212-650-5372
c.chen@ccny.cuny.edu

Hongmian Gong
Associate Professor
Geography
Hunter College - CUNY
212-772-4658
gong@hunter.cuny.edu

February 2010

UNIVERSITY AT ALBANY
State University of New York

Disclaimer

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Advanced Applications of Person-based GPS in an Urban Environment		5. Report Date February 29, 2010
		6. Performing Organization Code
7. Author(s) Catherine T. Lawson, University at Albany Cynthia Chen, City College Hongmian Gong, Hunter College		8. Performing Organization Report No. 49111-14-21
9. Performing Organization Name and Address University at Albany 1400 Washington Albany, New York 12222		10. Work Unit No.
		11. Contract or Grant No.
12. Sponsoring Agency Name and Address University Transportation Research Center City College of New York 138th Street & Convent New York, NY 10031	Federal Highway Administration U.S. Department of Transportation Washington, D.C.	13. Type of Report and Period Covered Final Report 1/1/2009 – 12/31/2009
		14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Traditional travel surveys provide essential information on travel patterns, but are time-consuming, expensive, and have seen declining rates of participation. Recently global positioning systems (GPS) technologies have been introduced to facilitate the data collection process. While GPS traces can provide accurate information on the location and time of travel, these traces do not contain explicit information on the mode used (e.g., walking, biking, transit or auto). Strategies under consideration for determining the mode, using the GPS trace data, have not been systematically investigated with respect to the needs of practitioners. A review of mode identification strategies reveals a mixed set of results from the three primary methods currently in use. These methods are rule-based, neural networks, and fuzzy logic. This research evaluates these three methods to determine their effectiveness for mode detection, their ease-of-application, and their overall performance. In order to move the concept of using GPS trace data as a substitute for traditional or modified travel surveys -- moving from small sample feasibility studies to agency-level implementation -- there needs to be sufficient evidence that GPS trace data can be handled with ease, and that the promise of reduced respondent and processing burden can be realized.

17. Key Words	18. Distribution Statement
Global positioning systems (GPS), travel surveys, GIS, neural networks, fuzzy logic	

19. Security Classif (of this report)	20. Security Classif. (of this page)	21. No of Pages	22. Price
Unclassified	Unclassified		

Form DOT F 1700.7 (8-69)

Table of Contents

List of Tables

Page

List of Figures

SECTION OVERVIEW

Part One: BACKGROUND information on travel surveys includes a short history of the efforts to gather information on household travel patterns and the challenges facing traditional surveying methodologies. Recently, global positioning systems (GPS) technologies have been introduced to facilitate the data collection process. While GPS traces can provide accurate information on the location and time of travel, these traces do not contain explicit information on the mode used to travel (e.g., walking, biking, transit or auto). Strategies under consideration for determining the mode using the GPS trace data have not been systematically investigated with respect to the needs of practitioners.

Part Two: METHODOLOGICAL ANALYSIS reveals a mixed set of results from the three primary methods currently in use: rule-based; neural networks; and fuzzy logic**.** Rule-based mode detection uses a series of hierarchical rules. The neural network method uses an input layer, an output layer and a hidden layer approach. Fuzzy logic uses approximations. Nine applications are reviewed.

Part Three: EVALUATION OF GPS MODE DETECTION STRATEGIES compares rule-based, neural networks and fuzzy logic approaches using an experimental dataset comprised of 101 trips. The effectiveness and appropriateness for practice of the three methods are evaluated.

PART ONE

BACKGROUND

Introduction

Finding an effective method for obtaining information on the travel patterns of household members is a critical need for transportation planners and researchers. This is particularly the case for Metropolitan Planning Organizations (MPOs), now being pressured to improve their travel demand forecasting capabilities. A recent review of transportation planning practice pointed out that current travel demand models are inadequate to address transportation policy concerns (e.g., lack of data for non-motorized trips) (Transportation Research Board 2007). Most transportation modeling efforts rely heavily on the data from household travel surveys. These surveys are expensive to implement and have only been conducted in a limited number of cities, very infrequently (Lawson, Fassman et al. 2007). These surveys also can place a large burden on the respondents to recall and report their travel patterns and on agencies with respect to resources (including staff time and consulting costs).

An emerging strategy for generating high-quality data is the use of Global Positioning Systems (GPS) trace data to collect travel patterns. While some challenges have been addressed in the collection and application of GPS trace data (Lawson, Chen et al. 2008), there is ample evidence that these data are more accurate than traditional household travel surveys that rely upon respondents to remember the time and location of their movements. The GPS trace data includes data elements that accurately track the location and time of travel, but not the essential mode information, such as walking, biking, transit or auto. Currently, this data is used as supplementary information, requiring survey participants to carry the GPS unit and fill out a traditional survey instrument, or provide information during a prompted recall interview over the phone. A modified methodology reduces the survey component and allows participants to enter mode information on the web, on the GPS device, or other electronic unit. An emerging area of research uses only the GPS trace data to infer mode. Such a strategy could greatly reduce respondent burden, and at the same time, provide a high quality source of travel behavior information.

Evolution of Travel Survey Methods

Since the 1950s, the methods used to collect household travel data have been expensive, with an excessive burden on respondents to self-report their travel activities for a specific time period (e.g., asking for all the travel activities of each household member, over a one or two day period). Stopher (2009) describes the original in-home interviews process where interviewers knocked on the front doors of a selected sample of homes, unannounced, to ask a set of questions about the previous day's travel.

By the end of the 1970s, in-home interviews were replaced with a strategy that included recruiting households by telephone, followed by a mail-out/mail-back survey to consenting households. Unfortunately, the response rate was very low, prompting researchers to rely on the telephone for retrieving information recorded in a written diary-style survey instrument, as well as the initial contact for recruitment. The use of "active" data (e.g., respondent reports data elements) continues to be problematic, with an increasing reluctance by household members to answer telephone calls from unrecognized sources (using caller ID services); increasing use of cell phones and corresponding reduction of "landline" telephone use; and increasing refusal rates by households (Stopher 2009).

In the mid-1990s, new technologies capable of collecting data without respondent participation (e.g., global position systems (GPS)), offered the promise of a methodology to replace the traditional household travel survey data approach. Bricka (2009) reports the use of GPS in regional travel surveys in the United States, beginning in Lexington, Kentucky, in 1996. Bricka and Bhat (2006) found the inclusion of GPS in travel surveys in Los Angeles; Pittsburgh; St. Louis; Kansas City; Reno; Portland; Seattle; Chicago; Washington, DC; and Baltimore. Since 1997, Texas has conducted six regional surveys that included a 10% GPS subsample. California and Ohio both used GPS in statewide surveys. However, all but the Portland and Chicago efforts were in-vehicle applications, rather than on-body GPS. The purpose of all of these GPS deployments was to audit trip rates.

Perdok et al. (1998) proposed a strategy for on-body data collection that required a person to carry a fairly bulky unit to record the position and time data, and the carrier to "input" the mode used and trip purpose into a portable device. Stopher and Greaves (2006) describe recent efforts to increase the use of passive data collection methods primarily to increase accuracy and reduce costs.

