A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys

Yijing Lu[1], Shanjiang Zhu[2], and Lei Zhang[3,*]
1. Graduate Research Assistant
2. Research Scientist
2. Assistant Professor (*Corresponding Author)
Department of Civil and Environmental Engineering, University of Maryland
1173 Glenn Martin Hall, College Park, MD 20742
301-405-2881, lei@umd.edu

1. Introduction

The principal data sources in travel behavior research are travel surveys that collect a series of trip-making characteristics including trip purpose, travel mode, trip length, origin and destination, time of day, week and month, and trip companions. The accuracy, completeness and timeliness of travel survey data also play an essential role in travel demand modeling. Travel researchers often require temporal-spatial travel data as accurate as possible. Meanwhile, they are faced with increasing difficulties in travel survey methods such as low response rates and underreported trips. Thanks to the new technologies which could help acquire travel data automatically, accurately and constantly, difficulties could be mitigated. GPS/GIS technology tracking individual travel behavior and offering detailed spatial structures of individual travel has obtained increasing interest and attention in travel data collection and travel behavior research. Compared with the conventional travel survey methods which collect individual travel data through mail or phone, GPS-based survey methods are able to improve data quality and provide researchers with more detailed and more accurate travel information for longer periods. Unlike conventional travel survey methods, GPS-based surveys record the individual's travel paths by a series of time, geo-coded points with corresponding variables such as time of day, week and month, speed, latitude, and longitude. Combined with GIS data and transportation network information, individual's travel mode, trip purposes and trip length could be derived, therefore, the GPS-based survey method could reduce respondent burden in the travel data collection and increase the response rate.

Along with the rapid development of technology, utilizing GPS technology in travel survey methods, supplementing or replacing the conventional travel data collection, is a trend. However, the biggest challenge of successfully utilizing GPS-based data is to develop efficient tools for data post-processing which could extract individual travel information including trip purpose, travel model, and trip length from the GPS raw data. Furthermore, as the GPS technology generates a large quantity of raw data, automated procedures for efficiently post-processing GPS data with low computation cost will become necessary and essential.

An increasing number of travel researchers are exploring GPS-based travel survey methods, and different methods have been developed to derive the trip purpose. Wolf et al. (2001) pioneered the procedures of trip purpose detection based on a set of deterministic rules and a sample of 19

respondents who both successfully collected travel data with the GPS data logger and returned a completed a paper trip diary and demonstrated the possibility of detecting trip purposes in Atlanta, Georgia, given a detailed GIS database of land use. Schönfelder et al. (2003) in Europe further developed the procedures. They used multi-stage hierarchical matching procedure, calculating a cluster center of stop ends by combining trip ends, identifying trips with obvious purposes, and establishing relationships between trip purposes and activity temporal information as well as the socio-demographics of the respondents. Stopher et al. (2008) presented a set of heuristic rules to derive trip purpose of 43 trips collected in Sydney with the help of not only the parcel-level land use data but also the geo-coded addresses of the respondent's workplace or school, and the two most frequently used grocery stores. Bohte et al (2008) developed a GPS-based travel data collection method combining GPS devices, GIS technology and a web-based validation procedure, and derived the trip purposes based on the heuristic rules. Chen (2010) followed Schönfelder's approach to cluster trip ends into activity locations, employing deterministic rules to derive trip purpose for low-density area and the Multinomial Logit model for trips in high-density area.

The method of deriving trip purpose based on GPS/GIS-based data was further explored with artificial intelligence or machine learning. Griffin et al. (2008) constructed a decision tree to derive trip purposes, and the procedure was implemented in the C4.5 environment with 50 randomly generated trips which are simulated following a series of assumptions. Deng et al. (2010) employed a number of attributes to construct a decision tree to derive the travel models and trip purposes. The decision tree is implemented in the C5.0 machine learning environment with a homogenous set of 226 GPS trip records collected from 36 respondents in Shanghai.

In this paper, we explore the feasibility of using machine learning method to automatically derive trip purpose based on in-vehicle GPS data collected by University of Minnesota, and examine the effects of different categories of input variables, different land use coding methods and different trip purpose categorizations on the trip purpose detection.

2. Methodology

The GPS/GIS-based trip purpose detection system is illustrated in Figure 1. Three dashed boxes represent input module, learning process module, and output module. GPS-based data, GPS-based travel recall survey, and GIS data form the input module. Learning process module which employs machine learning method is the core of trip purpose detection. It consists of trip purpose estimation and decision tree pruning. Once the trip purposes are derived based on the machine learning method, they are forwarded to validation part to evaluate the classifier performance.

