LOGISTIC REGRESSION MODELS FOR PREDICTING TRIP REPORTING ACCURACY IN GPS-ENHANCED HOUSEHOLD TRAVEL SURVEYS

A Thesis

by

TIMOTHY LEE FORREST

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2005

Major Subject: Geography

LOGISTIC REGRESSION MODELS FOR PREDICTING TRIP REPORTING ACCURACY IN GPS-ENHANCED HOUSEHOLD TRAVEL SURVEYS

A Thesis

by

TIMOTHY LEE FORREST

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Andrew Klein
Committee Members,	Dennis Perkinson
	Daniel Sui
Head of Department,	Douglas Sherman

December 2005

Major Subject: Geography

ABSTRACT

Logistic Regression Models for Predicting Trip
Reporting Accuracy in GPS-Enhanced Household
Travel Surveys. (December 2005)
Timothy Lee Forrest, B.S., Texas A&M University
Chair of Advisory Committee: Dr. Andrew Klein

This thesis presents a methodology for conducting logistic regression modeling of trip and household information obtained from household travel surveys and vehicle trip information obtained from global positioning systems (GPS) to better understand the trip underreporting that occurs. The methodology presented here builds on previous research by adding additional variables to the logistic regression model that might be significant in contributing to underreporting, specifically, trip purpose. Understanding the trip purpose is crucial in transportation planning because many of the transportation models used today are based on the number of trips in a given area by the purpose of a trip.

The methodology used here was applied to two study areas in Texas, Laredo and Tyler-Longview. In these two study areas, household travel survey data and GPS-based vehicle tracking data was collected over a 24-hour period for 254 households and 388 vehicles. From these 254 households, a total of 2,795 trips were made, averaging 11.0 trips per household. By comparing the trips reported in the household travel survey with

those recorded by the GPS unit, trips not reported in the household travel survey were identified.

Logistic regression was shown to be effective in determining which household- and trip-related variables significantly contributed to the likelihood of a trip being reported. Although different variables were identified as significant in each of the models tested, one variable was found to be significant in all of them – trip purpose. It was also found that the household residence type and the use of household vehicles for commercial purposes did not significantly affect reporting rates in any of the models tested. The results shown here support the need for modeling trips by trip purpose, but also indicate that, from urban area to urban area, there are different factors contributing to the level of underreporting that occurs. An analysis of additional significant variables in each urban area found combinations that yielded trip reporting rates of 0%. Similar to the results of Zmud and Wolf (2003), trip duration and the number of vehicles available were also found to be significant in a full model encompassing both study areas.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In the State of Texas, travel surveys administered through the Department of Transportation have been in existence for over fifty years. During that time, their primary purpose has been to assist transportation modelers and planners in urban areas around the state. Various forms of travel data have been collected to develop urban travel demand models. As defined by Pearson and Dresser (*1*), urban travel demand can be measured by "the number of trips that people make or desire to make within an urban area." Travel demand models can be used for estimating the existing travel demand in an urban area, as well as forecasting the future travel demand. They also provide methods of assessing the impact of changes to the transportation system, such as the addition of new roadways or lanes, and can also be used in evaluating the impacts of the transportation system on air quality (*2*).

Specifically, household travel surveys have been used to collect a variety of socioeconomic characteristics for individual households, as well as information relating to trips made by each individual in that household in a 24-hour period. Until the mid-to-late 1980s, these surveys were administered as in-home surveys, in which surveyors would travel to thousands of homes in each urban area to collect travel information for each member of the household during a previous day. Since that time, however, improvements in computing technology have given rise to an alternative method of survey data collection,

---

This thesis follows the style of the *Transportation Research Record – Journal of the Transportation Research Board.*

called the computer assisted telephone interview (CATI). The CATI method of data collection has been the one primarily used in household travel surveys. Through the CATI, travel survey data for households is collected entirely through the respondent's home telephone, and no live personal interview is needed. Typically, paper diaries are also provided to respondents to assist in the recall process performed over the telephone (*3*). The use of the CATI survey method has significantly reduced the costs associated with conducting household surveys, as well as the time spent collecting the data. Despite this, continued advances in transportation modeling have resulted in the quality of household travel data obtained from the CATI being increasingly scrutinized in recent years. In turn, a variety of data correction methods have been developed. Weighting or other normalization procedures of survey data are used to better fit the travel demand models to the real-world situations from which they were obtained.

The application of global positioning systems (GPS) has become prevalent in a variety of transportation-related studies in recent years, such as travel time surveys, traffic and congestion management systems, and the above-mentioned household travel surveys. For household travel surveys, the addition of GPS data can provide a spatial dimension to the existing information, specifically that relating to route choice (*3*), as well as identify potential missed and unreported trips. In recent years, the use of GPS enhancement has been prevalent in household travel surveys as a means of comparing second-by-second vehicle position to trip information obtained using standard methods of house travel data retrieval, such as the CATI (*4-12*). In every one of these studies, a comparison of these two forms of household travel data revealed that there is some

degree of trip underreporting in the CATI survey that occurs. The majority of these studies compared the total trips made with the measured vehicle miles traveled (VMT). The overall potential impact of underreporting was analyzed with respect to the study area, and identifying specific household- and trip-related factors that might affect why certain trips are underreported was not done. In 2003, Zmud and Wolf used logistic regression to identify these specific household- and trip-related factors to trip underreporting using GPS-enhanced household surveys conducted in three urban areas in California (*10*).

The research presented in this thesis will apply a similar methodology as demonstrated by Zmud and Wolf (*10*), but will also test additional household- and trip-related variables that were not considered in their original logistic regression model. The additional variables that will be tested include the day of the week the survey was conducted, the time of day that the trip occurred, the purpose of the trip, and the distance traveled on the trip, the total number of vehicle trips take by the survey household on the survey day, the household residence type, and whether the surveyed vehicle was also used for commercial purposes. Two study areas, Laredo, TX and Tyler-Longview, TX will be used for investigation. Logistic regression models will be developed and interpreted for each of the two study areas, in addition to a single model encompassing both study areas.

RESEARCH OBJECTIVES

The research in this paper will explore the nature of trip underreporting that occurs in household travel surveys in two urban areas in Texas. In doing this, logistic regression models will be used in determining what are the significant factors affecting trip underreporting. These factors relate to both the nature of the household, such as household size and household residence type, as well as the nature of the trip, such as trip duration and trip purpose. The developed logistic regression model will be presented by comparing them to each other and to the results found in a similar study performed in California. These models will be analyzed using a variety of summary statistics and diagnostics, including analysis of variable effects, R-squares, goodness-of-fit measures, and association statistics. Additionally, a qualitative assessment of the models will also be made, taking into account the real-world implications of each variable on the response. The results from these models will then be placed in the context of their potential impact on existing urban travel demand models in use by the metropolitan planning organizations for each study area.

LITERATURE REVIEW

An article published by Frihida et al. in 2002 in *Transactions in GIS* dealing with new spatio-temporal models for disaggregate travel behavior immediately recognized that a large variety of survey data is collected for existing models. These include trip description surveys, one-day origin-destination surveys, real-time vehicle movement monitoring, long-term panel surveys, and GPS vehicle tracking (*13*). Because the information originating from these surveys is both spatial and temporal in nature, GIS technology is ideal for exploiting it, and is capable of allowing for identifying spatial links occurring between individual trips, transportation networks, and the distribution of activities within a region (*13*). In another article appearing in the same journal, Jiang and Claramunt (2002) take an more expansive view, recognizing that an integration between GIS and urban morphology can exist in such a way that the models and modeling concepts used in GIS account for the alternative cognitive-oriented models that can support human interaction with their environment (*14*). It is also argued that the modeling of "space syntax" into a GIS will provide a variety of new applications to planners for any urban studies in which the structure and function of a city are relevant.

It has also been shown that the locational referencing capabilities provided by a GIS framework makes it ideal for the incorporation of transportation-related studies, and that the combination of these two fields brings geography "to a full circle as it is re-discovering the primacy of space and time, two concepts that launched the systematic study of transportation in Geography and Regional Science in the 1950s" (*15*). Although it is important to recognize the need for such applications of GIS from a long-

term perspective, the specific relevance of these papers to the proposed research is only appropriate for providing an outline for future urban models which may incorporate all aspects of transportation data and its associated research. There does exist, however, a large amount of research which deals with a specific topic within transportation and geography that is extremely relevant to the research topic of this thesis – GPS and travel surveys.

In 2000, Taylor et al. recognized that GPS receivers are capable of providing a fast and convenient data collection which is easily be integrated into a GIS (*16*). In a synthesis report presented by the National Cooperative Highway Research Program in 2002, however, the integration of GIS and GPS for the purposes of mapping applications was identified as a significant problem. It was immediately identified that a major hurdle impacting the integration of GIS and GPS in handling travel surveys involved "locating travel routes within a digital base map when provided with a route generated from in-vehicle GPS collected points and identifying points that are trip origins and destinations" (*17*).

This viewpoint was supported by Denstadli and Hjorthol (*18*), claiming that the effectiveness of travel surveys to collect input data for transport modeling requires that survey respondents provide "adequate geographical information on their trips." Additionally, it was found that the use of a telephone interview for obtaining travel information was more accurate than a provided self-completion questionnaire. The higher accuracy of the interview was found because during the telephone interview, the interviewer was often able to help clarify and questions the respondent may have had

with regard to interpreting the survey questions (*18*). These types of user-interaction problems resulting from the use of existing survey techniques further supports the use of GPS devices as a supplemental means of collecting travel data, because it can assist in revealing these problems at a much finer spatio-temporal scale.

This belief is also supported by the reasons presented by Wolf et al. (*19*), in which the advantages of using GPS for travel survey data collection are clear. These include the collection of trip origin, destination, route taken, trip start and end times, and trip length, all obtained without questioning the respondent. Additionally, the collected GPS data can be compared against the reported trip for accuracy (*19*). Wolf et al. has also argued that the use of passive GPS data collection can also be effective if the primary interest is in completely eliminating the travel diary. This would reduce costs relating to telephone interview surveys and interview length time, allow for surveys extending over longer periods of time, improve the accuracy and completeness of existing travel models, and facilitate the collection of new data elements that can contribute to travel model validation, calibration, or update. In their study, Wolf et al. presented the first research that demonstrated the feasibility of deriving the purpose of the trip only by means of the in-vehicle GPS log combined with a spatially accurate GIS land use data base. Despite this claim, Wolf et al. also recognize that "there may always be a need for certain follow up questions regarding the derived travel data during the CATI household retrieval call" (*20*).

Through the use of a Personal Digital Assistant (PDA) combined with a GPS, Murakami and Wagner asked the question "can using GPS improve trip reporting?" (*6*).

Although they found that distances reported in the survey were over 50% greater than those recorded by the GPS, the authors still felt that the research was successful. It was stated that the use of both survey methods provided the best results, giving planners a variety of travel information, including trip purpose, occupancy, route choice, and travel speed, and that the combination of these survey techniques together should provide planners with the capability to evaluate transportation management systems, design intelligent transportation systems, as well as address other important issues (*6*). In assessing other likely technologies to be used for travel surveys in the future, Wolf (*21*) identified Assisted-GPS (A-GPS) as a likely choice. The use of A-GPS as means of travel data collection involves the use of any wireless network with its own GPS receivers, such as a cellular phone network (*21*). In this scenario, personal cellular phones could be used for monitoring a person for a 24-hour period or longer, provided the phone remains on (or relatively close to) the person at all times.

In 2003, Wolf et al. recognized the problem of trip-underreporting that occurs in household travel surveys, and present it as the result of a series of situations that arise during the survey process. "Memory decay, failure to understand or to follow survey instructions, unwillingness to report full travel details of travel, and simple carelessness have all contributed to the incomplete collection of travel data in self-reporting surveys." Wolf et al. also present some of the procedures that have been used over the years to attempt to correct for these issues, such as imputation, regression analysis, weighting procedures, and item substitution. It is also recognized that as the number of large-scale

surveys featuring a GPS component increases, the ability to calculate correction factors based on independent observations should provide more accurate results (*11*).

By comparing diary and GPS data, Pierce et al. (*7*) found that characteristics relating to the household can affect the level of under-reporting that occurs. Specifically, it was shown that smaller-sized and lower-income households tend to exhibit higher levels of under-reporting. Additionally, a comparison of the two methods of travel data collection revealed that approximately 30 percent of all vehicle and person trips per household were not reported (*7*). It is believed that many of trips commonly not reported may be short trips. The reasons affecting why people make short trips has been explored by Mackett (*22*). In surveying drivers who made short trips, it was found that the main specific reasons for making short trips are shopping, giving passengers rides, shortage of time, and because the car is needed for another trip prior to the return home (22).

Gliebe and Koppelman (*23*) examined household survey data from the Puget Sound Transportation Panel in Seattle, Washington to determine ways in which household members might share their travel during joint activities throughout the day. This survey was a two-day diary collected by the Puget Sound Regional Council. Using structural discrete choice models, the authors found strong evidence indicating that work schedules, commuting distances, automobile availability, and the presence of children have strong influence on both joint and independent activity patterns (*23*).

The effectiveness of logistic regression for modeling categorical variables in transportation-related issues has been demonstrated in recent years. In 2001, Li

presented a paper that used logistic regression to model which factors contribute to a person's likelihood to use the HOT (high occupancy toll) lanes on State Route 91 in California. In this study, riders on the HOT lane were stopped and surveyed for a variety of household and trip characteristics. These included household income, trip purpose, vehicle occupancy, and other related characteristics. To perform the logistic regression, each variable was assigned a categorical classification scheme, and the scheme was applied to each variable to discretize the data. The dependent variable tested in the model was a dummy variable stating whether or not the person used the HOT lane on their most recent trip that occurred during peak periods. The logistic regression analysis showed that household income, vehicle occupancy, and age were significant factors that contributed to a person's likelihood to use HOT lanes. It was also shown that gender, trip length, trip frequency, and other household characteristics played little or no role in determining whether or not a person was likely to use the HOT lanes (*24*).

In a study by Zmud and Wolf (*10*), trip data was collected in the form of household travel surveys and GPS vehicle surveys for three major urban areas in California. By comparing GPS data from a participant's vehicle to survey data taken form the same participant via CATI, underreported trips were identified. Zmud and Wolf selected ten variables relating to household and trip characteristics for a logistic regression model analysis. Each of the ten variables was discretized into three of four categories per variable. The response, or dependent variable, was a dummy variable indicated by a "0" for correctly reported trips and a "1" for incorrectly reported or non-

reported trips. By applying a logistic regression model to the ten independent predictors against the dummy dependent variable for missed trips, Zmud and Wolf were able to identify four of the ten variables that were significantly associated with trip underreporting. Zmud and Wolf were also identified in their conclusions that trip purpose could be a significant correlate of trip underreporting, but was not available in their data to be tested (*10*).