.

More recently, the New York Metropolitan Transportation Council (NYMTC) sponsored a special study for the use of GPS for their upcoming large scale household travel surveying effort (Lawson, Chen et al. 2008). NYMTC staff was particularly interested in "research to practice" methods that were transparent, cost-effective, could be replicated, and, where possible, implemented by in-house staff. It was demonstrated in this effort that staff could deploy the GPS equipment and participate in the data collection effort. Currently, the data can provide extremely dense and accurate data on the location and time of a travel event, but lacks explicit information on the mode of travel. Developing an effective mode identification strategy from the GPS trace data is critical in the preparation of a practice-ready dataset[1].

[1] It should be noted that while it is theoretically possible for a mode identification strategy to be implemented at any time during the post-processing operations (e.g., before or after each trip segment is identified, or simultaneously with other data enhancements, such as trip purpose identification), for the purposes of evaluation, the mode identification step will be assumed to be implemented after a trip segmentation step has been performed.

Methodologies used in previous studies

Thirteen studies were reviewed to establish the state of the practice using GPS trace data for travel surveying or related efforts. Table 1 lists the characteristics and outcomes of these studies.

Table 1: Characteristics and Outcomes of the Studies Reviewed

First Author (year)	Location of study	Number of cases	Processing steps prior to mode identification	Study Outcomes
Bialostozky (2009)	New York Metropolitan Region	24 students	Define trip between two clusters of points before beginning mode identification.	Develops a computerized methodology using geographic information systems (GIS) to determine mode.
Bothe (2009)	Amersfoort, Veenendaal, & Zeewolde, Netherlands	1104 participants	Remove unreliable trackpoints and divide into trips by categories.	Develops a methodology using GPS data, to determine modes of travel and a validation process that uses a web-based user interface.
Chung (2005)	Toronto, Canada	60 students	Identification of road links using GIS layer of streets.	Using replicated travel patterns, develops a tool for automatic processing of GPS data using a rule-based algorithm with attributes of modes.
Clifford (2008)	Adelaide, Australia	Not reported	GPS data is assumed to be divided into trips.	Develops procedure for identifying mode using GIS and heuristics based on speed, route of travel, and demographic information.
de Jong (2003)	No application	No application	GPS data is "cleaned" and then mapped in GIS.	Develops a series of rules for determining mode from GPS data with additional data from recall surveys.
Flamm (2007)	N/A	N/A	N/A	Describes process to be used to collect data from 30 persons using GPS, with recall prompt from an interviewer to acquire mode information.
Gilani (2005)	University of Southern Florida area	100 trips, 25 of each type taken by author	Critical points are determined in the GPS data.	To develop an automated process for determining mode from GPS enabled phone data.
Gonzalez (2008)	Tampa, Florida	114 trips	GPS-enabled mobile phones programmed to output location every four seconds.	Demonstrates the feasibility of using GPS-enabled phone data and neural networks to automatically detect mode.
Lawson (2008)	New York Metropolitan Region	5 students using 6 GPS units	Data is cleaned.	Relevant portion of report demonstrates the feasibility of using an algorithm in Excel to develop mode profiles.

(Table 1 continued)

First Author (year)	**Location of study**	**Number of cases**	**Processing steps prior to mode identification**	**Study Purpose**
Schuessler (2009)	Zurich, Winterthur, and Geneva, Switzerland	4,882 participants with on-body GPS receivers	Data is cleaned and then trips and activities are determined.	Develops a post-processing procedure using only GPS data to identify mode.
Stopher (2008a)	N/A	Not provided	Data is cleaned and segmented into trips	Develops software to identify modes.
Tsui (2006)	Toronto, Canada	9 volunteers (109 trips)	Data is "cleaned", followed by identification of activity and trip.	Develops an integrated process using GPS, GPS and GIS, and a fuzzy logic-based mode identification algorithms.
Upadhyay (2008)	N/A	N/A	N/A	Using fuzzy logic based algorithm (see Schuessler and Axhausen 2009), develops a methodology for identifying a set of parameter values.
Zheng (2008)	14 cities in China	45 participants	Segmentation method used on raw GPS data.	Develops a methodology to identify mode for personal use, particularly outdoor movements.

The majority of these studies used volunteers or very small samples. Both of the large studies were conducted in Europe (Bothe and Maat 2009; Schuessler and Axhausen 2009). Most of the studies were experimental in nature, either focused on examining the feasibility of a method or developing the method itself.

Table 2 provides a detailed description of the various algorithms developed, modes detected, and strategies used for mode detection. Comparing and contrasting these studies presents a challenge, as they vary in approach, and for the most part, each has been conducted with a unique dataset. To facilitate a comparison of these approaches, it is necessary to categorize the methods used in a manner that would allow MPOs, and other transportation organizations considering the use of the GPS for travel surveys, to evaluate respondent, deployment, and processing burdens.

Table 2: Identified Modes and Description of Strategies Implemented

1st Author (year)	Modes Identified	Strategy Implemented
Bialostozky (2009)	Auto, bus, subway, commuter rail, walk, and underground transfer	Designate underground travel when 2 consecutive points are more than 120 s and 250 m apart; Designate walk segment at every modal transfer (at least 60 s; maximum speed <= 10 km/h; average speed <= 6 km/h); Identify segment as walk, non-walk aboveground, or non-walk underground; If non-walk segment, identify above ground subway or rail. GPS data is mapped using GIS layers with subway lines and rail lines; If non-walk on streets, bus segment must start and end near a bus stop, mapping to GIS layer of bus stops and be located on streets with bus routes, at <= 55 mph, with acceleration <= 1.5 m/s2; Signal gaps indicate underground segments, identified on GIS layers near subway station is subway trip; and Gaps in car or bus trips are identified as tunnels with GIS layer.
Bothe (2009)	Auto, bike, walk, train, bus or metro	Calculate average and maximum speed to determine mode as walking, cycling, or auto. Train trip is identified if > 50% of trackpoints are within 50 m of center of a rail line and maximum speed is at least 20 km/h. GIS layer with train station locations is used to map train trip start and end points.
Chung (2005)	Walk, bicycle, auto and bus (or street car)	To set the modality of a trip – use the following rules: IF average trip_speed < 10 km/h AND max trip_speed < 14 km/h THEN set modality = 'foot'; ELSE IF average trip_speed < 25 km/h AND max trip_speed < 45 km/h THEN set modality = 'bicycle'; ELSE IF average trip_speed < 200 km/h AND THEN set modality = 'car'; IF trackpoint is within railarea (line element of 100 meters width following rail tracks), THEN trackpoint = railpoint; IF nr trackpoints within trip>= 20 AND max trip_speed > 20 km/h AND 2*nr railpoints within trip > nr trackpoints within trip THEN set modality – 'train' AND set category = 'railwaystation'
Clifford (2008) Stopher (2008a,b)	Foot (walk), bicycle (bike), private vehicle (auto) and public transport (rail, bus, tram and ferry)	Requires GIS layers: street network; all public transport routes (including rail and subway lines); all bus stops and station locations/Walk is identified by low speeds for the entire trip. Off-network public transport modes (rail and ferry) are then indentified using GIS layer of routes for these modes. Underground rail trips appear as gaps – repaired in ad hoc process. Speed, acceleration and deceleration are reduced to 85th percentile to remove extremes. Bus is based on appropriate speeds and proximity to bus stops in GIS layer. Deceleration is expected near these locations. Demographic data is used to identify bike trips, based on ownership of at least one bike per household. Bike trips must begin from home location or continuation of previous bike segment and have appropriate speeds, acceleration and deceleration patterns. Remaining trips are auto by default, with check for appropriate speed, acceleration and deceleration patterns. Passenger is identified as any person without a current license (obtained from demographic information).