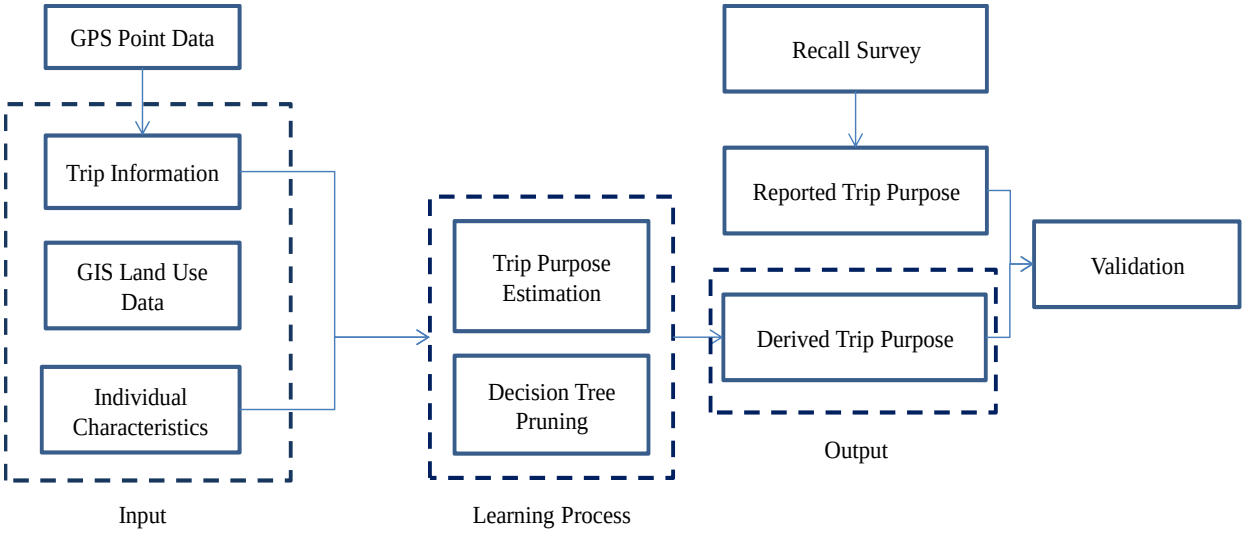

Figure 1 Trip Purpose Imputation Based on GPS/GIS Data and Machine Learning Methods

Learning Process and Validation

The decision tree is constructed and employed to automate the trip purpose detection. The input attributes include individual's trip characteristics derived from the GPS data such as trip start/end time, trip destination location, and activity duration, GIS-based land use type, as well as individual's social-demographic attributes.

The widely used decision tree algorithm in practice is C4.5 introduced by J. Ross Quinlan in 1993, which employs the information gain to split each node, choosing the attribute at each node that produces the purest daughter node to split on. The information is a measurement of purity. The daughter nodes in the sub-tree will be split based on the same procedure, until all the instances at a node reach the same classification.

Pruning a decision tree is a technique that reduces the size of the tree by cutting off some nodes from the tree which have litter power in instances classification. Employing pruning in decision tree model could improve the computational efficiency and accuracy, reduce the complexity of the tree and avoid the problem of the data set over-fitting. The pruning methods applied to the trip purpose decision tree in the research are post-pruning and on-line pruning. Once the decision tree is constructed and pruned, the 10-fold cross-validation is employed to estimate the error rate of machine learning technique.

Input Attributes

The input variables used to derive trip purpose can be divided into three categories and five subcategories (Table 1).

Table 1 Input Variables of Trip Purpose Detection by Category

Category	Subcategory	Variables
GPS Trip Attributes	Trip Attributes	Current Trip Start Time
		Current Trip End Time
		Current Activity Duration
		Day of Week (Weekday or Weekend)
		Trip Type
	Next & Previous Trip Attributes	Previous Trip Start Time
		Previous Trip End Time
		Previous Activity Duration
		Next Trip Start Time
		Next Trip End Time
		Next Activity Duration
Respondent's Characteristics	Respondent's Social Demographic	Respondent Income Level
		Respondent Education Degree
		Respondent Race
		Respondent Age
Land Use Data	Parcel-level Land Use Type	Land Use Type of Current Trip Destination
		Land Use Type of Previous Trip Destination
		Land Use Type of Next Trip Destination
	Trip End Location	Locations of Current, Previous and Next Trips Destination (Home, Work, Other Place)

3. Data

GPS-based data used in this study were collected from a 13-week long study targeting behavioral reactions to the I-35W Bridge reopening on September 18th, 2008. Details about this behavioral study and data collection process are discussed in Zhu et al. (2010). Participants were randomly selected commuters in Minneapolis, Saint Paul, and Minnesota metropolitan area (Twin Cities). The vehicle trajectories were divided into separate trips based on the engine-on and engine-off events. The origin and destination of each trip can then be identified as the first point and last point along the trajectory of each trip. Land use data used in this study is the 2005 Generalized Land Use dataset for the seven counties of Twin Cities Metropolitan area in Minnesota developed by the Metropolitan Council. The area was divided into different polygons with most of them consistent with street block boundary.