Similar to Zmud and Wolf, this thesis will also use logistic regression modeling to test independent predictor variables against trip underreporting. The research presented in this thesis will improve upon the Zmud and Wolf model by including trip purpose, as well is creating alternative models with fewer variables using stepwise variable selection procedures. Because the Zmud and Wolf model included data from multiple study areas, a similar model including both study areas will be used for the comparison.

LOGISTIC REGRESSION

Background

In the ordinary least squares (OLS) regression model, it is assumed that the dependent variable is continuous or quantitative in nature. With this assumption, a continuous dependent variable such as income level can be modeled against a series of both continuous and discrete independent variables – education level, sex, race, and unemployment rate. In many social phenomena, however, the dependent variable of interest does not occur continuously, but rather as a discrete choice that is qualitative in nature.

For example, it might of interest to determine whether or not a person is likely to vote yes or no in the next election. In this example, there are only two discrete outcomes that can occur for the dependent variable from the model. In other words, the dependent variable is dichotomous. Dichotomous dependent variables in which the responses are "yes" and "no" are most frequently modeled with a value of 1 for "yes" and 0 for "no." The regression model presented in this thesis also uses a dichotomous dependent variable, whether or not a household reported a trip through the CATI survey that was detected in the GPS survey. If the trip was accurately reported in the CATI survey, the dependent variable was assigned a value of 1. If the trip was not accurately reported in the CATI survey or not reported at all, the dependent variable was assigned a value of 0. In both of the above-mentioned examples, the ideal modeling situation would be to, given a series of independent variables of interest, predict an outcome of 0 or 1 for the dependent variable that can be used to predict one outcome or the other. For regression

modeling involving this kind of dependent variable, it can be shown why the use of the logistic form is more appropriate than the OLS form for predicting probabilities.

The first issue with the OLS form in handling dichotomous dependent variables involves issues with the prediction floor and ceiling. Conceptually, the use of OLS deals with continuous dependent variables that have no prediction floor and ceiling. If OLS were used to model an outcome as a predicted probability, any value less than 0 or greater than 1 would not be intuitive with respect to the dependent variable of interest. Logistic regression, however, has overcome this conceptual problem, and only models values between 0 and 1. The second issue with the OLS form deals with the problem of nonlinearity. Because the final model values for the dependent variable are only 0 and 1, attempting to fit a straight line through the two values of points seems inappropriate (Figure 1). Additionally, any line that does not have a slope other than zero will eventually fall below 0 and exceed 1. The nonlinear s-curve of logistic regression (Figure 1) bends slowly and smoothly as it approaches 0 and 1. As values get closer to 0 and 1, the relationship between the dependent variable and independent variables requires a larger change to have the same impact as a smaller change in the independent variable at the middle of the curve (*25*).

The third issue with the OLS form compared to the logistic form in handling dichotomous dependent variables deals with the assumption of normality of errors. In logistic regression, because there are only two dependent variable outcomes for the model, only two residuals can exist for any single give independent value in OLS. Although this violation of normality of errors can cause problems, they are found

to be minimal when dealing with large samples (*25*). On the other hand, the violation of the assumption of homoscedasticity can have much more serious effects. The assumption of homoscedasticity states that at any given value of the independent variable, there is an equal variance for the dependent variable values. This assumption is violated because at upper and lower extreme values of the independent variables, the residuals are relatively small, and near the middle values of the independent variables, the residuals are relatively large. In other words the error variance is not constant along the independent variable. This violation of homoscedasticity results in the overestimating of sample variance for regression coefficients. This leads to standard errors of the sampling estimates being biased, making significance tests invalid (*25*).

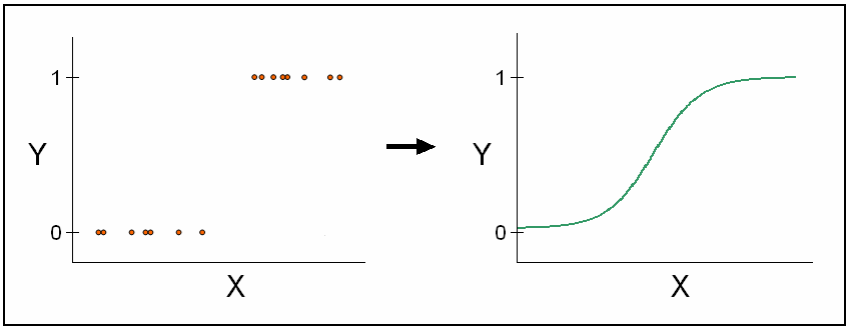

FIGURE 1 Example of logistic regression S-curve.

In logistic regression, where the dependent variable to be modeled is 0 or 1, it is important to understand that the final model estimate for each individual sample normally does not result in exactly 0 or exactly 1, but rather a predicted probability falling somewhere in between 0 and 1. For example, with a given set of independent variable inputs to a model, the dependent variable might result in a predicted probability

of 0.750 that a person might answer yes to a particular survey question. This indicates that using the available sample data, the model estimates there is a 75% likelihood that someone who matches that set of independent variable inputs will answer yes to that question. To determine 0's and 1's, a cutoff value is assigned to each model for the final dichotomous prediction. A cutoff value of 0.5 is normally used, indicating that any value of 0.5 or higher will be predicted as a 1, and any value lower than 0.5 will be predicted as a 0. Considering that probabilities ranging between 0 and 1 are capable of being estimated by a given logistic regression model, it becomes clearer as to why the s-shaped curve of the logistic regression model is appropriate (Figure 1).

Logit Transformation

To remove the effects of ceiling and floor in predicting probabilities, the logistic regression model uses a logit transformation. Through the logit transformation, dependent values can be less than 0 and greater than 1. The logit transformation linearizes the nonlinear relationship between the independent variables and the original probabilities. The logit transformation is a two-step process, the first step being the transformation of probabilities into odds. Given the probability of an event *i*, defined as P_i, the odds of an event can be seen in Equation 1.

$$O_i = P_i /(1 - P_i) \tag{1}$$

In this equation, the transformation of probabilities to odds eliminates the ceiling of 1 for dependent values. For example, if the probability of an event is 0.8, than the odds are

equal to 0.8/0.2, or 4. If the probability of an event is 0.2, than the odds are equal to 0.2/0.8, or 0.25. The second step in the logit transformation is to take the log of the odds, as seen in Equation 2. In this equation, L_i is the value of the logit for an event *i*.

$$L_i = \ln(O_i) = \ln[P_i/(1-P_i)] \quad (2)$$

By taking the natural logarithm of the odds, shown in equation 2, the floor of 0 is also removed for dependent values. For example, if the predicted probability of an event is 0.8, the logit is equal to ln(0.8/0.2), or 1.39. If the probability of an event is 0.2, the logit is equal to ln(0.2/0.8), or -1.39. In other words, the combination of the two steps shown in Equations 1 and 2 in transforming the predicted probability of a dependent variable value into a logit results in a continuous and linear form from which it can be modeled (*25*).

Using the logit transformation of the dependent variable, L_i, the regression coefficients can be modeled using the standard OLS form, seen in Equation 3. Once the regression coefficients (*b* values) are obtained using the logit transformation of the dependent variable, Equations 4 through 6 show how the predicted probability can be calculated from them.

$$L_i = \ln[P_i/(1-P_i)] = b_0 + b_1 X_i \quad (3)$$

$$e^{L_i} = P_i/(1-P_i) = e^{b_0+b_1X_i} = e^{b_0} * e^{b_1X_i} \quad (4)$$

$$P_i = (e^{b_0+b_1X_i})/(1+e^{b_0+b_1X_i}) \quad (5)$$

$$P_i = e^{L_i} / (1 + e^{L_i}) \quad (6)$$

From the above equations, it can be seen how logistic regression modeling is able to use a series of transformations to the dependent variable in order to allow it become continuous value with neither a ceiling or floor, but yet still be able to estimate a predicted probability that only falls between 0 and 1 (*25*).

When dealing with independent variables in logistic regression modeling that are categorical, using a single *b* value to represent multiple discrete categories within a single variable is not possible. To handle this, one category within each independent variable is assigned as the reference category. Using the reference category, every other category within an independent variable is given its own regression coefficient, expressed in relation to the reference category.

STUDY AREAS

The research presented in this thesis dealt with two primary study areas, both located within the State of Texas. The first study area, Laredo, is a medium-sized city with a 2005 population estimated at 215,375 according to the Laredo Development Corporation. The U.S. Census Bureau has ranked Laredo as the 23rd fastest growing city in the Unites States with a 3.3 percent annual growth rate, based on population estimates from July 1, 2003 to July 1, 2004 (*26*). Located in Webb County on the Mexican border, this city is adjacent to the Rio Grande River, and is a major thoroughfare for international commercial and non-commercial traffic. The U.S. Census Bureau estimates the Webb County population was 219,464 in 2004 (*26*). In other words, virtually of the travel of interest that occurs in Webb County happens within the city of Laredo. For this research, only travel within the city of Laredo will be considered. In other words, only those trips in which either the trip start or trip end location occurred in Laredo will be analyzed.

East of Dallas/Fort Worth, the second study area is the five-county area surrounding the cities of Tyler and Longview. The five counties encompassing this study are Gregg, Harrison, Rusk, Smith and Upshur. The U.S. Census estimates the combined population for these five counties as of July 1, 2004 was 449,546 (*27*), up 1.0 percent from the same date one year earlier. Similar to Laredo, only travel associated directly with the five-county study area surrounding the cities of Tyler and Longview will be considered in this research Only those trips in which at least one trip end occurred in within the five counties will be analyzed.

For the Laredo study area, a total of 83 households were selected comprising 116 vehicles, yielding an average of 1.40 vehicles per household. In the Tyler-Longview study area, a total of 171 households were selected comprising 272 vehicles, yielding an average of 1.59 vehicles per household. The map in Figure 2 shows the locations of the two study areas with respect to some of the major cities in Texas.

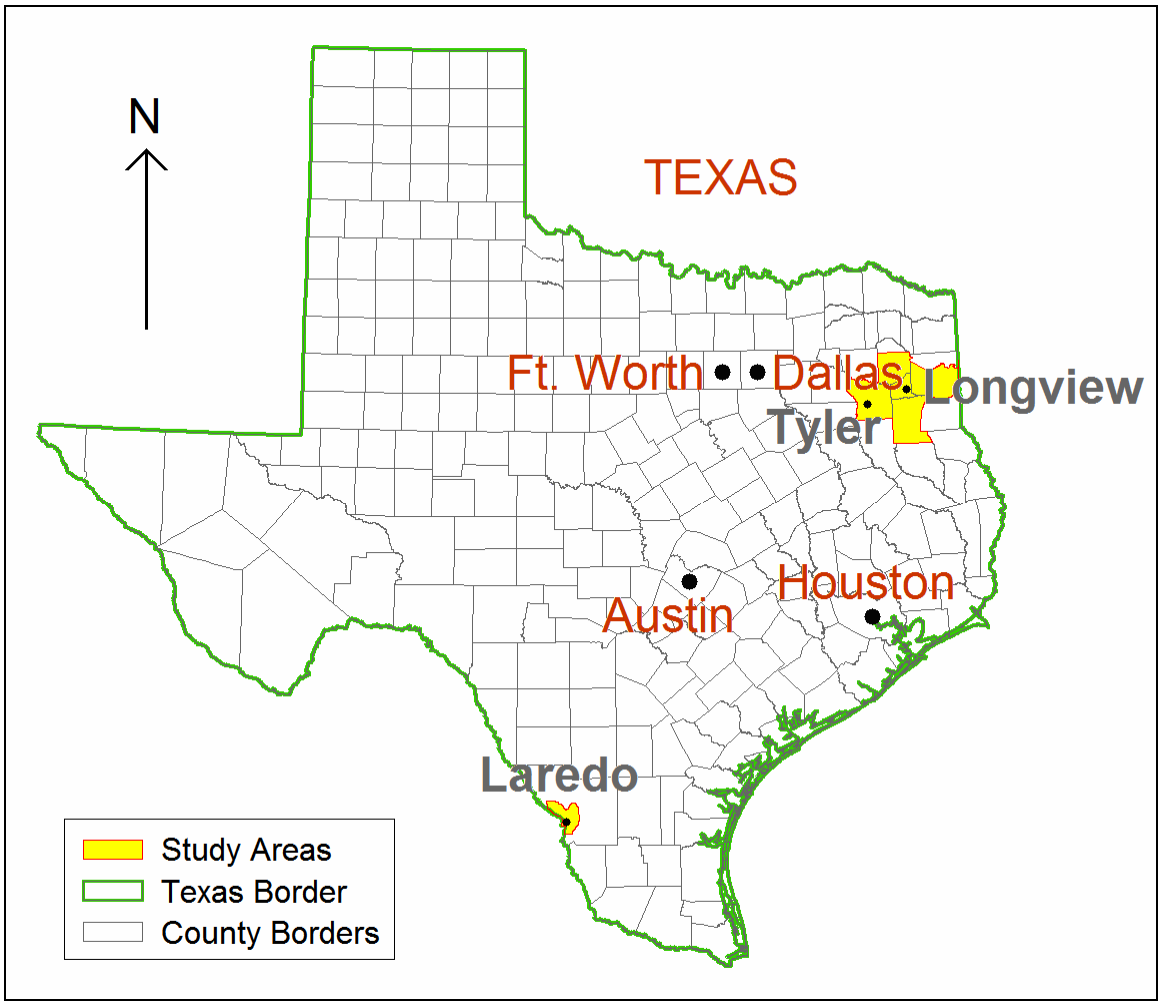

FIGURE 2 Map of study areas.

DESCRIPTION OF DATA

For this study, a variety of survey and GIS data were used during the analysis. These data were provided primarily through the Texas Department of Transportation's (TxDOT) Travel Survey Program (TSP) and the metropolitan planning organizations (MPO) in each of the two study areas involved. Table 1 lists all of the data obtained for this study, as well as the agency supplying them (data source), and who was the initial collecting organization of the data was (data collected by).