(Table 2 continued)

1st Author (year)	Modes Identified	Process Implemented
de Jong (2003)	Walk, train, ferry, bus, car	Walk: no fixed route; no fixed stops; rare unwanted stops; max speed 7 to 15 km/h; average speed 4 km/h; no unique infrastructure. Train: fixed route; fixed stops; rare unwanted stops; max speed 110 km/h; average speed 60 km/h; requires own infrastructure. Ferry: fixed route; fixed stops; rare unwanted stops; maximum speed 25 to 45 km/h; average speed 20 to 30 km/h; required own infrastructure. Bus: fixed route; fixed stops; often makes unexpected stops; maximum speed 100 km/h; average speed 30 km/h; requires no unique infrastructure. Auto: no fixed routes; no fixed stops; makes many unexpected stops; maximum speed 120 km/h; average speed 40 km/h; requires no unique infrastructure.
Gilani (2005)	Walk, bike, bus, car	Uses algorithm that predicts mode as soon as data comes to the server using travel speed. Classifies the percentage of a trip that is below walk and bike top speed as walk percentage and bike percentage, respectively. Where walk percentage is > 90%, walk is identified, followed by bus is > 70%, if not bus, then bike if > 90% is bike percentage, otherwise, auto mode.
Gonzalez (2008)	Car, bus, walk	Method 1 uses neural network, a Multi-Layer Perception. Inputs include: average speed, maximum speed, estimated horizontal accuracy uncertainty, percent Cell-ID fixes, standard deviation of distance between stop locations and average dwell time. Method 2 uses only critical points and the following inputs: Average acceleration, maximum acceleration, average speed, maximum speed, ratio of the number of critical points over the total distance of the trips, ratio of the number of critical points over the total time of the trip, total distance, and average distance between critical points. Method 3 uses a modified version of the critical points that collected the velocity and time of the GPS point immediately before or after a currently identified critical point.
Lawson (2008)	Walk, train, auto, bus	Original kml GPS file is converted and imported to Excel as comma-delimited file and "cleaned" for points with reasonable positional accuracy. Computed cumulative distance of travel from start of a trip and plot this distance over time using colors to represent data quality. A new kml is created from "good" points with time stamp references. Output to Google and compare with traveler narratives. Associate patterns with modes. If current mode is "walking": Find the data point where slope (computed as avg. speed over 2 minutes from a given data point) changes to above 4 km/hr, or when the signal disappears. If current mode is "train" and current data point is bad, find the next good data point. If current mode is "train" and current data point is not bad (not underground), or if the current mode is "vehicle" (bus or train not yet determined), find the data point where slope reduces to 2 km/hr, and then further determine the exact data point where speed is zero for at least 20 seconds, or when the signal disappears.

(Table 2 continued)

1st Author (year)	Modes Identified	Process Implemented
Lawson (2008) – continued		Compute time difference, cumulative distance traveled and average speed for segment. If current mode is "train", and speed at segment end is not zero, then go to step a. to look for a new segment, end. The train has gone from underground to above ground, but the mode is still "train". If current mode is "vehicle", then track change in heading over 5 second intervals for the entire segment. If the heading changes sharply (less than 50 degrees (or greater than 310 degrees) over 5 seconds, then the vehicle is not a train. Set mode to "bus". Record current segment mode, start and end times. Set current segment end as next segment start and proceed with the loop. Once the entire data set has been looped through, check the output trip summary for completeness and accuracy. If multiple sequential segments are of the same mode and have less than 1 minute intervals between them, then combine them into one mode.
Schuessler (2009) Upadhyay (2008)	Walk, cycle, auto, public transport (bus/tram), rail	Open source fuzzy engine uses the following: median of the speed distribution with four membership functions, and 95th percentiles of the speed and acceleration distributions, with three membership functions each. Fuzzy rules are derived to characterize modes, followed by defuzzify method that calculates the likelihood for each mode based on all mode scores. The third step determines the reasonability of the derived mode chains.

PART TWO

METHODOLOGICAL ANALYSIS

Introduction

Across the various studies, the methodologies used to identify modes from GPS trace data fall into three basic methodologies. These methodologies are rule-based, neural networks, and fuzzy logic. These methods each have distinctive characteristics, with advantages and disadvantages associated with their use and effectiveness.

Methodology Descriptions

Rule-based mode detection can be accomplished using a series of hierarchical rules. These rules first attempt to identify the most prevalent and most easily identifiable modes (e.g., walking), then progressively bring in more data and complicated analysis to identify the more difficult modes (e.g., subway and rail). This approach to mode detection is the most prevalent in the literature (Chung and Shalaby 2005; Gilani 2005; Clifford, Zhang et al. 2008; Stopher, Clifford et al. 2008a; Bialostozky 2009), but the specific rules in each model differ drastically. The input parameters and datasets are often distinct in the different mode detection methodologies. Most notably, numerous methods use geographic data incorporated in a GIS system (Stopher, FitzGerald et al. 2008b; Bialostozky 2009), such as the location of bus stops/routes and subway lines. Other approaches (de Jong and Mensonides 2003; Zheng, Liu et al. 2008) apply a less data-intensive approach and use rules associated with the speed, acceleration, time of day and the distance between arrival and destination points.

Neural networks are distinctly different from traditional rule-based approaches. One method in the literature implemented a neural network to determine the travel mode (Gonzalez, Weinstein et al. 2008). Neural networks are a common technique used in artificial intelligence applications, based on a simplified model of neuron function in the brain. The main advantage of using neural networks, compared to other techniques, is adaptability across a wide range of different datasets and ability to exploit subtle patterns to make predictions, given a well-formed training dataset.

A neural network typically consists of three layers: input, output and the hidden layer. An input layer contains the values of the different variables thought to be useful in determining the output, such as average speed, acceleration, etc. An output layer reflects the possible range of results, such as different mode choices (e.g., walking, auto, bus, subway). An additional layer, called the hidden layer, serves as the location for intermediate connection nodes between the input layer and the output layer. Proper training is vitally important in effective use of a neural network. During training, a technique known as back propagation is used, in which the relative weights between the different nodes in each layer are adjusted based on correct or incorrect matching between the predicted output and the correct output. If a certain configuration of weights between the different nodes successfully predicts the correct output, the weights are strengthened, while an incorrect response negatively adjusts the values of the weights. After sufficient and comprehensive training, particular adjustments of weights are chosen to reflect the highest probability of successfully predicting the correct mode choice, based on a set of input parameters.