To complement the GPS-based survey, an online travel diary survey was conducted once a week during the study period. An email, sent to each participant at the end of a randomly select day (one day a week), invited them to visit our survey website to complete the survey.

The GPS raw data is temporal-spatial track point data which needs to be processed and transformed into trip information such as trip start/end time and activity duration for the detected trip purpose. It is hypothesized that the trip ends land use data and its coding method are critical for the quality of trip purpose detection. In the research, GIS data used to derive trip purposes

mainly consists of the multiple parcel-level land use database and respondent's home and work geo-coded addresses. Trip end land use deployment can be executed in two steps: first, any trip end falls within a land use parcel will be assigned the corresponding land use type; second, if a trip destination within a buffer of 500m from home or work address, the trip end will be assumed the home or work location.

4. Model Estimation and Conclusion

Specifically, the trip purposes are coded into 10 categories (home, work, shopping, daycare, dining, driving others, services, school, social/recreation, and other) and decoded into five categories including Home-base Work, Home-base Shopping, Home-base Social/Recreation, Other home-base, and Not Home-base based trip.

The data preparation procedure yields a sample of 3188 trips for the trip purpose decision tree learning. Furthermore, sensitivity analysis is carried out to assist us in analyzing the essential role of trip end land use coding method in deriving trip purpose and the contributions of various input variables to the trip purpose detection. Different subcategories of input variables are separately added into the trip purpose decision tree gradually. The recursive procedure of trip purpose identification and the corresponding classifier performance are illustrated in Figure 2. The final classifier with all the three input modules reaches the highest accuracy rate of 73.37%. Module 2 with Next & Previous Trip Attributes and Module 3 with Trip End Location can have classifier performance 60.57% and 72.30% separately. Moreover, the data is also trained and tested in C4.5 environment based on the reported 10 categories of trip purposes. All the same input attributes as those in the final 5-trip-purpose model are incorporated into the learning process and 62.77% of classification accuracy is given.