TABLE 1 Data Type, Data Source, and Data Collector

Study Area	Data Type	Data Source	Data Collected By
Laredo	CATI survey data	TxDOT TSP	NuStats, Inc.
	GPS survey data	TxDOT TSP	GeoStats, Inc.
	Laredo GIS transportation network	Laredo MPO	Laredo MPO
Tyler-Longview	CATI survey data	TxDOT TSP	MORPACE Intl.
	GPS survey data	TxDOT TSP	GeoStats, Inc.
	Tyler GIS transportation network	Tyler MPO	Tyler MPO
	Longview GIS transportation network	Longview MPO	Longview MPO
	County Files	TxDOT TSP	TxDOT

Survey Data

The two types of survey data used in this research were the CATI and GPS survey data. For both study areas, these survey data were obtained from the TSP. This program coordinates, finances, and is ultimately responsible for the collection of all travel data in Texas used in statewide and regional planning. In addition to the household survey data collected via CATI and GPS, other types of travel surveys are also performed around the state that are managed through the TSP, such as external station surveys, workplace surveys, commercial vehicles surveys, and special generator surveys. The CATI and GPS household surveys for the Laredo study area were

performed between March 25, 2002 and May 29, 2002. The surveys for the Tyler-Longview study area were performed between September 2, 2003 and November 25, 2003.

For each of the two study areas, the CATI survey data was provided as two fixed-width text files containing four record types. The first text file contained three of the four records – the household record, the person record, and the vehicle record. Each record contains all of the survey data relating to the record type. The household record contains all of the household information for each surveyed household, such as number of persons, number of persons employed, and household income. The person record contains specific information about each person in each of the surveyed households, such as age, sex, race, and employment status. The vehicle record contains specific information about the vehicles owned by each of the surveyed households. The second text file provided as part of the CATI survey data contains the fourth record – the trip record. This record lists all of the trips made by each household member on the assigned survey day. This includes information such as trip start time, trip end time, vehicle used on trip, trip purpose, and trip origin and destination (longitude-latitude pairs geocoded from a street address). In each of the four records, each household is assigned a household identifier to maintain consistency in each of the four record types. Additionally, within each household, all persons and vehicles in each household were also given a unique ID number. This ensures that in any analysis across multiple records, the household, person, and vehicle ID numbers will always correspond. Examples of the data file formats for each of the four records of the CATI survey data

can be found in Appendix A, Tables A-1 through A-8. The data file format for Laredo was selected for the examples in the Appendix. There are variations in the CATI survey between the two study areas, with additional questions being added for border-related issues in Laredo.

The second type of survey data collected, the GPS data, was provided as individual tracking files for each vehicle from the TSP. Each of these files was identified by the household ID and vehicle number, allowing for a simple means of comparison to the CATI survey trip data. The GPS trip data was recorded in each vehicle on a second-by-second basis and stored in a comma-delimited text file. Each GPS receiver was connected through the cigarette lighter in the vehicle for a continuous power supply, and only collected data when the vehicle engine was running.

Additional Data

To complement the spatial components of the CATI and GPS survey data within the GIS, it was also necessary to obtain the transportation networks for each of the two study areas. These transportation networks were not used for calculating any distances, but rather as simply a reference base map which could be used for assisting in the visual identification of trip ends from the CATI and GPS survey data. In the Tyler-Longview study area, a combination of two transportation networks was used. The first transportation network used was created by the Tyler and Longview MPOs. This network contained all of the roadways within each of the two MPO boundaries. The second transportation network that was used was provided by TxDOT TSP as a part of

their County Files package, a statewide dataset which includes all road network segments and other transportation features for every county in the state of Texas. For this data, only the network segments outside of the MPO boundaries were used to fill in the remaining sections of the five-county study area encompassing Tyler-Longview. These two networks were combined to form the single transportation network used for analysis. Because the Laredo study area of interest was only the Laredo MPO, the additional street network data available from TxDOT TSP was not necessary.

METHODOLOGY

Analysis Software

Three primary software packages were used to analyze the data. The first one was the Microsoft® Office suite, specifically Excel and Access. Microsoft Excel was used for converting the plain-text data files (CATI) into a more user-friendly spreadsheet format. It was also used with the GPS data for making preliminary predictions of trip ends, as well as second-by-second distance calculations from the GPS data. To calculate distance using the GPS data, second-by-second velocity readings (meters/second) were summed over time, for the entire length of each trip. These velocities were calculated by the GPS unit and stored in the second-by-second data. During the trip analysis, all analysis tables were created with Excel. As the analysis progressed and results became richer with detail, these tables were modified as necessary to maintain the highest level of detail. Excel also acted as a medium for converting the tabular data into a database format compatible with Microsoft® Access. This was crucial because Access can create geo-databases of the trip data that are compatible with the geographical information system used for this research, ESRI's ArcGIS™. Using geodatabases created with Access allowed the data to be managed in larger (but fewer) files, eliminating the need to maintain hundreds of individual data files.

The second software package, ArcGIS, is a powerful geographical information system that can be used for a variety of analysis within the transportation field. For this study, ArcGIS offered a framework within which all of the street networks, GPS-collected latitude/longitude pairs, and geo-coded addresses were stored, viewed,

manipulated, and analyzed for virtually all aspects of the research. All maps produced in this thesis were created with ArcGIS.

The third software package used was the SAS® System, developed by the SAS Institute, Inc. The SAS System provides a statistical framework in which logistic regression analysis can be performed. The use of SAS allows for a variety of parameters to be fed into the model prior to testing, allowing for the creation of more sophisticated models, alternative testing procedures, and a variety of customized significance tests and regression model outputs.

Data Preparation

Preparation of the survey data prior to performing the analysis ensured that it could be properly accessed, displayed, and edited as necessary. As discussed above, spreadsheets were used to initially import the plain-text data files for both survey data types (GPS and CATI) into a spreadsheet format.

During the GPS survey data (Table A-9) import to a spreadsheet format, 24-hour distance calculations were also made for each vehicle by summing the distances between each one-second increment. Initially, this was measured as the total distance traveled for each vehicle for the entire 24-hour survey period. As trips were identified within the data sets, however, the total distance was determined for individual trips. For the GPS data, each spreadsheet was maintained as a single worksheet, and imported into Access into individual tables, where the latitude/longitude pairs were used as part of a geodatabase.

CATI Survey Data Trip Determination

Once the CATI trip data was converted and prepared in a spreadsheet format, a determination of the trip ends was made. When reporting their travel information as part of the CATI, the trip data reported by each household member was recorded in the trip record (Tables A-7). Because this study is only interested in travel patterns as they relate to the study area, only those trips where at least one trip end was within the study areas identified in Figure 2 were included in the analysis. Any trip in which both trip ends were outside of the study area were excluded from analysis.

To determine the trips made for each vehicle within the household, a simple sorting process was used. First, each of the household members' trips was sorted by vehicle. Next, vehicle trips were sorted by the time of day, starting with the first trip that ended at 12 a.m. or later on the assigned travel day and ending with the last trip that began before 12 a.m. on the day after the assigned travel day. After sorting, duplicate trips (the same trip made in the same vehicle by more than one member of the household) were removed from the list for each vehicle. Each household member also classified each trip that they took based on activity and trip purpose codes, selected from predetermined lists located in Table A-8. These two pieces of information provided enough detail about each trip to determine the trip purpose as either home-based work, home-based non-work, or non-home-based. Home-based work (HBW) trips are any trip, regardless of direction, in which one trip end is the drivers' home location and the other trip end is the drivers' work location. Home-based non-work (HBNW) trips are any trip, regardless of directions, in which one trip end is the drivers' home location and the other

trip end is any non-work location for the driver. Non-home-based (NHB) trips are any trips, regardless of directions, in which both trip ends are not the home location. Once determined, the trip purpose information, as well as the reported trip start and end times, were stored in the trip comparison table. The trip comparison table was used to store trip information for all the vehicles of households that participated in both surveys. This table was used to compare the trips identified from the two survey sets and determine which trips that were found in the GPS survey were not reported in the CATI survey. The format of the trip comparison table will be discussed in the section *Comparing GPS and CATI Trips*.

GPS Survey Data Trip Determination

The complexities involved in determining trip ends using only point-based, second-by-second GPS data required the use of a multi-step heuristic procedure. This ensured not only that identified trip ends could be determined as accurately as possible, but that the process was accomplished in an efficient manner. The first step was performed entirely within the spreadsheet, and was based on vehicle velocities recorded by the GPS unit. The second step used the GIS to make visual detection of missing and false trip ends determined from the first step.

For most vehicles, the GPS unit was able to begin data acquisition from GPS satellites within a few seconds of the vehicle engine start-up, resulting in little or no loss of data for trip starts. For the initial step of the heuristic procedure, trip starts were

identified as the initial second that, immediately after vehicle start-up, the velocity was reported by the GPS unit as greater than zero.

In identifying trip ends, a dwell time threshold of 120 seconds was selected to make the initial determination. This meant that for the first step of the heuristic procedure, trip ends were identified at any locations where the vehicle velocity remained at zero for 120 seconds or more. Previous studies have shown this to be an appropriate dwell time threshold for making an initial determination of trip ends (*5, 10, 12*).

For this study, however, the dwell time threshold was modified slightly in attempt to eliminate additional potentially false-positive trip end detections. The modified dwell time threshold used for these study areas allowed vehicles to remain at a zero velocity for up to 135 seconds (as opposed to 120 seconds) without a trip end being placed at that location, as long as the vehicle heading (direction the vehicle was facing) did not change by more that five degrees during the final ten seconds immediately prior to the velocity reaching zero. This modification to the dwell time threshold implied that despite the vehicle dwell time being as long as 135 seconds, the vehicle did make a stop because there was an insufficient change in heading to indicate that the vehicle came to rest off of the road and made a stop. It was found that this modification did remove a small number of false-positives associated with delay at traffic control devices, many of which were found to be in the 120-135 second range. Despite this, it is still possible that this threshold setting could miss some trip ends where the driver parked or made a stop along the road without having to turn off. Additionally, extremely small changes (+/- 0.00001 decimal degrees) in GPS readings due to satellite shifting occasionally

resulted in the velocity being calculated by the GPS unit as greater than zero when in fact there was no vehicle movement, and these small velocities were treated as zero to ensure correct dwell and trip time calculations. Any changes in latitude/longitude position greater than +/- 0.00001 decimal degrees were considered vehicle movements.

After the initial estimates of trip ends were made in the spreadsheet, the second step of the GPS survey data trip determination procedure required that the GPS data be exported to the GIS in a longitude-latitude format via a geodatabase. The use of the GIS facilitated viewing the GPS data spatially, overlaid with a road network of the study area. This was necessary to visually identify missed or false trip ends determined from the first step. Each trip identified in the first step was placed on a separate layer within the GIS, allowing them to be viewed one at a time over the road network. From each of these individual trip layers, visual identification of missed trip ends was performed by locating consecutive GPS points along a trip that appeared to represent the path of a vehicle at a trip end or passenger drop-off was made. These were typically seen as small turns off of roads into driveways and parking lots, but occasionally seen as longer detours off of a road to reach the trip end, followed by a return to the previous road the trip was made on. Examples of these visual identifications of trip ends using the second step of the procedure can be seen in Figure 3. In this figure, the trip ends are circled in red. Once identified visually, the trip start and trip end estimates made from the first step were corrected, and a final list of trip ends for each vehicle was determined.

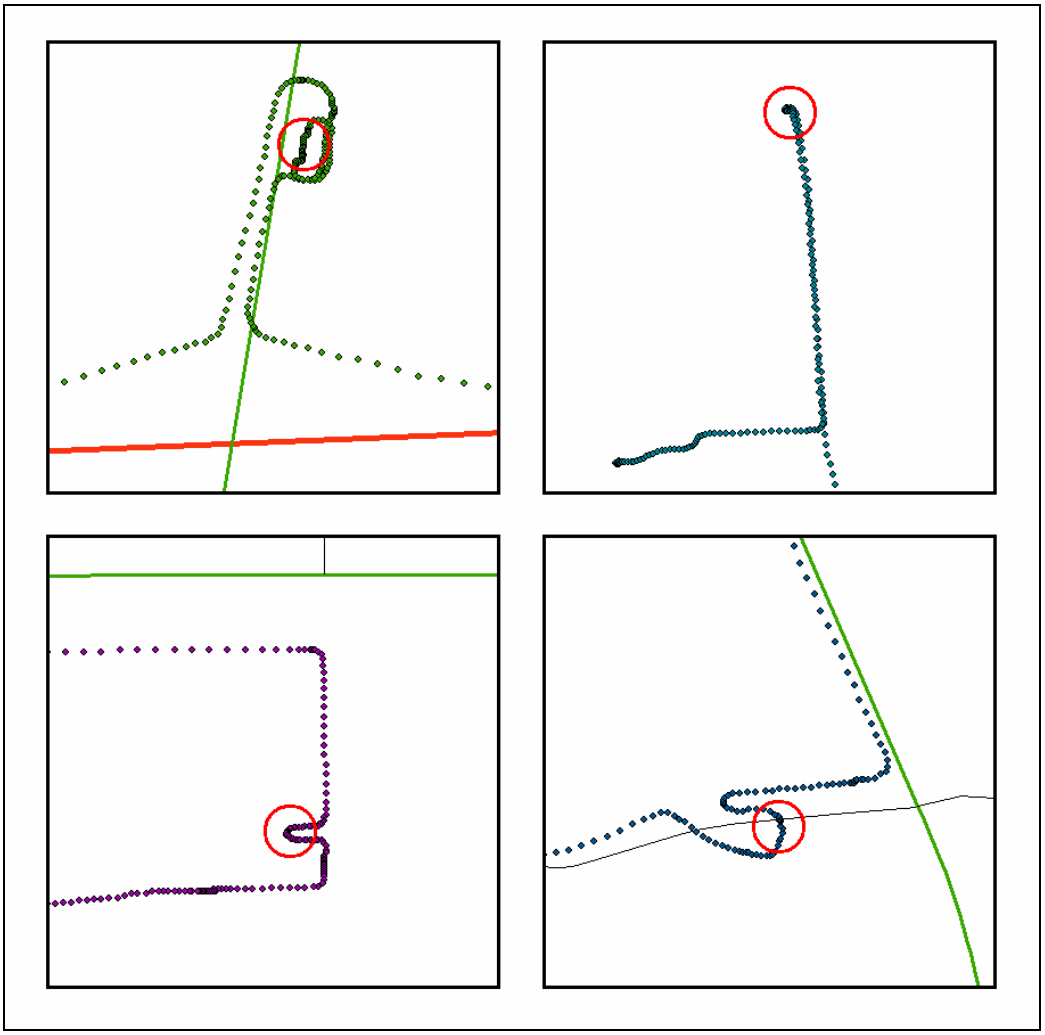

FIGURE 3 Examples of visual identification of missed trip ends.