The third methodology, fuzzy logic, is derived from the fact that the methodology does not view the world in terms of absolutes. Fuzzy logic was designed for imprecise situations, which cannot be defined with typical Boolean logic of true and false. Fuzzy logic allows for approximating values between the values of zero and one. For instance, for a GPS trip, the median travel speed can be defined in terms of slow, medium, or fast. In everyday situations, people typically do not reason that there is an exact cut-off between someone traveling slowly and fast. In other words, it would not be logical to consider that someone traveling at 20 miles per hour is traveling slowly and someone traveling at 25 miles per hour is traveling fast. Instead, everyday reasoning consists of various degrees or levels. In fuzzy logic this might be describe someone traveling at 25 miles per hour as 0.6 fast and 0.4 slow.

Table 3 reports the mode identification effectiveness with rule-based methods, the most prevalent methodology. The vast majority of the rule-based algorithms were implemented using information gathered from geographic information systems (GIS), (e.g., the location of bus stops or rail lines). This is in contrast to research conducted by Tsui and Shalaby (2006) which found no difference in performance between GIS-based approaches and less data-intensive algorithms that do not rely on GIS. Relatively simple rules appear to predict about 70% of the travel modes (Bohte and Maat 2009). In particular, using basic rules derived from maximum speed and acceleration, one can easily identify walking trips with greater than 95% accuracy (Lawson, Chen et al. 2008; Stopher, Clifford et al. 2008a; Bialostozky 2009). A variation of rule-based methods using fuzzy-logic was also found in the literature review (Tsui and Shalaby 2006; Upadhyay, Schuessler et al. 2008).

There is little consistency in results between the different methodologies shown in the literature (rule-based, neural networks, and fuzzy logic). Comparing the methodologies for effectiveness, given the variation in the data components and mode choice opportunities, creates some uncertainty with respect to the best performing strategy. There are also concerns where only the overall effectiveness is reported (e.g., Stopher, FitzGerald et al. 2008b). The rule-based results range from a low of 70% overall, to 97%. The neural networks study includes only three possible modes (Gonzalez 2008), while the fuzzy logic, with and without GIS, appears to have the same overall outcomes (Tsui and Shalaby 2006). Because of these variations, it is necessary to conduct a controlled experiment, using a single dataset, and applying the three possible methodologies to the same dataset, measuring the accuracy and ease-of-use of the difference algorithmic approaches.

Table 3. Effectiveness of Reported Mode Identification

1st author (year)	Method	**Over-all**	Walk	Bike	Bus	Street-car	Subway	Rail	Und T*	Auto	Other
Bialostozky (2009)	Rule-based with GIS	**79%**	92%	--	54%	--	68%	29%	83%	96%	--
Bothe (2009)	Rule-based with GIS	**70%**	68%	72%	0%	--	--	34%	--	75%	7%
Chung (2005)	Rule-based with GIS	**92%**	100%	100%	80%	--	--	--	--	88%	--
Giliani (2005)	Rule-based with GIS	**97%**	100%	96%	92%	--	--	--	--	100 %	--
Gonzalez (2008)	Neural Networks	**91%**	100%	--	82%	--	--	--	--	92%	--
Stopher (2008b)	Rule-based with GIS	**95%**	NR**	NR	NR	--	NR	NR	NR	NR	NR
Tsui (method 1) (2006)	Fuzzy Logic no GIS	**91%**	97%	86%	76%	--	--	--	--	97%	--
Tsui (method 2a) (2006)	Fuzzy Logic with GIS	**91%**	98%	72%	80%	88%	--	--	--	97%	--
Tsui (method 2b) (2006)	Fuzzy Logic with GIS	**94%**	98%	86%	80%	88%	--	--	--	99%	--

*Underground transfer
**Not Reported

PART THREE

EVALUATION OF GPS MODE DETECTION STRATEGIES

Introduction

Comparing the methodologies for effectiveness, given the variation in the data components and mode choice opportunities, creates some uncertainty with respect to the best-performing strategy. There are also concerns where only the overall effectiveness is reported (e.g., Stopher, FitzGerald et al. 2008b). In order to control for the variation in datasets, an experiment was performed using a single dataset applied to a rule-based method, a neural networks method and a fuzzy logic method.

Data Collection and Preparation

The first step for any study that relies upon GPS trace data is to select the appropriate unit for gathering the data. A review of a variety of equipment (GPS loggers) and an extensive test of two of them by Lawson, Chen et al. (2008) found the high-quality chipset in the I-Blue 747 received satellite signals and logged location points successfully, except in underground subways and tunnels. The demonstrated accuracy of the I-Blue 747 was from 2.5 to 3 meters, although the accuracy was affected by the urban canyon effect in downtown and midtown Manhattan. One of the outstanding aspects of the I-Blue 747 was its ability to provide many data fields, (e.g., number of satellites used and horizontal dilution of precision). These fields can be used to remove low quality of location points.

The second step is to recruit survey participants. The principal criterion for sample selection is to ensure complete coverage of all modes, not necessarily requiring a randomly sampled population. For this study, 24 individuals from the City University of New York were recruited. They were asked to carry an on-body GPS unit for a period of five weekdays. Each GPS unit was configured to automatically log the person’s position every 5 seconds, along with the 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 at all times. Additionally, survey participants were also asked to fill out a travel log for one designated travel survey day. Data collected in the travel log included each person’s contact information, auto ownership, frequently visited locations, and all the activities and trips made during that day. For each activity and trip, detailed information was requested, (e.g., the location name, exact address or intersection, mode taken to access the activity site, departure and arrival time, and purpose of the trip).

The primary purpose of requiring survey participants to fill out a travel log was to generate a validation set, which contained the self-reported travel modes taken for each trip. Special post-processing of the travel log was carried out to verify the reported travel patterns; if unverified, the travel diary may be prone to errors. Close visual inspection of all completed travel logs was performed to identify problems (e.g., missed trips and illogical mode sequences, etc.), as complete and correct information was needed for this research.

GPS points of poor quality were removed from the data set in a preprocessing step. To clean the data, a four-layer data filtering process was utilized, based on the work of Tsui and Shalaby (2006). This approach was designed to systematically remove points with significantly high error. For example, the first layer removes GPS points with fewer than three satellites in view; then the second layer removes records that have a horizontal dilution of precision (HDOP) high than five. This ensures that the satellites were well dispersed. To further ensure that the speed and heading information was accurate, records that have a zero heading and speed were also removed. The fourth filter, originally developed by Chung and Shalaby (2005), removed GPS points that were likely to be erroneous due to the "urban canyon effect", which caused fictitious jumps in the GPS trace data. Since the experimental dataset was collected in a highly concentrated setting in the New York City metropolitan area, it was especially important to remove these points. Therefore, GPS records that exhibit illogical jumps were also removed.