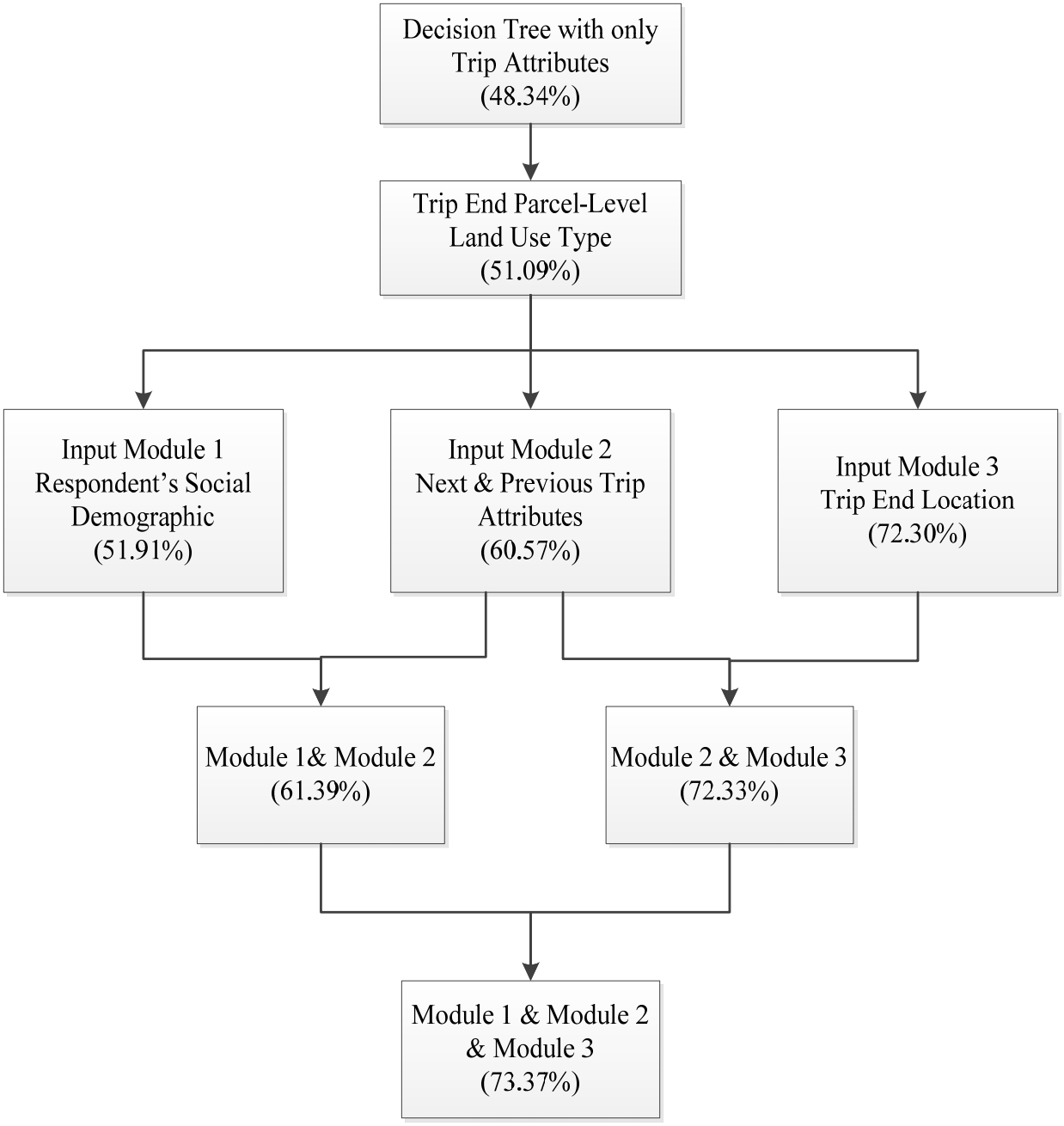

Figure 2 Trip Purpose Detection Procedure and Accuracy Results

The results indicated that the four subcategories of input attributes except for the current trip information have different contributions to trip purpose classification. Respondent's social-demographic variables, which can only improve the classifier performance less than 1%, have little power in trip purpose detection, whereas next and previous trip information can enhance the trip purpose classification accuracy by more than 10%, which proves the hypothesis that people usually make their trips strategically. Among all the four subcategories of input attributes, the trip end locations have the strongest impact on trip purpose classification with an improvement of the classifier accuracy by more than 20%. Moreover, the comparison of the performance rates

between the 10-trip-purpose categorization and the 5-trip-purpose categorization shows that the complex typology with more classes can deteriorate the classification accuracy by more than 10%. Mixed-use land type containing multiple units in the area such as restaurants, commercial shops, and childcare facility is a challenge for trip purpose derivation. POI (Point of Interest) offering geo-coded locations of business units, services units and etc. and being used as an improved land use coding method is expected to be much helpful for trip purpose detection.

Acknowledgments

The author would like to thank Drs. David Levinson, Henry Liu, and Kathleen Harder at the University of Minnesota for providing the GPS dataset. The authors are solely responsible for all statements in the paper.

References

Axhausen, KW.,Schonfelder, S., Wolf, J., Oliveira, M., Samaga, U., 2003. 80 weeks of GPS-traces: approaches to enriching the trip information. Transportation Research Record, 1870, 46 - 54

Bohte, W., Maat, K., 2009. Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transportation Research Part C 17, 285–297

Chen, C., Gong, H., b, Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A 44, 830–840

Deng, Z., Ji, M., Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. 2010. Traffic and Transportation Studies. p.768-777

Gonzalez, A.P., Weinstein, S.J., Barbeau, J.S, Labrador, A.M., Winters, L.P., Georggi, L.N., Perez, R., Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. 2008. 15th World Congress on Intelligent Transportation Systems, New York, NY. Paper # 30267

Griffin, T., Huang, Y., 2005. A Decision Tree Based Classification Model to Automate Trip Purpose Derivation. In the Proceedings of the 18th International Conference on Computer Applications in Industry and Engineering, Honolulu, Hawaii

Schönfelder, S., K. Axhausen, N. Antille, M. Bierlaire, and E. Lausanne. 2002 . Exploring the potentials of automatically collected GPS data for travel behavior analysis - a Swedish data source. GI-Technologien für Verkehr und Logistik 13, 155-179.

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

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

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

Wolf, J., R. Guensler, and W. Bachman (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transportation Research Record: Journal of the Transportation Research Board 1768 (1), 125-134.

Zhu, S., D. Levinson, and H. Liu. 2010. Measuring Winners and Losers from the new I-35W Mississippi River Bridge. In The 89th Annual Conference of Transportation Research Board, Washington D.C.