Because each home and work location in the CATI trip data was geocoded into a longitude-latitude format, a determination of the trip purpose as HBW, HBNW, or NHB for each trip found in the GPS data. A screenshot of GIS software showing GPS second-by-second data and CATI geocoded location data from one vehicle within a household can be seen in Figure 4. In this screenshot, the numbered point locations represent geocoded locations of the addresses specified as trip ends in the CATI survey. This

figure also shows an example of the GPS second-by-second trip data proceeding from one CATI trip end location to the next. In Figure 4, GPS Trip 6 travels from CATI Geocoded Trip Location 8 to CATI Geocoded Trip Location 9. As each GPS trip was identified by its trip ends and classified by trip purpose, this information was stored in the trip comparison table along with the CATI trip data. By using this table to compare the trips found within each survey, the trips in the GPS survey not reported within the CATI survey could be identified.

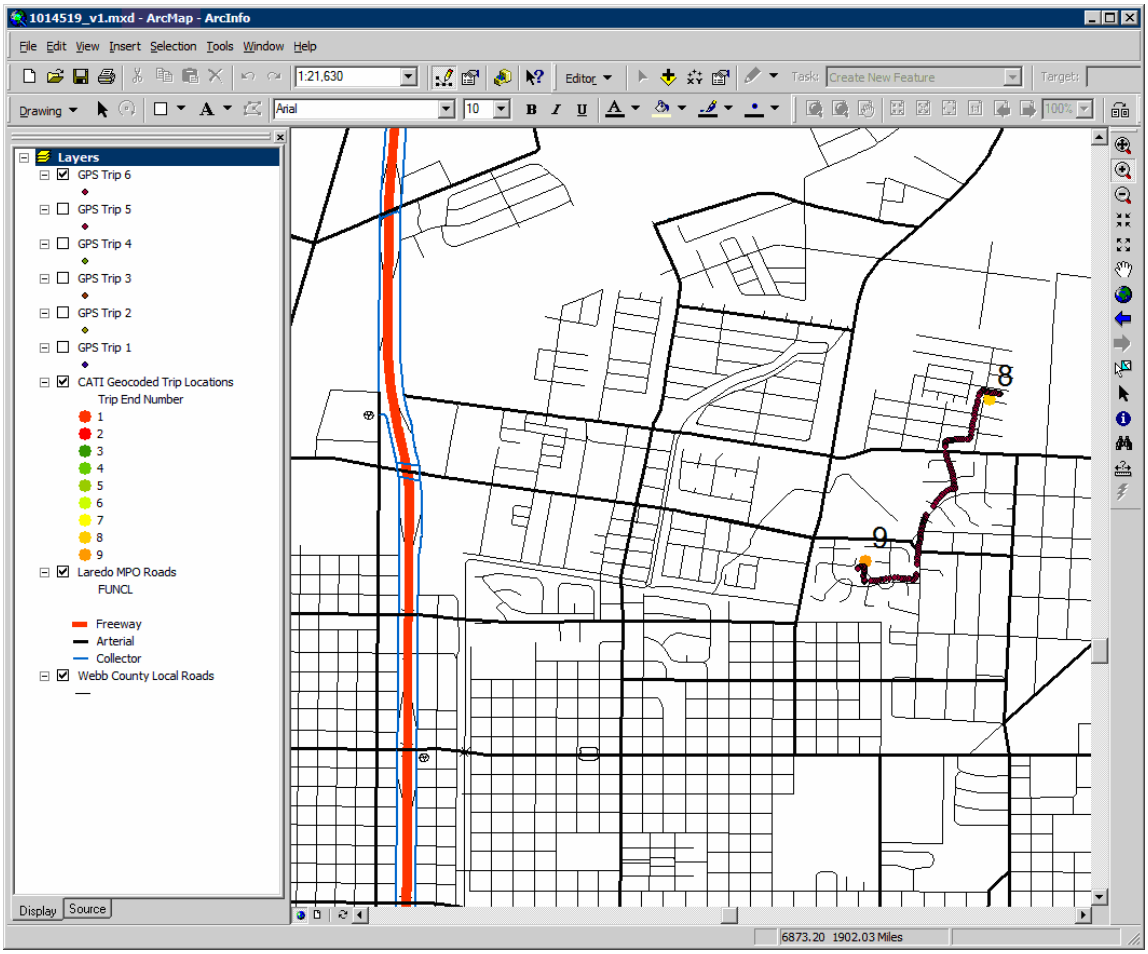

FIGURE 4 Example of GPS trip data with CATI trip locations.

Comparing CATI and GPS Trips

In order to compare the trips for each household and vehicle from the survey, the trip comparison table was created. The fields for this table are in Table A-10 in Appendix A. From this table, each trip start and end time found in the GPS survey was compared against the trip start and end times reported in the CATI. Because survey respondents in the CATI frequently rounded trip start and end times to the near 5, 10, 15, or 30 minute increments, a +/- 30 minute threshold was used for comparing trip times between the two surveys.

A GPS trip was classified as a reported trip in the CATI if it met the following two guidelines: 1) the trip start and end times are both within 30 minutes of the start and end times for a trip reported in the CATI; and 2) the location of the trip start and trip end correspond to geocoded locations specified by the survey respondent in the CATI. In other words, the GPS trip must closely match the CATI trip both temporally and spatially. Using this criteria, each GPS trip was classified as reported or not reported, based on the CATI survey data. This indicator was used as the dependent variable for regression modeling. Figure 5 shows an example of comparing GPS to CATI trips. In this figure, the trips highlighted in red indicate GPS trips that were not found in the CATI. Trips highlighted in green indicate GPS trips that were found in the CATI. In this example, it can be seen that the vehicle made the first two trips as reported in the CATI, but then proceeded to make five additional trips later in the day that were not reported in the CATI.

Trip No.	GPS Trip Time		Classification				Classification			CATI Trip Time		Classification			Classification		
	Begin	End	H	W	I	N	HBW	HBNW	NHB	Begin	End	H	W	O	HBW	HBNW	NHB
1	8:46	8:50	⊗→				X			9:00	9:05	⊗→			X		
2	11:46	11:51	←⊗				X			11:00	11:05	←⊗			X		
3	13:32	13:36	⊗→					X									
4	14:03	14:09				⊗			X								
5	14:14	14:18				⊗			X								
6	14:45	14:46				⊗			X								
7	14:56	15:04	←⊗					X									
Totals							**2**	**2**	**3**						**2**	**0**	**0**

FIGURE 5 GPS and CATI trip comparison.

Logistic Regression – Data Preparation

Once each GPS trip was classified as being reported or not reported, this classification was used as the dichotomous dependent variable for the model. From the trip comparison table, it was necessary to create a new table with additional fields containing all of the independent variables that were tested in the models. All of the independent variables used in the models can be seen in Table 2, and the dependent variable in Table 3. In this table, the first six independent variables (HHTRIPS through DAYWEEK) were already stored in the trip comparison table, and remained in the new table. The remaining six independent variables (NUMPER through DELVEH) were collected and stored as part of the CATI household record (Table A-1). To link the values associated with these remaining six independent variables to their corresponding households, a simple query was designed in the database to add these fields to the newly created table. By linking the household ID number shared between the two tables, a simple query can be used to append fields to the table containing the additional variables.

TABLE 2 Independent Variables Used in the Logistic Regression Model

Independent Variables	Variable Name	Variable Format – Coding Scheme
Number of trips taken during day by household	HHTRIPS	Continuous – Integer
Purpose of trip	PURPOSE	Categorical – HBW, HBNW, or NHB
Time of day trip occurred	BEGIN	Categorical – 12a-4a, 4a-8a, 8a-12p, 12p-4p, 4p-8p, or 8p-12a
Length of trip in minutes	TIME	Continuous – Floating
Length of trip in miles	DIST	Continuous – Floating
Day of week survey was conducted	DAYWEEK	Categorical – M, T, W, R (Thurs.), or F
Number of persons in household	NUMPER	Continuous – Integer
Number of employed persons in household	NUMEMP	Continuous – Integer
Income of household	INCOME	Continuous – Integer
Number of vehicles available to household	VEHAVAIL	Continuous – Integer
Household residence type	RESTYPE	Categorical – Single or Multi
Is vehicle used for commercial purposes	DELVEH	Categorical – Yes (1) or No (1)

TABLE 3 Dependent Variable Used in the Logistic Regression Model

Dependent Variable	Variable Name	Variable Format – Coding Scheme
Was trip reported correctly?	REPORTED	Categorical – Yes (1) or No (1)

With the exception of the time of the day the trip occurred and the household residence type, the remaining ten independent variables were kept in their original numerical or categorical format for use in the regression model. For these models, the time of day the trip occurred was divided into six equally spaced time periods throughout the day, starting with 12:00 a.m. to 4:00 a.m., and ending with 8:00 p.m. to 12:00 a.m. The start time for each trip was used to place it into one of these six groups. The impact of splitting the trip start time variable into six equally spaced groups is that, if this variable is found to be significant, underreporting can be more easily understood in the context of what time periods during the day generate higher rates of underreporting. For household residence type, there were initially six classes, shown in

Table A-2. These six original classes were aggregated into two larger classes, single-family households and multi-family households. A single-household residence is any unattached single-family home or mobile home. A multi-household residence is any apartment, condo, duplex, or other form of attached housing.

Using the variables in Tables 2 and 3, a spreadsheet was created for each of the study areas, as well as one for the combined study areas. This spreadsheet was exported into a tab-delimited format, where it could be placed directly into a SAS program as modeling data.

Logistic Regression – Modeling with SAS

For performing the logistic regression modeling discussed here, the PROC LOGISTIC module within SAS was used. This module can be used to develop multi-variate logistic regression models, and allows for a variety of input parameters for model building and significance testing. The SAS programs that were created for the logistic regression modeling in this research are listed in Appendix B.

Within the PROC LOGISTIC module, the DESCENDING parameter was specified, indicating that the selected variables will be fitted to REPORTED=1 (a reported trip). This implies that SAS will fit the sample data to correctly reported trips as opposed to incorrectly reported trips. If this parameter where not included, REPORTED=0 (a non-reported trip) would be modeled instead. Either option will yield the same model results, with the only difference being opposite signs for the regression coefficients.

Since the models created here use both continuous and categorical independent variables, it was necessary to specify in SAS which of the variables were categorical using the CLASS statement. Any variable not defined as categorical using the CLASS statement was assumed to be a continuous variable. The CLASS statement was used to define a reference category within each categorical independent variable. The reference category is used in logistic regression modeling as a way of redefining categorical variables as a series of dichotomous variables. For example, the independent variable trip purpose has three values – HBW, HBNW, and NHB. By assigning NHB as the reference category, each of the other two variables is treated as a dichotomous variable against NHB. In other words, the three-way category becomes two simpler dichotomous variables – HBW (1) vs. NHB (0) and HBNW (1) vs. NHB (0). For any independent categorical value with *n* categories, the use of the CLASS statement in PROC LOGISTIC will convert the independent variable to a series of *n*-1 dichotomous variables.

The MODEL statement in PROC LOGISTIC is used to specify the modeling parameters, as well as conduct any additional significance tests. For each of the three sets of data used for modeling (Laredo, Tyler-Longview, combined), two models types were formed – a stepwise model with specified significance levels required for a variable to be entered into the model, and a full model using all twelve of the selected variables. After the MODEL statement, the dependent variable to be tested is listed, followed by an equal sign, and followed by the full list of independent variables to be tested in the model, separated by spaces. The MODEL statement supports a variety of additional

options which instruct SAS to perform additional tests of significance on the data. To obtain the Hosmer and Lemeshow goodness-of-fit test, the LACKFIT option was used. The RSQAURE option displays calculated R-square and max-adjusted R-square values as a part of the SAS output. A combination of the options SCALE=NONE and AGGREGATE prompt SAS to also display the Pearson goodness-of-fit test and the Deviance goodness-of-fit test. Although these two tests are appropriate for assessing logistic regression models in SAS, they typically perform better when all of the variables are continuous rather than categorical. For models involving categorical variables, the Hosmer and Lemeshow test obtained from the LACKFIT option is more appropriate. Despite this, the Pearson and Deviance goodness-of-fit tests were also performed to show the confounding results that can be obtained from these two significance tests when categorical variables are used (*28*).

For each of the study areas, two models were created – a stepwise model and a full model. In the stepwise selection model, the final regression model is created by starting with a flat intercept model with no variables, and successively adding the independent variables one at a time to the model in the order of significance. By specifying p-value cutoff levels within SAS, it can be determined at what point no more variables will be entered into the model, based on significance level. Additionally, a p-value cutoff level can also be specified as required for a p-value that is already in the model, to remain in the model. In other words, a variable can be removed from the model at a later step if it no longer found to be significant when placed with other variables. For the stepwise selection model, it was necessary to use the

SELECTION=STEPWISE option. For these models, a p-value of 0.001 was used as a cutoff for both entrance into the model and as a cutoff for staying in the model. The extremely small p-value for these models was selected to make the models as simple as possible. The options SLENTRY=0.001 and SLSTAY=0.001 were used to define these cutoff levels.

For the full model, all of the variables are entered into the model on the first step, and remain in the model. No additional variable removal or addition procedures are performed as with the stepwise selection model. This is the default option (same as SELECTION=NONE) in PROC LOGISTIC, so no option statements are needed to define the variable selection method and associated cutoff levels (*28*). Once the programs were created, SAS was used to run the programs and provide the resulting output. The modeling output for each program was saved to a text file so that it could be accessed by any basic word processor. Selected portions of the regression model outputs can be found in Appendix C.

The full model developed for each of the three areas of interest (Laredo, Tyler-Longview, and combined) will be used to compare against the Zmud and Wolf (*10*) model, which also used a regression of all of the selected variables without variable selection. The stepwise model for each of the three areas of interest will be used to look at trip underreporting rates as they relate to the variables chosen using the stepwise selection process.

RESULTS AND INTERPRETATION

The results of the logistic regression models developed in this thesis are presented in this section. The SAS output obtained through the processes of variable selection, modeling parameter selection, and final model testing will be discussed. The models obtained from each study area will be compared to one another, as well as to the model obtained by Zmud and Wolf (*10*). A single, unified model incorporating the data from both study areas will also be analyzed. This model will also be compared to the models developed for each study area, as well as Zmud and Wolf model. The effects of the logistic regression models in terms of the potential impact on calibration to travel demand models will also be discussed. The SAS programs created for the modeling of the data from each study area can be found in Appendix B. The data has been removed from the programs, but would be located at <DATA> within each program listed in the Appendix.