After preprocessing the data to remove inaccurate GPS locations, an additional preprocessing step was performed to segment the original GPS data into individual trips. A trip end was defined in this study as occurring any time there was no movement for a period of 120 seconds or more (Stopher, Clifford et al. 2008a). Software was written using Visual Basic for Applications (VBA) in ArcGIS 9.3, to identify the trip ends. To determine whether or not there was movement, each point was buffered by 50 meters. If no GPS point was outside the buffer within a 120 second duration, then the centroid of the points within the buffer was identified as the trip end.

After uploading the survey participant's GPS trace data and running the data cleaning and trip segmentation steps, each trip was compared directly against the travel log. Not all trips identified from the GPS unit were present in the travel log; conversely, not all trips in the travel log were present in the GPS data. These inconsistencies may have been due to a failure of the participants to record trips in the travel log, or there was imprecise information derived from the GPS units. GPS errors also result from the urban canyon effect, cold-start problems, or the individual not remaining at a destination for a prolonged period of time (at least 120 seconds), as is sometimes the case when picking up a cup of coffee or fast food.

A total of 101 trips were identified that accurately matched information written in the travel logs. Other trips were omitted due to the inability to determine the correct travel mode based on travel logs and manual inspection of the GPS data. Bialostozky (2009) also used this dataset in his work. Figure 1 provides an example of the data features: the GPS trace, the reported frequently visited locations and the locations recorded in the logs.

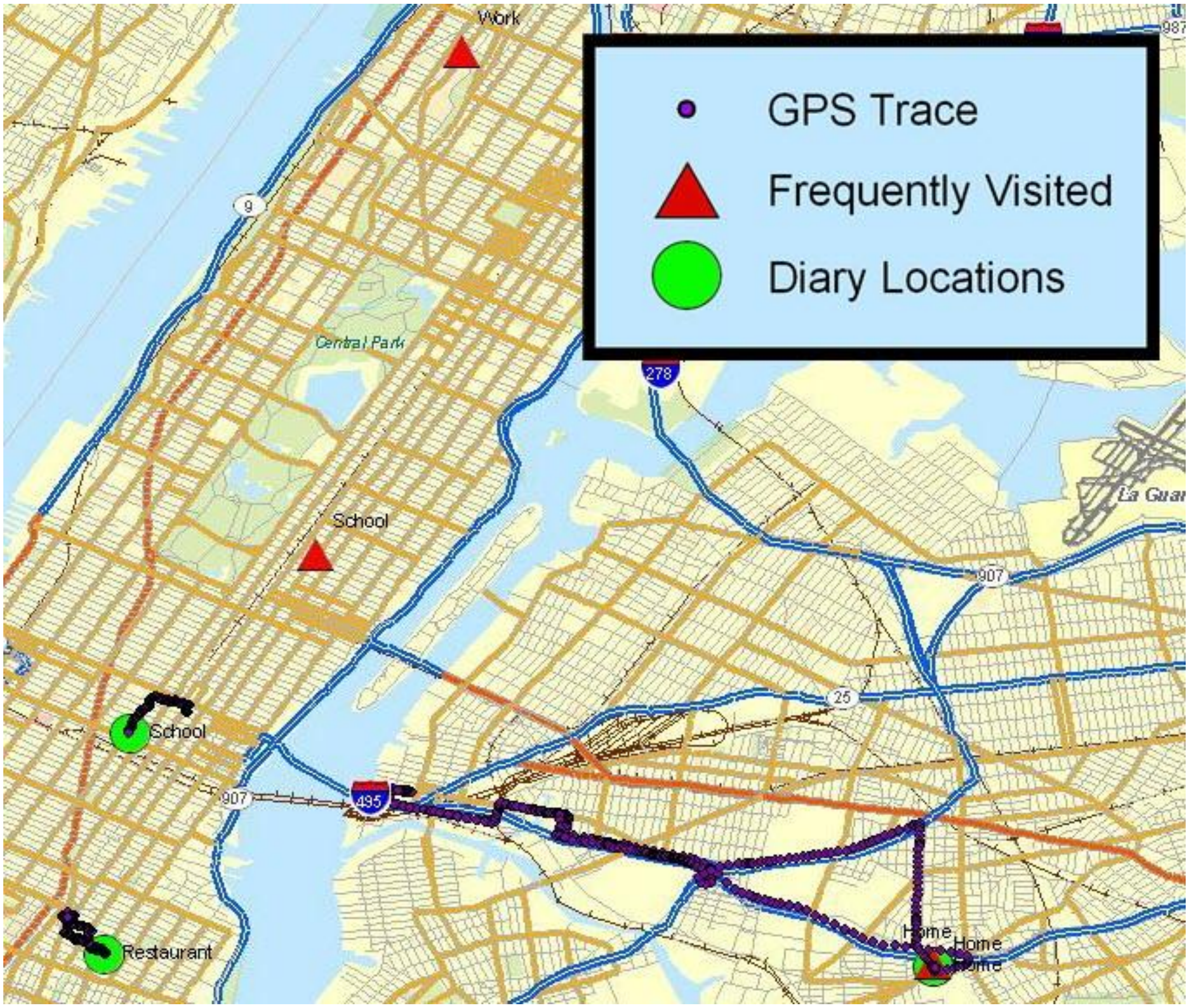

Figure 1. Display of data elements collected

Method One: Rule-based GIS Approach

The rule-based approach was primarily based on the work of Bialostozky (2009) and relied on a hierarchical-rule based detection scheme that depends heavily on data derived from GIS. GPS traces (shown in Figure 2) were processed in GIS software to detect travel patterns. Once trip segments were identified, the developed algorithm attempted to detect the mode used on each segment, according to a set of criteria based on the characteristics of each mode. The algorithm first attempted to identify walking segments. This was because walking segment identification was also essential in trip segmentation. Detection of the walking mode relied on three primary characteristics: a) a pedestrian must walk on a pedestrian-accessible link, which requires map-matching; b) the travel time must be longer than 60s (Schuessler and Axhausen 2009); and c) the walking speed cannot exceed 10 km/hr (Stopher, FitzGerald et al. 2008b). After walking segments were identified, the algorithm proceeded to determine if any of the remaining segments were 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 auto and bus. Subway or rail trips must occur on subway or rail links; the change of mode must occur near a subway stop or a rail stop; and some parts of the trips may be underground and thus received no GPS signals.

The primary criterion to distinguish between auto and bus trips was that buses travel on bus routes and stop at bus stops. However, the latter criterion can be complicated by congestion and signal delays, during which the bus will stop. In fact, these slow-moving and stopping bus segments were likely to be classified as walking 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 walking segment is less than 5 minutes long and then combines them into one street mode (bus or auto). This criterion was applied based on the reality that a trip that involves both auto and bus modes separated by a brief walking segment were rare -- but could occur in New York City, indicating a transfer from one bus to another.