Laredo

The stepwise method of variable selection for logistic regression modeling identified two variables (of the twelve variables tested) as being significant to the model at the 0.001-level. The significance test for variables for entrance into the model is the Score Chi-square statistic. Once entered into the model, the Wald Chi-Square statistic is used to test variables for staying in the model. The two variables identified as significant using this method of variable selection are the trip purpose and the day of the

week for the assigned survey day. Using the specified 0.001 cutoff level for entry into the model, the remaining variables were not selected.

For this model, the R-square value was calculated at 0.1032, with a max-rescaled R-square value of 0.1379. The max-rescaled R-square statistic calculated in SAS exhibits similar traits to Pearson's R-square statistic, but applies itself to logistic regression rather than linear regression. The max-rescaled R-square is directly based on the R-square proposed by Cox and Snell (*28*) for logistic regression. It was found that the Cox and Snell R-square did not always have a true range of 0 and 1. A correct this, a max-rescaled R-square was proposed by Nagelkerke (*28*), who developed a way to normalize the Cox and Snell range to 0 and 1 by dividing the calculated Cox and Snell R-square by the maximum Cox and Snell R-square value that could have been obtained from the given model situation. Similar to Pearson's R Square, the max-rescaled R-square is logistic regression's version of the "coefficient of determination", in that it shows the percent of variability in the data that can be explained by the regression model.

The max-rescaled R-square value indicates that a model containing the two selected variables can explain roughly 14% of the variation that occurs when one chooses to report a trip. Although this is not a very high percentage, it does indicate that an additional analysis of the response (dependent) variable with respect to these two variables will still provide usable results and account for some of the variation that occurs. The Deviance and Pearson goodness-of-fit statistics are appropriate for checking for a significant lack of fit in the model when dealing with continuous variables (*29*). In

this model, however, categorical variables exist. The ineffectiveness in using the Deviance and Pearson goodness-of-fit tests is confirmed for this model, in which the Deviance statistic is found to be highly significant (<0.0001) and the Pearson statistic is found to be not significant (0.4922). For logistic regression models containing categorical variables, the most appropriate method for making an assessment of the goodness-of-fit is the Hosmer and Lemeshow goodness-of-fit test. For this model, the chi-square value of 13.4856 with 9 degrees of freedom did not indicate a significant lack of fit for the model (p-value=0.1418).

It was found that a comparison of the predicted probabilities to the actual responses for a pair of observations, known as "Association Statistics" in SAS, revealed that 61.5% of the pairs of observations were found to be concordant, 29.7% were found discordant, and 8.8% were found to be tied. Although only 61.5% concordant pairs were found, not a significant majority of the total observed pairs, it is higher than the baseline probability of occurrence (495 non-reported/917 observed = 54.0%), indicating that the selected variables are helping to improve the predictive capabilities of the model.

By examining the level of trip reporting that occurred with respect to the two variables selected in the model, specific sources of underreporting were identified. Table 4 shows the trip reporting accuracy of households participating in this survey by comparing the day of week the for the assigned travel day against the trip purpose. This data is presented graphically in Figure 5. Reporting accuracies on Wednesday and Thursday are lower than for other days of the week for each of the three trip purposes.

Specifically, Wednesdays were found to have the lowest trip reporting accuracy for both HBW and NHB trips.

TABLE 4 Laredo Trip Reporting Accuracy – Survey Day of Week vs. Trip Purpose

	Mon	Tue	Wed	Thu	Fri
HBW	93%	100%	62%	100%	100%
HBNW	49%	62%	47%	34%	45%
NHB	44%	34%	22%	20%	57%

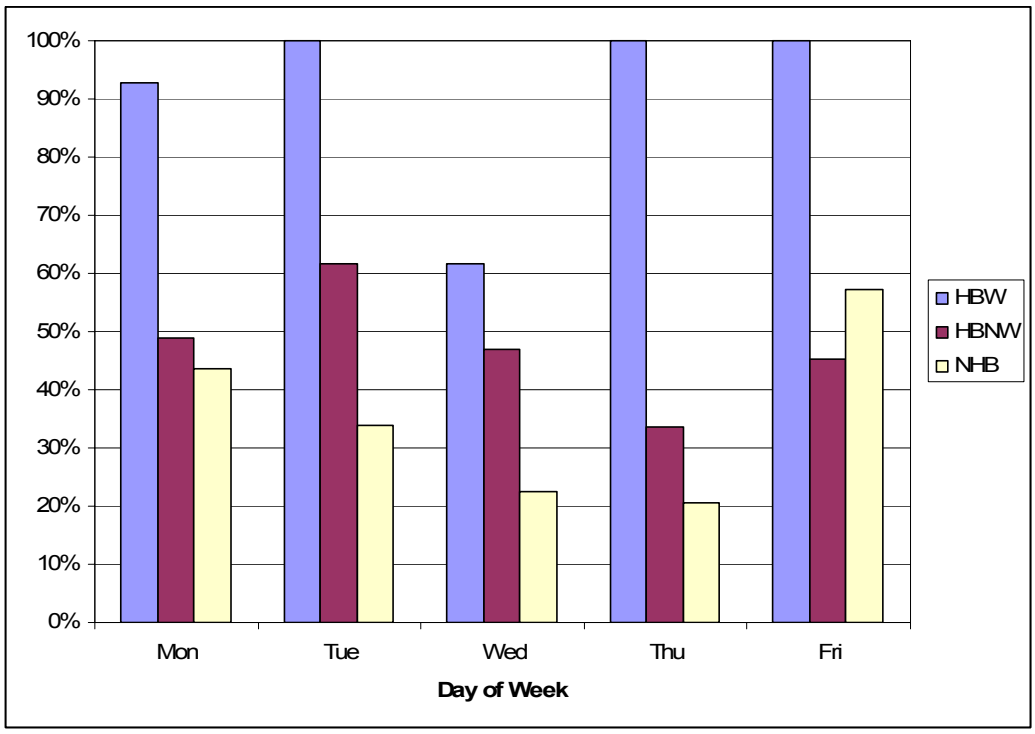

FIGURE 6 Laredo trip reporting accuracy – survey day of week vs. trip purpose.

Tyler-Longview

The stepwise method of variable selection for logistic regression modeling identified four variables (of the twelve variables tested) as being significant to the model

at the 0.001-level. The four variables identified as significant using this method of variable selection are the number of trips made by the household on the survey day, the trip purpose, the duration of the trip (in minutes), and the number of persons in the household. Using the specified 0.001 cutoff level for entry into the model, the remaining variables were not selected.

For this model, the R-square value was calculated at 0.1291, with a max-rescaled R-square value of 0.1814. The max-rescaled R-square statistic calculated in SAS exhibits similar traits to Pearson's R-square statistic, but applies itself to logistic regression rather than linear regression. The max-rescaled R-square statistic shows the percent of variability in the data that can be explained by the regression model. The max-rescaled R-square value for this model indicates that using the four variables selected can explain roughly 18% of the variation that occurs when one chooses to report a trip. This value is 4% greater than the stepwise model created for Laredo, and indicates that these four variables can improve prediction accuracy for trip reporting, and account for some of the variation that occurs. Similar to Laredo, the ineffectiveness in using the Deviance and Pearson goodness-of-fit tests is also confirmed for this model. The Deviance statistic is found to be highly significant (0.0006) and the Pearson statistic is found to be not significant (0.3724). For logistic regression models containing categorical variables, the most appropriate method for making an assessment of the goodness-of-fit is the Hosmer and Lemeshow goodness-of-fit test. For this model, the chi-square value of 14.1433 with 8 degrees of freedom did not indicate a significant lack of fit for the model (p-value=0.0781).

It was found that a comparison of the predicted probabilities to the actual responses for a pair of observations, revealed that 71.9% of the pairs of observations were found to be concordant, 27.7% were found discordant, and 0.3% were found to be tied. Although only 71.9% concordant pairs were found is not an overwhelming majority of the total observed pairs, it is higher than the baseline probability of occurrence (1289 reported/1878 observed = 68.6%), indicating that the selected variables are helping to improve the predictive capabilities of the model.

By examining the level of trip reporting that occurred with respect to the four variables selected in the model, specific sources of underreporting were identified. Table 5 shows the specific groups created from the combination of each variable that had a trip reporting accuracy of less than 50%. In this table, only cells in which the total number of observed GPS trips was ten or more ($n \geq 10$) were selected. From this table, it can be seen that there are certain variable classifications that are resulting in low accuracy rates. Specifically, non-home-based (NHB) trips originating in households that took a large number of trips (21+) during the survey day were the most underreported. Additionally, larger household sizes (3-4 and 5+ persons) also appear to have lower reporting accuracy for these trips than do smaller household sizes (1-2 persons).

TABLE 5 Tyler-Longview Trip Reporting Accuracy by Significant Variables

Number of trips by household	Trip Purpose	Trip Duration (mins)	Number of persons in household	Reporting Accuracy
21+	NHB	15+	3-4	12%
21+	NHB	5-10	5+	30%
21+	NHB	0-5	5+	36%
21+	NHB	0-5	3-4	37%
21+	NHB	5-10	3-4	38%
21+	NHB	10-15	3-4	38%
1-10	NHB	0-5	3-4	41%
11-20	NHB	0-5	1-2	41%

Combined Study Areas

The stepwise method of variable selection for logistic regression modeling identified six variables (of the twelve variables tested) as being significant to the model at the 0.001-level. The six variables identified as significant using this method of variable selection are the number of trips made by the household on the survey day, the trip purpose, the time of day the trip began, the number of employees in the household, and the number of vehicles available. Using the specified 0.001 cutoff level for entry into the model, the remaining variables were not selected.

For this model, the R-square value was calculated at 0.1150, with a max-rescaled R-square value of 0.1560. The max-rescaled R-square statistic calculated in SAS exhibits similar traits to Pearson's R-square statistic, but applies itself to logistic regression rather than linear regression. The max-rescaled R-square statistic shows the percent of variability in the data that can be explained by the regression model. The max-rescaled R-square value for this model indicates that using the six variables selected can explain roughly 16% of the variation that occurs when one chooses to report a trip. This value is 2% greater than the stepwise model created for Laredo, and 2% less than

the stepwise model created for Tyler-Longview. This also indicates that these six variables can improve prediction accuracy for trip reporting, and account for some of the variation that occurs. Similar to Laredo and Tyler-Longview, the ineffectiveness in using the Deviance and Pearson goodness-of-fit tests is also confirmed for this model. The Deviance statistic is found to be highly significant (<0.0001) and the Pearson statistic is found to be not significant (0.3564). For logistic regression models containing categorical variables, the most appropriate method for making an assessment of the goodness-of-fit is the Hosmer and Lemeshow goodness-of-fit test. For this model, the chi-square value of 12.2749 with 8 degrees of freedom did not indicate a significant lack of fit for the model (p-value=0.1394).

It was found that a comparison of the predicted probabilities to the actual responses for a pair of observations, revealed that 69.6% of the pairs of observations were found to be concordant, 30.0% were found discordant, and 0.4% were found to be tied. Although only 69.6% concordant pairs were found is not an overwhelming majority of the total observed pairs, it is higher than the baseline probability of occurrence (1711 reported/2795 observed = 61.2%), indicating that the selected variables are helping to improve the predictive capabilities of the model.

TABLE 6 Combined Study Areas Trip Reporting Accuracy by Significant Variables

Number of trips by household	Trip Purpose	Time of Day Trip Began	Day of Week	Number of Employed Persons	Number of Vehicles Available	Reporting Accuracy
11-20	NHB	8p-12a	M	1-2	1-2	0%
21+	NHB	4a-8a	W	1-2	3+	0%
21+	NHB	12p-4p	W	1-2	3+	6%
21+	NHB	8a-12p	T	1-2	1-2	12%
21+	NHB	12p-4p	F	1-2	3+	17%
11-20	NHB	8a-12p	W	0	1-2	24%
21+	NHB	12p-4p	T	1-2	1-2	26%
21+	NHB	4p-8p	F	1-2	3+	27%
11-20	NHB	4a-8a	M	1-2	1-2	27%
11-20	NHB	12p-4p	R	3+	3+	29%

Table 6 shows the combination of significant variables that resulted in the lowest reporting accuracies. In this table, only cells in which the total number of observed GPS trips was ten or more ($n \geq 10$) were selected. It is clear from this table that, similar to each individual study area, NHB trips are the most underreported of any trip purpose. Looking at the number of trips per household, it can also be seen that households that made more trips during the day (11-20 and 20+) had a low reporting accuracy more frequently than households that made fewer trips (1-10). In fact, households that made few trips (1-10) did not have cells in which trip reporting accuracies were low enough to be included in this table.

Comparing to Other Models

The stepwise selection models were used to identify variables found to be significant by using cutoff p-values for entry into the model and cutoff p-values for staying in the model. Variables not found to be significant at the 0.001-level were not

entered into the model. To compare modeling results for this research to those found by Zmud and Wolf (*10*), it was necessary to create a full model. This was because Zmud and Wolf only used a full model for their regression model. For each study area, a full model was created that utilized every variable, even those not found to be significant. These models exhibited slightly higher R-square values than their stepwise counterparts, but this came at a price – the much larger number of independent variable inputs required for prediction. For example, the Tyler-Longview stepwise model yielded a max-rescaled R-square value of 0.1814 using only four variables, and the full model yielded a max-rescaled R-square value of 0.2120 using all 12 variables. In other words, the eight additional independent variables to the stepwise model only improved the model's account for variation by 3% - not a drastic improvement considering the required number of variables to the model increased threefold.

Table 7 identifies which variables were found to be significant at the 0.001-level for each of the models tested and for the Zmud and Wolf model (*10*). From this table, it can be seen that trip purpose was found to be significant in all six of the models that were developed. None of the remaining eleven variables were found to be significant in all six models.