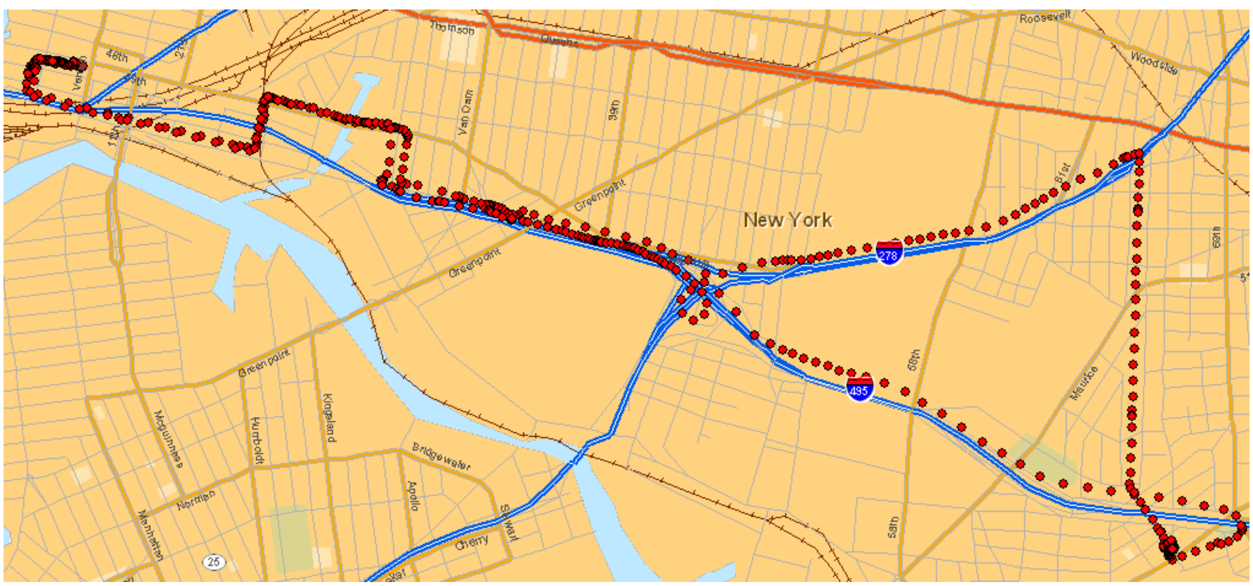

Figure 2: GPS trace (red) in GIS software

Method One: Results

Among the five modes, the auto mode achieved the highest success rate, 96% (*see Table 3*). Trip segments made by walking also had a high success rate; 92%, or 161 out of 176 trip segments were correctly identified. Subway and bus, serving similar trips in terms of trip distance, received similar success rates, 68% and 53% respectively. Rail achieved the lowest success rate, only 29%. Out of 14 rail segments, only 4 were correctly diagnosed. When an underground subway trip segment immediately followed or preceded an underground rail segment (e.g., Grand Central Station), no signal was received during the entire course during which a mode transfer was made. These segments were designated as "Und. T", to indicate "underground transfer." Using GPS traces collected near the subway and rail stations when a subject entered or exited from a station, the developed GPS algorithm was able to identify 83% of these underground transfer segments.

Method Two: Neural Networks

To test the effectiveness of a neural network on detecting mode choices, a neural network was implemented to replicate research previously described by Gonzalez et al. (2008). The Multi-Layer Perception neural network was utilized using open-sourced software call Weka (Witten, Frank et al. 1999), commonly used in data mining applications. The input layer consists of four input parameters:

Input Parameters:

- Average Speed
- Maximum Speed
- Estimated Horizontal Accuracy Uncertainty
- Distance between the start and end of trip

The estimated horizontal accuracy uncertainty is a measurement of the estimated confidence of a calculated GPS position. Different modes of transportation have different values for estimated horizontal accuracy uncertainty. For example, a subway has the worst estimated horizontal accuracy uncertainty since it is underground and therefore obstructs the GPS signals. On the other hand, a walking trip has the best estimated horizontal accuracy uncertainty since typical GPS signal obstruction is most likely caused only by clothing, bag or purse.

The input parameters differ slightly from the variables outlined in Gonzalez, Weinstein et al. (2008), to reflect the complex, urban setting of the experimental dataset. Gonzalez, Weinstein et al. (2008) included two additional variables, the standard deviation of the distances between stop locations and the average dwell time. When these same variables were used with the experimental dataset, the results did not yield a higher success rate. These values were included primarily to distinguish between auto and bus modes. However, buses and autos were found to follow similar stop and go patterns in New York City, due to the high volume of traffic. Therefore, the neural network was not able to accurately distinguish between these two modes. To address this, an additional variable of distance between the start and end of the trip was found to be effective in determining the mode, since individuals tend to take autos and subway/rail farther than buses. A depiction of the neural network used is shown in Figure 3.

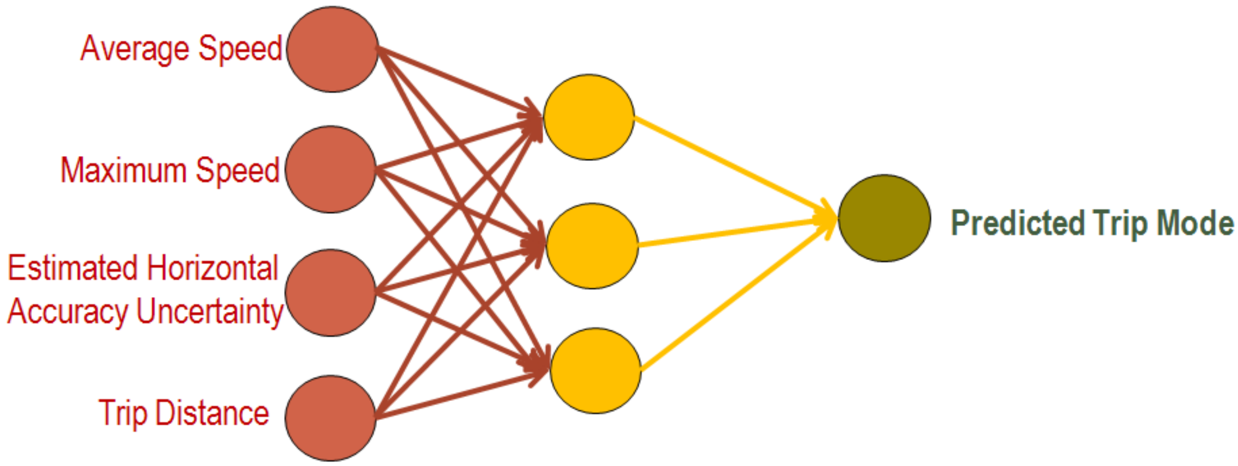

Figure 3: Depiction of the neural network used to predict travel mode

To train the neural network, a file was loaded into the Multi-Layer Perception neural network that contained the input parameters of each GPS trip, along with the correct answer. A 10-fold cross validation technique was used to assign weights and test the effectiveness of the network. The N-fold cross-validation randomly partitions the data into N-folds. Each fold served as the testing data set, while the remaining N-1 folds were used as the training data set. The process was repeated N times – once for each fold (Kohavi 1995). The result was the average success rate from the model from the N iterations.

Two input parameters, learning rate and training time, were varied to determine the effect on resulting accuracy in predicting trip mode. The learning rate corresponds to the degree of change occurring in the weights between nodes in the neural network. Possible values for the learning rate range between zero and one, with zero indicating no change to the edge weights during the training iterations, and one indicating complete change. The greater the learning rate, the greater the degree of change during the training process. The training time was the number of times the entire training dataset was used to adjust the weights in the neural network. A single iteration of the entire training dataset was considered one epoch. A large training time (number of epochs) corresponded to more processing time and usually more accurate results.

Method Two: Results

The overall success rate of predicting mode choice was very similar to the approach described in Gonzalez, Weinstein et al. (2008). Using input parameters of 0.3 as the learning rate and a training time of 500 epochs, the overall success rate of our implementation was 84% compared to 85.4% in the Gonzalez research, using the same parameters. Some input parameters hindered the overall success rate; when using a 0.1 learning rate and a training time of 300 epochs, the overall performance was 82% compared to 88.6% for Gonzalez. The consistency of performance between the two implementations provides additional evidence that the neural network technique was replicable in a distinctly different geographic location. However, some drawbacks were apparent, as shown in Table 4. For instance, the success rate for auto and bus trips was typically around 60%, and depending on the neural network training parameters, the results could be much lower. Typically, the neural network confused auto and bus trips, since the input parameters followed a similar pattern. Walking was the most frequent travel mode in our dataset, consisting of 60% of the trips in the dataset. Walking was also the most easily distinguishable travel mode, while subway trips were detected with the second highest success rate.

Method Three: Fuzzy Logic

Research has been conducted in using fuzzy logic to automatically determine travel modes based on GPS trajectory data (Tsui and Shalaby 2006; Schuessler and Axhausen 2009). To define this logic, a series of input variables are described and modeled in terms of their fuzzy logic values. Attributes determined from each GPS trip are used to predict the travel mode. The Schuessler-Axhausen algorithm used three input variables:

Schuessler-Axhausen input variables:

- Median Speed
- 95th Percentile Speed
- 95th Percentile Acceleration

A trapezoidal membership function represented the different levels of the fuzzy variables. This function was defined by four control points, the starting point, the left top corner, the right top corner and the end point (Upadhyay, Schuessler et al. 2008), and is shown in Figure 4. Each of the input variables were described using three or four membership functions. Once the variables were described, using the membership functions, rules were used to predict the travel mode.

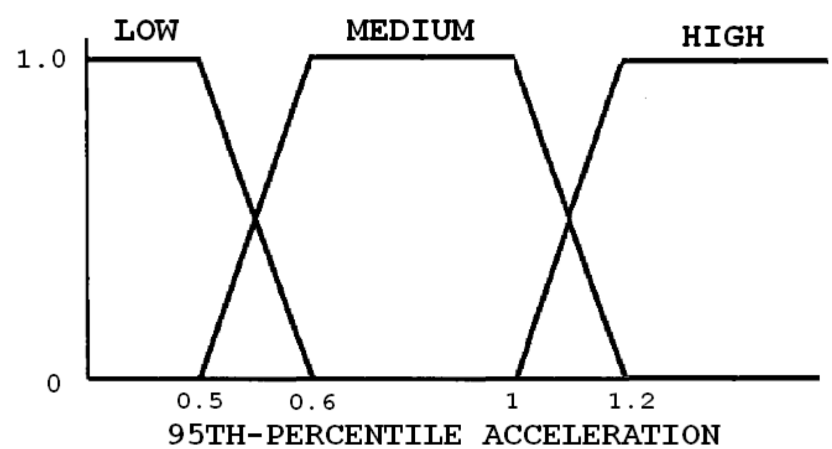

Figure 4: Example of a fuzzy logic rule to classify acceleration

The fuzzy logic technique, described in the Schuessler-Axhausen algorithm, was used and compared, using the experimental dataset. The input variables and membership functions were the same as those described by Schuessler and Axhausen (2009). Some of the rules needed to be modified to reflect the standardized dataset, however. Since the experimental dataset contained no bike trips, rules representing biking outcomes were changed to walking.

Method Three: Results

The fuzzy logic rules appear to work reasonably well in detecting the walking trips, as indicated with a success rate of 91%. However, results in detecting subway trips were very poor, with only a 26% success rate. Furthermore, due to the similar driving behavior of autos and buses in New York City (in terms of speed and acceleration), the fuzzy logic rules could not accurately distinguish between the two modes (*see Fuzzy Logic, method 1 - Table 4*). When combining autos and buses into a single mode called "vehicle", the success rate was 85% (*see Fuzzy Logic, method 2 - Table 4*).

Comparative Results

The mode detection algorithms produced various success rates for detecting different modes. In some respects, the most effective algorithm was the neural network approach (method 1), with an overall success rate of 84% and the ability to distinguish between four different modes. This is slightly higher than the overall success rate of the rule-based GIS approach (79%). However, the rule-based algorithm had the ability to detect a larger number of mode choices, such as rail and underground transfers. The fuzzy logic implementation had the highest success rate, but was unable to distinguish between bus and auto, and had a very low subway success rate. For individual modes, the neural networks (method 1) had the highest success rate for walking, bus, and subway, but a lower success rate for autos. Additional metrics beyond simply measuring the overall success rate are needed accurately rate the effectiveness of each approach. These metrics are discussed in the following section.

Table 4: Success Rate of Mode Detection Algorithms

Method	**Overall**	**Walk**	**Bus**	**Auto**	**Subway**	**Rail**	**Und T***
Rule-based w/ GIS	**79%**	92%	54%	96%	68%	29%	83%
Neural Networks (method 1)	**84%**	97%	60%	59%	88%	--	--
Neural Networks (method 2)	**82%**	96.6%	23.4%	70.6%	88.0%	--	--
Fuzzy Logic (method 1)	**67%**	86%	21%	53%	17%	--	--
Fuzzy Logic (method 2)	**87%**	91%	85%**		26%	--	--

*Underground transfer

** Vehicle

Discussion

In addition to the success rates listed in Table 4, there are other trade-offs regarding the ease-of-use for each approach. Data-intensive approaches, (e.g., Bialostozky 2009), could prevent wide-spread adoption of this methodology. This approach requires extensive data pre-processing, using geographic information systems (GIS), and the development of a complex series of rules, fine-tuned to fit a particular dataset. These rules could easily fail to achieve high success rates when applied to another geographic region. In addition, geographic features need to be included in the processing (e.g., stops and routes of all buses, subways and rail locations). In contrast, the neural network approach, outlined by Gonzalez, Weinstein et al. (2008) and replicated in this research, does not require extensive data preparation. Instead, a substantial training dataset, that is representative of the input data, is required.