TABLE 7 Significance Levels of Variables in Each Model

Independent Variable	Laredo		Tyler-Longview		Combined		Zmud and Wolf (10)
	Stepwise	Full	Stepwise	Full	Stepwise	Full	
Number of trips taken during day by household	-	0.001	**0.000**	**0.000**	**0.000**	**0.000**	not tested
Purpose of trip	**0.000***	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	not tested
Time of day trip occurred	-	0.002	-	0.176	**0.000**	**0.001**	not tested
Length of trip in minutes	-	**0.000**	**0.000**	0.007	-	**0.000**	**0.000**
Length of trip in miles	-	**0.000**	-	0.077	-	**0.001**	not tested
Day of week survey was conducted	**0.000**	**0.000**	-	0.070	**0.000**	**0.000**	not tested
Number of persons in household	-	0.864	**0.001**	**0.000**	-	0.992	0.002
Number of employed persons in household	-	0.090	-	**0.000**	**0.000**	**0.000**	0.864
Income of household	-	0.298	-	**0.001**	-	0.017	**0.000**
Number of vehicles available to household	-	0.046	-	0.023	**0.000**	**0.001**	**0.000**
Household residence type	-	0.038	-	0.117	-	0.154	not tested
Is vehicle used for commercial purposes	-	0.098	-	0.213	-	0.293	not tested

* bold indicates significant at 0.001-level

A comparison of the Zmud and Wolf model to the full models for each study area did not identify any consistent patterns. The duration of a trip was found to be significant in the Laredo full model, the combined full model, and the Zmud and Wolf model, but not in the Tyler-Longview full model. The household income was found to be significant only in the Tyler-Longview full model and the Zmud and Wolf model, but not the Laredo full model and the combined full model. The number of vehicles available to a household was found to be significant only in the combined full model and the Zmud and Wolf model, but not in the individual full models for each study area.

SUMMARY AND CONCLUSIONS

In this thesis, a methodology for conducting logistic regression modeling to identify household- and trip-related variables contributing to trip underreporting is presented. The methodology shown here provides a framework for data analysis of both types of survey data, in order to allow for a determination of the underreported trips to be identified. Although specific software packages are discussed here, the processes applied in each are not specific to the software package used, and alternative software is also appropriate. Additionally, the methodology shown here has built on previous research in this area by incorporating new variables into the model for testing that have not been tested previously. This section summarizes the methodology presented in this thesis; discusses the results obtained from it; and identifies additional research that may of interest to the thesis topic.

Methodology

The analysis software used in this thesis provides tools that can be used to collect, maintain, and analyze travel survey data. For handling the two forms of spatial data provided by the two surveys, the use of a geodatabase significantly reduced the hard disk storage requirements, and facilitated a much simpler form of data management during the analysis. Although the multi-step heuristic procedure presented here for identifying vehicle trips from GPS data is effective, it should be noted that the need for accurate visual identification of trip ends is also required. Because there is no precise way in which this can be done, some level of judgment is always required in making the

visual assessment of trip ends. An alternative dwell time threshold value of 135 seconds was used in making a determination of trip ends from the GPS data, and it was found that this higher threshold eliminated some of the false-positive trip end detections that would have been made with using the original threshold of 120 seconds.

The trip comparison table was presented as a method of organizing and displaying trip data obtained from both survey types. This is crucial is making a comparison of the two sets of travel data in order to determine the vehicle trips found in the GPS data that were not reported by the household in the CATI.

The statistical software SAS was used to conduct the logistic regression modeling of the data. A total of twelve unique independent variables (Table 2) were tested against a dichotomous dependent variable (Table 3), whether or not a trip was reported. Through the creation of SAS programs for analyzing the data, both a stepwise model and full model were created for each study area. The stepwise model added significant variables to the model as they were found, and stopping when no additional significant variables were found. The full model used all twelve of the independent variables for modeling, and did not attempt to add or remove any of them. Additional modeling parameters, such as p-value cutoff levels and goodness-of-fit tests were added to the SAS programs to customize the modeling procedure and display significance tests of interest.

Discussion of Results

For each study area, a stepwise selection model was used to identify independent variables that are significant to trip underreporting at the 0.001-level. This modeling scenario identified two variables in Laredo as significant to trip underreporting – trip purpose and the day of the week the survey was conducted. An additional analysis of the underreporting that occurred revealed that NHB trips were underreported every day of the week. It was also found that HBW trips were reported accurately every day of the week except for Wednesday (Table 4 and Figure 5). It is possible that this lower reporting level on Wednesdays could be due to households losing track of which trips were made on which day during the week. It could also be due to survey participants more easily remembering trips on days that occurred early or late in the week (Monday or Friday), because they could be more easily tied to the beginning and ending of weekend activities.

For the Tyler-Longview study area, four of the independent variables tested in the regression model were found to be significant. These included trip purpose, the number of trips made by the household, the duration of the trip, and the number of persons in the household. Dissimilar to Laredo, the day of the week the survey was conducted was not found to be a significant factor contributing to underreporting. The trip purpose, however, was found to be highly significant to the model. It was also found that NHB trips were the most underreported trip. Every possible combination of the four variables selected to the model revealed that those in which the overall accuracy

was less than 50%, were based on NHB trips (Table 5). It was also found that the lowest reporting accuracies were in households that had a high number (21+) of trips.

The stepwise logistic regression model created from the data for both study areas found six variables to be significant at the 0.001-level. These variables were trip purpose, the number of trips made by the household, the time of day the trip began, the day of the week the survey was conducted, the number of employed persons in the household, and the number of vehicles available to the household. The six remaining variables were not found to be significant at the 0.001-level. Further analysis of trip reporting accuracy with respect to these six variables indicated that certain combinations of them yield a reporting accuracy of 0% (Table 6). In other words, there were specific sets circumstances found in the data from both study areas in which none of the trips were reported.

It was found that trip purpose was considered highly significant in all of the models tested, which confirms the previous assertions that trip purpose is a likely significant indicator in determining whether or a not a trip is accurately reported. A comparison of the Zmud and Wolf model to the full model for the combined study area (Table 7) showed that there were two variables that were found to be significant in both models – the trip duration and the number of vehicles available to a household. This similarity in significant variables for the two models strongly indicates that the effect of these variables on reporting accuracy could exist in all urban areas. Considering the dependent variable being modeled, it is logical that these variables would impact it. The

longer a trip lasts, the more likely a survey respondent should be able to remember the trip, and thus be able to report it in the CATI survey.

Additionally, the variables household residence type and the use of a household vehicle for commercial purposes were not found to be significant in any of the models tested. The lack of significance for household residence type in all of the models indicates that there is generally no difference in reporting accuracy for households living in single-family residences (homes and mobile homes) and households living in multi-family residences (apartments, condos, dorms, etc.). Similarly, the lack of significance for the use of the vehicle for commercial purposes indicates that households who use their vehicle for reasons other than personal also show no difference in reporting accuracy than households who do not.

Additional Research

The research presented in this thesis provided an analytical framework in which a variety of additional research may be performed. In addition to looking at the number of trips made by each vehicle and household, other details relating to individual trips that may be of interest include route choice and the impact of trip reporting on activity modeling.

The vehicle miles traveled (VMT) in an urban area is often used as a calibration guide for estimating travel demand. Based on this, additional research that explores the average trip length for reported and non-reported trips could potentially reveal deficiencies in VMT estimations which are used to calibrate travel demand models. The

research presented here shows that trip purpose is an important factor when looking at trip underreporting. Although an analysis of the reporting accuracy for each trip purpose can be used to expand the total number of trips estimated for each trip purpose within an urban area, it is also important to look at the average trip length for each of these trip purposes. For example, if it is shown that HBW trips that are not reported are found to be significantly longer than HBW trips that are reported, this can result in a severe underestimation of VMT generated from HBW trips despite properly estimating the number of trips in the urban area from the reporting accuracy.

The impact on VMT can also be viewed in terms of trip chaining. Trip chaining occurs by linking together multiple consecutive trips together to form a single, larger trip. When looking at trip underreporting, it is often found that many of the trips that are underreported are simply quick stops that occur along the route to a farther destination. In the CATI, the respondent often only reports the final destination as the trip end. Although a comparison of this CATI trip information to that found in the GPS reveals that the respondent failed to report some of the trips it made, a comparison of the VMT generated from each of the two methods shows that they were very similar. The impact of these trip chaining effects should be considered when using reporting accuracy rates.

The logistic regression modeling performed in this research also provides a variety of additional research topics that may be of interest. In the research presented here, only the trip start time variable was categorized – the remaining variables were kept in their original continuous or categorical form. This was done to minimize model

complexity and maximize the model specifications. Categorizing variables can lead to improved model results, but also leads to a more generalized interpretation of the categorized variable in the model. Additionally, the new interpretation of the model parameter does not conform to the original variable measurement units.

The level at which variables are categorized can effect their significance to a model. A detailed analysis of alternative levels of variable categorization could provide insight into how these variables should be categorized for a logistic regression model. This analysis may also reveal that the simplification (decreasing the number of categories for the respondent to choose from) of certain questions on the CATI survey could provide better modeling results than using the original categories. This can lead to better survey results, and a more appropriate interpretation of survey respondent answers.

In additional to exploring alternative modeling options within logistic regression, the application of other methods of data analysis, such as decision trees and artificial neural networks may also be appropriate. These alternative methods of data analysis can also be used to explore a dependent variable in relation to a series of independent variables, and can identify which of the independent variables have the highest impact on the dependent variable.

REFERENCES

1. Pearson, D.F. and G.B. Dresser. *Urban Travel Demand Modeling Data*. 1994, Texas Transportation Institute: College Station, TX. pp. 122.

2. Pearson, D.F., A.F. Gamble, and M. Salami. *Urban Travel in Texas: an Evaluation of Travel Surveys*. 1996, Texas Transportation Institute: College Station, TX. pp. 528.

3. Wolf, J., R. Guensler, S. Washington, and L. Frank. *Use of Electronic Travel Diaries and Vehicle Instrumentation Packages in the Year 2000.* TRB Transportation Research Circular, 2000(E-C026): pp. 413-429.

4. Pearson, D.F. Global positioning system (GPS) and travel surveys: results from the 1997 Austin household survey. In *Eighth Conference on the Application of Transportation Planning Methods*. 2001. Corpus Christi, TX.

5. NuStats. *Kansas City Regional Household Travel Survey: GPS Study Final Report*, S. Bricka, Ed. 2003, NuStats: Austin, TX. pp. 19.

6. Murakami, E. and D.P. Wagner. *Can using global positioning system (GPS) improve trip reporting?* Transportation Research Part C: Emerging Technologies, 1999. **7**(2-3): pp. 149-165.

7. Pierce, B., J. Casas, and G. Giaimo. Estimating trip rate under-reporting: preliminary results from the Ohio Household Travel Survey. In *TRB 2003 Annual Meeting*. 2003. Washington, DC.

8. Zhou, J. and R. Golledge. A GPS-based analysis of household travel behavior. In *WRSA Annual Meeting*. 1999. Kauai, HI.

9. Zito, R. and M.A.P. Taylor. The use of GPS in travel-time surveys. *Traffic Engineering & Control*, 1994. 35(12): pp. 685-690.

10. Zmud, J. and J. Wolf. Identifying the correlates of trip misreporting - results from the California statewide household travel survey GPS study. In *10th International Conference on Travel Behaviour Research*. 2003. Lucerne, Switzerland.

11. Wolf, J., M. Loechl, J. Myers, and C. Arce. Trip rate analysis in GPS-enhanced personal travel surveys. In *Transport Survey Quality and Innovation*. Elsevier: London. 2003. pp. 483-498.

12. Wolf, J., R. Guensler, and W. Bachman. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In *TRB 2001 Annual Meeting*. 2001. Washington, DC.

13. Frihida, A., D.J. Marceau, and M. Theriault. Spatio-temporal object-oriented data model for disaggregate travel behavior. *Transactions in GIS*, 2002. 6(3): pp. 277-294.

14. Jiang, B. and C. Claramunt. Integration of space syntax into GIS: new perspectives for urban morphology. *Transactions in GIS*, 2002. 6(3): pp. 295-309.

15. Thill, J.C. Geographic information systems for transportation in perspective. *Transportation Research Part C: Emerging Technologies*, 2000. 8(1-6): pp. 3-12.

16. Taylor, M.A.P., J.E. Woolley, and R. Zito. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C: Emerging Technologies*, 2000. 8(1-6): pp. 257-285.

17. Czerniak, R.J., R.L. Genrich, K.E. Johnson, B. Lewis, R.W. McCrary, et al. *Collecting, processing, and integrating GPS data into GIS. A Synthesis of Highway Practice.* 2002, Washington, DC: National Academy Press. p. 66.

18. Denstadli, J.M. and R.J. Hjorthol. Testing the accuracy of collected geoinformation in the Norwegian personal travel survey-experiences from a pilot study. *Journal of Transport Geography*, 2003. 11(1): pp. 47-54.

19. Wolf, J., S. Hallmark, M. Oliveira, R. Guensler, and W. Sarasua. Accuracy issues with route choice data collection by using global positioning system. *Transportation Research Record*, 1999. 1660: pp. 66-74.

20. 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 Record*, 2001. 1768: pp. 125-134.

21. Wolf, J. Applications of new technologies in travel surveys. In *2004 International Conference on Transport Survey Quality and Innovation*. 2004. Costa Rica.

22. Mackett, R.L. Why do people use their cars for short trips? *Transportation*, 2003. 30: pp. 329-349.

23. Gliebe, J.P. and F.S. Koppelman. Modeling household activity-travel interactions as parallel constrained choices. *Transportation*, 2005. 32: pp. 449-471.

24. Li, J. Explaining high-occupancy-toll lane use. *Transportation Research Part D*, 2001. 6: pp. 61-74.

25. Pampel, F.C. *Logistic Regression - A Primer*. 1st edition. M.S. Lewis-Beck ed. Vol. 132. 2000, Thousand Oaks, CA: Sage Publications, Inc. p. 86.

26. *Laredo Market Report: Demographics*. 2005, Texas A&M University, Real Estate Center: College Station, TX.

27. U.S. Census. *Annual Estimates of the Population for Counties: April 1, 2000 to July 1, 2004*. 2005. http://www.census.gov/popest/counties/CO-EST2004-01.html

28. SAS Institute, Inc. *SAS/STAT User's Guide, Version 8*. 2000, Cary, NC: SAS Institute Inc.

29. Stokes, M.E., C.S. Davis, and G.G. Koch. *Categorical Data Analysis Using the SAS System*. 1995, Cary, NC: SAS Institute Inc. p. 499.

30. *Laredo Survey Data Specifications*. 2002, Texas Department of Transportation, Travel Survey Program: Austin, TX.