In this study, neural networks were shown to be the most replicable method implemented based on ease-of-use and overall effectiveness. A similar success rate to Gonzalez, Weinstein et al. (2008) was obtained simply by extracting the variables, training the network, and running the neural network software called *Weka*. The entire process took a mere two days. In contrast, implementation of the rule-based GIS method, outlined in Bialostozky (2009) and similar to Clifford, Zhang et al. (2008), took numerous researchers months of labor and data-intensive work to develop a series of effective rules. To replicate the GIS-intensive results in another region would require extensive data collection and pre-processing.

The success of neural networks in predicting travel mode, along with the methods ease-of-use, emphasizes the need for other machine learning approaches to be applied to this area of research. Machine learning techniques have been shown to be applicable across a wide variety of datasets, by using automated training methods. Future research will need to consider a hybrid approach to providing a methodology that simultaneously reduces respondent burden, processing burden, and resource consumption.

Conclusions

Determining travel mode is important in a variety of transportation planning applications, particularly transportation models. Traditionally, this data has been collected in travel diaries and retrieved by telephone, requiring survey participants to remember and report their travel behavior. Recent work suggests that most of these data could be collected by replacing or supplementing traditional survey methods with passive data collection using GPS trace data. By leveraging GPS technology, the accuracy of transportation models could be improved by eliminating human error in reporting. While research has been conducted to facilitate automatic determination of travel mode, results were mixed and algorithm replication was rare.

This study replicated three algorithms, using a dataset collected in the complex urban environment of New York City. By replicating these algorithms, our results indicate that methodologies that rely heavily on pre-defined rules were often ineffective in achieving high success rate on different datasets. Machine learning techniques, such as neural networks, could often achieve high success rates across diverse datasets. These approaches avoid the manually-intensive efforts often associated with data-centric algorithms that rely heavily on geographic variables and combinations of spatial datasets in order to detect modes from GPS trace data.

To implement the use of GPS trace data as a substitute for traditional or modified travel surveys -- moving from small sample feasibility studies to agency-level implementation -- there needs to be sufficient evidence that GPS trace data will meet the recognized challenges. It will be necessary to prove that GPS trace data can be handled with ease, and that the promise of reduced respondent and processing burden, can be realized.

Acknowledgements

Special thanks to Haiyun Lin, Jonathan Muckell, and Monica Rodriguez for their assistance on this project. This work is supported in part by the Research and Innovation Technology Administration of the U. S. Department of Transportation through the Region 2 – University Transportation Research Centers (UTRC) Program.

References

Bialostozky, E. (2009). Development of a Mode Detection Algorithm for GPS-Based Personal Travel Surveys in New York City. Presentation materials for the New York Metropolitan Transportation Council.

Bohte, W. and K. Maat. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C* **17**: 285-197.

Bricka, S. (2009). What is Different About Non-Response in GPS-Aided Surveys? In P. Bonnel, M. Lee-Gosselin, J. Zmud and J. Madre (Eds.), *Transport Survey Methods: Keeping up with a Changing World*. United Kingdom, Emerald Group Publishing Limited: 403-427.

Bricka, S. and C. Bhat. (2006). A comparative analysis of GPS-based and travel survey-based data. *Transportation Research Record* 1972: 9-20.

Chung, E. and A. Shalaby. (2005). A trips reconstruction tool for GPS-based Personal Travel Surveys. *Transportation Planning and Technology* 28(5): 381-401.

Clifford, E., J. Zhang, and P. Stopher. (2008). Determining trip information using GPS data. *Working Paper (Institute of Transport and Logistics Studies); ITLS-WP-08-01*.

de Jong, R., and W. Mensonides. (2003). *Wearable GPS Device as a Data Collection Method for Travel Research*, University of Sydney and Monash University.

Flamm, M., C. Jemelin, and V. Kaufman. (2007). Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions. A paper presented at the 7th Swiss Transport Research Conference in Monte Verita/Ascona, Switzerland on September 12 - 14, 2007.

Gilani, H. (2005). Automatically Determining Route and Mode of Transport Using a GPS Enabled Phone. An unpublished paper.

Gonzalez, P., J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi, and R. Perez. (2008). Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-Enabled Mobile Phones. A paper presented at the 15th World Congress on Intelligent Transportation Systems. New York, New York.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International joint conference on artificial intelligence* 14(2): 1137-1145.

Lawson, C., C. Chen, H. Gong, S. Karthikeyan, A. Kornhauser, J. Cepler, M. Chau, H. Lin, and S. Dziekan. (2008). *GPS Pilot Study Phase Three: "Mixed Mode" Data Collection and Identification of Special Population Segment*. Prepared for the New York Metropolitan Transportation Council.

Lawson, C., C. Fassman, and M. Chau. (2007). *Household Travel Survey Research.* Prepared for the New York Metropolitan Transportation Council, New York, New York.

Perdok, J., G. Draijer, and N. Kalfs. (1998). Possible Application of Global Positioning System (GPS) of Collecting Travel Data. *European Transport Conference.*

Schuessler, N. and K. W. Axhausen. (2009). Processing Raw Data from Global Positioning Systems Without Additional Information, *Transportation Research Record: Journal of the Transportation Research Board*, Transportation Research Board of the National Academies.

Stopher, P. (2009). The Travel Survey Toolkit: Where to From Here? In P. Bonnel, M. Lee-Gosselin, J. Zmud and J. Madre (Eds.), *Transport Survey Methods: Keeping Up With a Changing World.* Emerald Group Publishing Limited**:** 15-46.

Stopher, P., E. Clifford, J. Zhang, and C. FitzGerald. (2008a). Deducing mode and purpose from GPS data. *ITLS - WP-08-06*.

Stopher, P., C. FitzGerald, and J. Zhang. (2008b). Search for a global positioning system device to measure person travel. *Transportation Research Part C* 16: 350-369.

Stopher, P. and S. Greaves. (2006). Where are we going? *Transportation Research Part A*(41): 367-381.

Transportation Research Board. (2007). *Metropolitan Travel Forecasting: Current Practice and Future Direction, Special Report 288*, Washington, D. C.: National Academy of Sciences.

Tsui, S. Y. A. and A. S. Shalaby. (2006). Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board* 1972: 38-45.

Upadhyay, D., N. Schuessler, K. W. Axhausen, M. Flamm and V. Kaufmann. (2008). *Optimal parameter values for GPS post-processing: An experiment*, ETH Zurich, Zurich.

Witten, I. H., E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham. (1999). Weka: Practical machine learning tools and techniques with Java implementations, *ICONIP/ANZIIS/ANNES'99 International Workshop*. H. Kasabov and K. Ko, eds., Dunedin.

Zheng, Y., L. Liu, L. Wang, and X. Xie. (2008). Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. Proceeding of the 17th international conference on World Wide Web, Beijing, China, Association for Computing Machinery.