APPENDIX A

TABLE A-1 Example of Household Data File Format (Laredo) (*30*)

Item	Begin	End	Type	Format	Description
1. Record Type	1	2	Numeric RJ	I2	Code indicating type of record. Here it should be 1.
2. Sample Number	3	9	Numeric RJ	I7	Unique non-zero number assigned to each household participating in survey.
3. Phone	10	21	Alphanum LJ	A12	Phone number of household.
4. Month	22	23	Numeric RJ	I2	Month of travel day.
5. Day	24	25	Numeric RJ	I2	Day of the month of travel.
6. Day of Week	26	26	Numeric RJ	I1	Day of the week travel was recorded; 1-Monday, 2-Tuesday, 3-Wednesday, 4-Thursday, 5-Friday.
7. Travel Assignment Number	27	29	Numeric RJ	I3	No description
8. Advance Letter	30	31	Numeric RJ	I2	Code indicating if household received advance letter; 1-Yes, 2-No, 98-Don't Know, 99-Refused.
9. Address	32	91	Alphanum LJ	A60	Street address or nearest cross streets of household.
10. City	92	121	Alphanum LJ	A30	City where household is located.
11. Zip Code	122	126	Numeric RJ	I5	Zip code of household address.
12. Zone	127	131	Numeric RJ	I5	Zone number where household is located. Unknown zones should be coded 8888**.**
13. Longitude	132	141	Numeric RJ	F10.0	Longitude of household address. If unknown, it should be coded 888.8888.
14. Latitude	142	151	Numeric RJ	F10.0	Latitude of household address. If unknown, it should be coded 888.8888.
15. Number Persons	152	153	Numeric RJ	I2	Number of persons living in residence.
16. Persons 5+ Age	154	155	Numeric RJ	I2	Number of persons 5 years of age or older living in household.
17. Number Employed	156	157	Numeric RJ	I2	Number of persons in household that are employed either full or part time.
18. Vehicles Available	158	159	Numeric RJ	I2	Number of cars, vans, light trucks, motorcycles available for use; 98-Don't Know, 99-Refused.
19. Bikes	160	161	Numeric RJ	I2	Number of working bicycles available for use by members of household; 98-Don't Know, 99-Refused.
20. Residence	162	163	Numeric RJ	I2	Code indicating the type of residence. See below for code definitions.
21. Other Residence	164	188	Alphanum LJ	A25	If residence is coded as "other", this field contains a description of the type of residence.
22. Tenure	189	190	Numeric RJ	I2	Code indicating number of years at residence; 0-<1yr, 1-one year, 2-two years, 3-three years, 4-four years, 5-five or more years.
23.Previous Residence	191	191	Numeric RJ	I1	If tenure was less than five years, this code indicates if previous residence was in the Laredo area; 1-Yes, 2-No.
24. Previous Zip Code	192	196	Numeric RJ	I5	If tenure was less than five years, this is the zip code of the previous residence.
25. HH Factors	197	198	Numeric RJ	I2	Code indicating factors that influenced their decision to locate in their current household. See code definitions.
26. Other Factors	199	228	Numeric RJ	A30	Other factors influencing their decision to locate in their current household.
27. Income	229	230	Numeric RJ	I2	Code indicating combined annual income of all household members. See codes below.
28. Sample HH Income	231	232	Numeric RJ	I2	Household income stratification for sampling quota. 1=<10k, 2=10k - <20k, 3=20k - <35k, 4=35k - <50k, 5= 50k or more.
29. Day Visitors	233	234	Numeric RJ	I2	Number of non-family persons that stopped at this residence for any reason on the travel day; 98-Don't Know, 99-Refused.
30. Overnight Visitors	235	236	Numeric RJ	I2	Number of overnight visitors at this residence during their travel day.

TABLE A-1 (Continued)

Item	Begin	End	Type	Format	Description
31. Delivery Vehicle	237	237	Numeric RJ	I1	Code indicating if someone in household drives a form of delivery vehicle; 1-Yes, 2-No, 9-Don't Know / Refused.
32. Number Delivery Driver	238	239	Numeric RJ	I2	Number of persons in household that are delivery drivers or travel within study area as part of their work.
33. Phone Service	240	241	Numeric RJ	I2	Number of times within past 12 months household was without telephone service.
34. Time Without	242	243	Numeric RJ	I2	Code indicating the average length of time household was without phone service. See code definitions below.
35. Share Phone	244	245	Numeric RJ	I2	Number of households that share a phone line with this household.
36. GPS House	246	246	Numeric RJ	I1	Code indicating if household vehicles had GPS equipment installed for GPS survey.
37. Total HH Trips	247	249	Numeric RJ	I3	The total combined number of all trips made by all persons in the household on the assigned travel day.

TABLE A-2 Example of Household Data File Format Codes (Laredo) (*30*)

Residence (Household Item 20)	**HH Factors (Household Item 25)**	**Time Without (Household Item 34)**
1 – Unattached Single Family Home	1 – Price of Property	1 – Less than one week
2 – Condo	2 – Taxes	2 – One week to less than two weeks
3 – Duplex	3 – Proximity to Work	3 – Two weeks to less than one month
4 – Apartment	4 – School District	4 – One month to less than four months
5 – Mobile Home	5 – Proximity to School	5 – Three months to less than six months
6 – Other	6 – Character of Neighborhood or Area	6 - Six months to less than one year
99 – Don't Know / Refused	7 – Access to Public Transportation	7 – One year or more
	97 – Other	97 – Don't know
	99 – Don't Know / Refused	99 – Refused
Household Income Codes (Household Item 27)		
1 – Less than $5,000	7 - $30,000 to $34,999	13 - $100,000 to $124,999
2 - $5,000 to $9,999	8 - $35,000 to $39,999	14 - $125,000 to $149,999
3 - $10,000 to $14,999	9 - $40,000 to $49,999	15 - $150,000 or more
4 - $15,000 to $19,999	10 - $50,000 to $59,999	99 – Refused
5 - $20,000 to $24,999	11 - $60,000 to $74,999	
6 - $25,000 to $29,999	12 - $75,000 to $99,999	

TABLE A-3 Example of Person Data File Format (Laredo) (*30*)

Item	Begin	End	Type	Format	Description
1. Record Type	1	2	Numeric RJ	I2	Code indicating type of record, here it should be 2.
2. Sample Number	3	9	Numeric RJ	I7	Unique non-zero number assigned to each household participating in survey. This number should match the sample number of the above record.
3. Person Number	10	12	Numeric RJ	I3	Number assigned to each person in the household with 0 assumed to be the head of household.
4. Relationship	13	14	Numeric RJ	I2	Code indicating relationship of person to the head of household. See code definitions below.
5. Sex	15	16	Numeric RJ	I2	Sex of person; 1-Male, 2-Female, 98- Don't Know, 99-Refused.
6. Ethnicity	17	18	Numeric RJ	I2	Race or ethnicity of person. See code definitions below.
7. Ethnicity Other	19	78	Alphanum RJ	A60	Description of other ethnicity which is not included in code definitions.
8. Age	79	81	Numeric RJ	I3	Age of person. 999-Don't know / Refused.
9. Licensed Driver	82	83	Numeric RJ	I2	Code indicating if person is a licensed driver; 1-Yes, 2-No, 98-Don't Know, 99-Refused.
10. Employment	84	85	Numeric RJ	I2	Code indicating if person is employed in a paying or volunteer job; 1-Yes, 2-No, 99-Refused.
11. Employment Status	86	87	Numeric RJ	I2	If person is employed, this is a code number indicating the person's employment status. See code definitions.
12. Hours	88	90	Numeric RJ	I3	On average, the number of hours worked per week. 999-varies from week to week.
13. Not Employed	91	92	Numeric RJ	I2	Code indicating current status if person is not employed. See code definitions below.
14.Not Employed Other	93	152	Alphanum LJ	A60	Description of employment status if none of the options in the employment status code are applicable.
15. Delivery	153	154	Numeric RJ	I2	Code indicating if person is a delivery driver or not; 1-Yes, 2-No, 99-Refused.
16. Flex Time	155	156	Numeric RJ	I2	Code indicating if person's employer allows them to work flexible hours or the hours are fixed; 1-Flexible / Variable, 2-Fixed / Unchanging, 99-Don't Know / Refused.
17. Job	157	158	Numeric RJ	I2	Code indicating if person has more than one paying job; 1-Yes, 2-No, 99-Refused.
18. Employer Name	159	218	Alphanum LJ	A60	Name of person's primary employer.
19. Workplace Type	219	220	Numeric RJ	I2	Code indicating type of workplace where person works. See code definitions below.
20. Other Workplace	221	250	Alphanum LJ	A30	Description of workplace type if "other" is coded.
21. Workplace Address	251	310	Alphanum LJ	A60	Street address of workplace or nearest intersecting street names.
22. Workplace City	311	340	Alphanum LJ	A30	City where workplace is located.
23. Workplace County	341	342	Numeric RJ	I2	Code indicating county where workplace is located; 1-Webb, 2-Mexico; 3 – Other; 99 – Refused/Unknown
24. Zip Code	343	347	Numeric RJ	I5	Zip code or workplace address.
25. Work Zone	348	352	Numeric RJ	I5	Zone where workplace is located. If unknown, it should be coded 8888. Locations in Mexico should be coded 7777 and addresses outside of Webb County but within Texas should be coded using the Statewide Zone System and preceded by the number 1 in column 348. Addresses outside of Texas and Mexico should be coded using 9999.
26. Longitude	353	362	Numeric RJ	F10.0	Longitude of workplace location. If unknown, it should be coded 888.8888. Workplaces in Mexico should be coded 777.7777 and workplace addresses outside of Webb County should be coded 999.9999.
27. Latitude	363	372	Numeric RJ	F10.0	Latitude of workplace location. If unknown, it should be coded 888.8888. Workplaces in Mexico should be coded 777.7777 and workplace addresses outside of Webb County should be coded 999.9999.

TABLE A-3 (Continued)

Item	Begin	End	Type	Format	Description
28. Days Worked	373	374	Numeric RJ	I2	Number of days per week person works. 98-Don't Know, 99-Refused.
29. Work at Home	375	376	Numeric RJ	I2	Out of the last seven days, the number of days worked at home instead of going to work. Valid responses 0-7, 98-Don't Know, 99-Refused.
30. Second Job Type	377	378	Numeric RJ	I2	Code indicating type of workplace where person works at second job. See code definitions below.
31. Other Job Type	379	438	Alphanum LJ	A60	Description of workplace type for second job if "other" is coded.
32. Total Hours	439	441	Numeric RJ	I3	Total hours on average person works per week at all jobs. 888-Don't Know, 999-Refused.
33. Student Status	442	443	Numeric RJ	I2	Code indicating if person is enrolled in any type of school; 1-Yes, 2-No, 98-Don't Know, 99-Refused.
34. School Type	444	445	Numeric RJ	I2	Code indicating type of school attended. See code definitions below.
35. School Type Other	446	505	Alphanum LJ	A60	Description of 'other' if other is coded as school type.
36. Hours Enrolled	506	507	Numeric RJ	I2	If person is enrolled in a college, trade school, etc., code indicates if person is enrolled for 12 or more hours; 1-Yes, 2-No, 98-Don't Know, 99-Refused.
37. Bike Use	508	509	Numeric RJ	I2	Number of days person rode bike in last seven days. 98-Don't Know, 99-Refused.
38. Bike Purpose	510	511	Numeric RJ	I2	Code indicating the most common trip purpose for person's bike trips. See code definitions below.
39. Disability	512	513	Numeric RJ	I2	Code indicating if person has transportation disability; 1-Yes, 2-No, 98-Don't Know, 99-Refused.
40. Travel	514	515	Numeric RJ	I2	Code indicating if person traveled on the designated travel day; 1-Yes, 2-No, 99-Indication person was out of town or away from the residence for the entire day and night of their travel day.
41 Person trips	516	518	Numeric RJ	I3	The total number of trips the person made on his/her travel day.
41. Why No Travel	519	578	Alphanum LJ	A60	Description of why the person did not make any trips on the travel day.
42. Diary Use	579	580	Numeric RJ	I2	Code indicating if person used diary or if information is based on memory or provided by a proxy. 1 – Used diary; 2 – Did not use diary; 3 – Do not know; 4. – Proxy provided information; 99 - Refused
43. Proxy	581	582	Numeric RJ	I2	1–Yes; 2–No;9–Don't Know/Refused
44. Date data was retrieved.	583	586	Numeric RJ	I4	The month and day the data was retrieved. Record all months as 2 digits and all days as 2 digits with the month preceding the day. Example: April 1st should be coded as 0401.

TABLE A-4 Example of Person Data File Format Codes (Laredo) (*30*)

Relationship (Person Item 4)	**Ethnicity (Person Item 6)**	**Status for not Employed (Person Item 13)**
0 – Head of Household 1 – Husband / Wife / Unmarried Partner 2 – Mother / Father / In-law 3 – Brother / Sister 4 – Grandfather / Grandmother 5 – Son / Daughter 6 – Aunt / Uncle 7 – Other Relative 8 – Other Non-Relative 9 – Household Help 99 – Don't Know / Refused	1 – Black / African American 2 – Hispanic / Mexican American 3 – Asian / Pacific Islander 4 – Native American 5 – White / Caucasian 6 – Other Group 99 – Don't Know / Refused	1 – Retired 2 – Disability Status 3 – Homemaker 4 – Looking for Work 5 – Not Looking for Work 6 – Student 98 – Other 99 – Refused
Type of Workplace (Person Item 19)	**School Type (Person Item 34)**	**Bike Trip Purpose (Person Item 38)**
1 – Office 2 – Retail 3 – Industrial / Manufacturing 4 – Medical 5 – Education – Day Care / K-12th 6 – Education – College, Trade, Other 7 – Government 8 – Residential Type Work Place 9 – Other 99 – Don't Know / Refused	1 – Day Care / Pre-School 2 – K-12th 3 – Post Secondary 4 – Other 99 – Don't Know / Refused	1 – Work 2 – School 3 – Shopping 4 – Visiting 5 – Recreation / Exercise 6 – Other 99 – Don't Know / Refused
Employment Status (Person Item 11)		
1 – Employed full time 30 or more hours per week 2 – Employed part time less than 30 hours per week 3 – Self employed full time 30 or more hours per week 4 – Self employed part time less than 30 hours per week 99 – Refused / Unknown		

TABLE A-5 Example of Vehicle Data File Format (Laredo) (*30*)

Item	Begin	End	Type	Format	Description
1. Record Type	1	2	Numeric RJ	I2	Code indicating type of record, here it should be 3.
2. Sample Number	3	9	Numeric RJ	I7	Unique non-zero number assigned to each household participating in survey.
3. Vehicle Number	10	11	Numeric RJ	I2	Unique non-zero number assigned to vehicle.
4. Type of Vehicle	12	13	Numeric RJ	I2	Code indicating type of vehicle. See code definitions below.
5. Other Vehicle Type	14	48	Alphanum LJ	A35	Other vehicle type not listed in vehicle code below.
5. Year	49	52	Numeric RJ	I4	Year vehicle was manufactured; 9998-Don't Know, 9999-Refused.
6. Make	53	54	Numeric RJ	I2	Make of vehicle. See vehicle make code below.
8. Other Make	55	114	Alphanum LJ	A60	Specify other make of vehicle if not included in vehicle make code below.
7. Model	115	174	Alphanum LJ	A60	Model of vehicle.
8. Type of Fuel	175	175	Numeric RJ	I1	Type of fuel used by vehicle; 1-Gasoline, 2-Diesel, 3-Other, 8-Don't Know, 9-Refused.
9. Other Fuel Type	176	190	Alphanum LJ	A15	Other type of fuel specified, e.g. propane, natural gas, electric, etc.
10. Classification	191	192	Numeric RJ	I2	Code indicating vehicle classification. See code definitions below.
11. Commercial Use	193	194	Numeric RJ	I2	Code indicating if vehicle is used for commercial purposes; 1-Yes, 2-No, 99-Don't Know / Refused.
12. Beginning Mileage	195	202	Numeric RJ	I8	Odometer reading on vehicle at beginning of travel day. Don't Know, 99999999. Refused, 99999998.
13. Ending Mileage	203	210	Numeric RJ	I8	Odometer reading on vehicle at end of travel day.

TABLE A-6 Example of Vehicle Data File Format Codes (Laredo) (*30*)

Vehicle Classification Codes (Vehicle Item 10)		**Type of Vehicle Codes (Vehicle Item 4)**
1 – Light Duty Gas Vehicle	6 – Light Duty Diesel Truck	1 – Motorcycle
2 – Light Duty Gas Truck Type 1	7 – Heavy Duty Diesel Truck	2 – Car
3 – Light Duty Gas Truck Type 2	8 – Motorcycle	3 – Van
4 – Heavy Duty Gas Truck	9 – Alternative Fuel Vehicle	4 – Sport Utility Vehicle
5 – Light Duty Diesel Vehicle	99 – Don't Know / Refused	5 – Pickup Truck
		6 – Cargo Van
		7 – Other
		99 – Refused / Unknown
Vehicle Make Codes (Vehicle Item 6)		
01 – Acura	18 – Jeep	35 – Subaru
02 – Audi	19 – Kawasaki	36 – Suzuki
03 – BMW	20 – KIA	37 – Toyota
04 – Buick	21 – Lexus	38 – Volkswagen
05 – Cadillac	22 – Lincoln	39 – Volvo
06 – Chevrolet	23 – Mazda	40 – Yamaha
07 – Chrysler	24 – Mercury	41 – Daewoo
08 – Dodge	25 – Mercedes-Benz	97 – Other
09 – Ford	26 – Mitzubitshi	98 – Don't Know
10 – Geo	27 – Nissan	99 – Refused
11 – GMC	28 – Oldsmobile	
12 – Harley Davidson	29 – Plymouth	
13 – Honda	30 – Pontiac	
14 – Hyundai	31 – Porsche	
15 – Infiniti	32 – Range Rover	
16 – Isuzu	33 – Saab	
17 – Jaguar	34 – Saturn	

TABLE A-7 Example of Trip Data File Format (Laredo) (*30*)

Item	Begin	End	Type	Format	Description
1. Record Type	1	2	Numeric RJ	I2	Code indicating type of record. Here it should be 4.
2. Sample Number	3	9	Numeric RJ	I7	Unique non-zero number assigned to each household participating in survey. This number must match the number used for the same household and recorded in the Household Data File.
3. Month	10	11	Numeric RJ	I2	Month of survey day.
4. Day	12	13	Numeric RJ	I2	Day of the month of the survey.
5. Person Number	14	15	Numeric RJ	I2	Number assigned to the person doing this activity.
6. Activity/Trip Number	16	17	Numeric RJ	I2	Activity number. First activity for each person will be recorded as 0 for where their day began. Each subsequent activity should be numbered sequentially as 1, 2, 3, etc.
7. Activity Type	18	19	Numeric RJ	I2	Code indicating the type of activity. See activity codes below. This may be posted coded. For activity 0 (where day began), this should be coded as a 1 if it began at home, 4 if day began at work, or as 20 if it began at another location. If this is coded as 20, the activity description should be included in item 8.
8. Activity Description	20	80	Alphanum LJ	A60	Description of Activity.
9. Location	81	110	Alphanum LJ	A30	Name of location where activity took place.
10. Location Address	111	170	Alphanum LJ	A60	Street address of location or name of nearest intersecting streets.
11. Location City	171	200	Alphanum LJ	A30	Name of city where location is.
12. Location County	201	202	Numeric RJ	I2	Code indicating county where location is; 1-Webb, 2-Mexico; 3 – Other; 99 – Unknown/Refused
13. Zip Code	203	207	Numeric RJ	I5	Zip code of location address.
14. Route	208	209	Numeric RJ	I2	Code indicating the road/route used if activity is outside of Webb County. See code definitions below.
15. Zone Number	210	214	Numeric RJ	I5	Zone number of location address. If unknown, it should be coded 8888. Locations in Mexico should be coded 7777 and addresses outside of Webb County but within Texas should be coded using the Statewide Zone System and preceded by the number 1 in column 210. Addresses outside of Texas and Mexico should be coded using 9999.
16. Longitude	215	224	Numeric RJ	F10.0	Longitude of location. If unknown, it should be coded 888.8888. Locations in Mexico should be coded 777.7777 and addresses outside of Webb County should be coded 999.9999.
17. Latitude	225	234	Numeric RJ	F10.0	Latitude of location. . If unknown, it should be coded 888.8888. Locations in Mexico should be coded 777.7777 and addresses outside of Webb County should be coded 999.9999.
18. Type of Place	235	236	Numeric RJ	I2	Code indicating the type of place at this location. If coded as “other”, specify in the next field. See code definitions below.
19. Other Place	237	256	Alphanum LJ	A20	Description of “other” type of place where activity occurred.
20. Purpose	257	258	Numeric RJ	I2	Purpose of trip, developed based on the activity type. See code definitions below.
21. Mode of Travel	259	260	Numeric RJ	I2	Code indicating mode of travel used in traveling to this location. See travel mode code definitions below.
22. Other Mode	261	290	Alphanum LJ	A30	If “other” is coded in mode of travel, this is the description of the “other” mode.
23. Number of People	291	292	Numeric RJ	I2	If travel was by private vehicle, this is the number of persons in the vehicle, including the person driving. A zero/blank should be recorded for non-private vehicle modes.
24. HH Members	293	294	Numeric RJ	I2	Of those in the vehicle, how many were household (HH) members.

TABLE A-7 (Continued)

Item	Begin	End	Type	Format	Description
25. Persons on Trip	295	304	Alphanum LJ	A10	Who was/were the HH members traveling with you?
26. Non HH Members	305	306	Numeric RJ	I2	Compute Non HH Members
27.HH Vehicle	307	307	Numeric RJ	I1	Was a HH vehicle used to make this trip? 1=Yes, 2=No, 9=Don't Know/Refused.
28. Vehicle Used	308	309	Numeric RJ	I2	If household vehicle was used for travel, this is the vehicle number (must correspond with vehicle number in household record). If other vehicle is used, this should be coded as 99.
29. Body Type	310	311	Numeric RJ	I2	See code set for body type.
30.Other Body Type	312	346	Alphanum LJ	A35	If body type is not in code set, describe body type.
31. Other Vehicle Year	347	350	Numeric RJ	I4	Year of "other" vehicle used for trip. 9998-Don't Know, 9999-Refused.
32. Other Vehicle Make	351	352	Numeric RJ	I2	Make of "other" vehicle used for trip. See code set.
33. Other Vehicle Make Description	353	412	Alphanum LJ	A60	If make of other vehicle is coded as other, this field contains a description of the vehicle make
34.Other Vehicle Model	413	472	Alphanum LJ	A60	Model of "other" vehicle used for trip.
35. Other Vehicle Fuel	473	474	Numeric RJ	I2	Code indicating type of fuel used by "other" vehicle; 1-Gasoline, 2-Diesel, 98-Other, 99-Don't Know / Refused.
36. Other Fuel	475	489	Alphanum LJ	A15	Description of "other" fuel for "other" vehicle, if not in fuel code above.
37. Other Vehicle Classification	490	491	Numeric RJ	I2	Code indicating EPA classification of other vehicle. See code definitions below.
38. Other Vehicle Commercial Use	492	493	Numeric RJ	I2	Code indicating if "other" vehicle used for commercial purposes; 1-Yes, 2-No, 99-Don't Know / Refused.
39. To Bus Stop	494	495	Numeric RJ	I2	Code indicating if they walked more than one block to get to bus stop; 1-Yes,2-No, 99-Don't Know / Refused.
40. To Activity	496	497	Numeric RJ	I2	Code indicating if they parked or got off bus more than one block from this activity; 1-Yes, 2-No, 99-Don't Know / Refused.
41. Off Bus Location	498	547	Alphanum LJ	A50	Street address or nearest intersecting streets where person got off of bus.
42. Parking Location	548	597	Alphanum LJ	A50	Street address of nearest intersecting streets where vehicle was parked.
43. Parking Cost	598	604	Numeric RJ	F7.2	Amount paid for parking.
44. Payment Method	605	606	Numeric RJ	I2	Time period for parking cost payment; 1-Hourly, 2-Daily, 3-Weekly, 4-Monthly, 5-Annually, 98-Other, 99-Don't Know / Refused.
45. Arrival Hour	607	608	Numeric RJ	I2	Hour that person arrived at this location. This hour should be in terms of military time. If this is activity 0, this should be blank since this is where they began their day.
46. Arrival Minute	609	610	Numeric RJ	I2	Minute that person arrived at this location. If this is activity 0, this should be blank since this is where they began their day.
47. Departure Hour	611	612	Numeric RJ	I2	Hour that person departed this location. This hour should be in terms of military time. If this is the last activity, this should be blank.
48. Departure Minute	613	614	Numeric RJ	I2	Minute that person departed this location. If this is the last activity for this person, this should be blank.

TABLE A-8 Example of Trip Data File Format Codes (Laredo) (*30*)

Route Codes (Trip Item 14)	**Type of Place Codes (Trip Item 18)**
1 – IH 35 (Juarez-Lincoln Bridge) at TX / Mexico Border	1 – Office Building
2 – IH 35 North at La Salle Co. Line	2 – Retail
3 – US 83 South at Zapata Co. Line	3 – Industrial / Manufacturing Site
4 – US 59 North at Duval Co. Line	4 – Medical
5 – US 83 North at Dimmit Co. Line	5 – Educational (12th grade or less)
6 – SH 44 East at Duval Co. Line	6 – Educational (College, trade, etc.)
7 – SH 359 East at Duval Co. Line	7 – Government
8 – FM 863 East at La Salle Co. Line	8 – Residential
9 – SH 44 West at La Salle Co. Line	9 – Other (Specify)
10 – FM 1472 (Laredo-Colombia Bridge) at TX / Mexico Border	10 – Airport
11 – FM 649 South at Jim Hogg Co. Line	99 – Don't Know / Refused
12 – Convent St. (Gateway to the Americas Bridge) at TX / Mexico Border	
13 – Loop 20 (World Trade Bridge) at TX/ Mexico Border	

Activity Type Codes (Trip Item 7)	
1 – At Home; primary job related	12– Other Services
2 – At Home; other	13– Social / Recreational
3 – At Home; job and non-job related	14– Eat Out
4 – Work	15– Civic Activities (including church)
5 – Work Related	16 – Pick-up / Drop-off Person at Work
6 – School; post secondary, college, trade	17 – Pick-up / Drop-off Person at School / Day Care
7 – School; secondary-day care, kindergarten, elementary, middle, high	18 – Pick-up / Drop-off Person at Other
8 – Incidental Shopping; gas, groceries, etc.	19 – Change Mode of Travel
9 – Major Shopping; clothes, appliances, etc.	20 – Other Activity (specify)
10 – Banking	99 – Don't Know / Refused
11– Personal Business; laundry, dry cleaning, barber, medical, etc	

Trip Purpose Codes (Trip Item 20)	**Mode of Travel Codes (Trip Item 21)**
1 – Home (Activity Type Codes 1,2,3)	1 – Walk
2 – Work (Activity Type Code 4)	2 – Auto / Van / Truck Driver
3 – Work Related (Activity Type Code 5)	3 – Auto / Van / Truck Passenger
4 – School; K thru 12 (Activity Type Code 7)	4 – Carpool Driver
5 – School; Post Secondary (Activity Type Code 6)	5 – Carpool Passenger
6 – Shopping (Activity Type Codes 8,9)	6 – Vanpool Driver
7 – Personal (Activity Type Codes 10,11,12,15)	7 – Vanpool Passenger
8 – Social / Recreation (Activity Type Codes 13,14)	8 – Commercial Vehicle Driver
9 – Pick-up Drop-off Other (Activity Type Codes 16,17,18)	9 – Commercial Vehicle Passenger
10 – Change Mode (Activity Type Code 19)	10 – Bus
11 – Other (Activity Type Code 20)	11 – School Bus
99 – Don't Know / Refused (Activity Type Code 99)	12 – Taxi / Paid Limo
	13 – Bicycle
	14 – Motorcycle / Moped
	15 – Other
	99 – Don't Know / Refused

Vehicle Classification Codes (Trip Item 29)	
1 – Light Duty Gas Vehicle	6 – Light Duty Diesel Truck
2 – Light Duty Gas Truck Type 1	7 – Heavy Duty Diesel Truck
3 – Light Duty Gas Truck Type 2	8 – Motorcycle
4 – Heavy Duty Gas Truck	9 – Alternative Fuel Vehicle
5 – Light Duty Diesel Vehicle	99 – Don't Know / Refused

TABLE A-9 Example of GPS Data File Format (Laredo) (*30*)

Variable Name	Variable Description	Data Type	Just.	Field Width	Values	Formal and Full Text
RECTYPE	Record Type	Integer	Right	1	GPS Record Type = 5	
GPS_ID	GPS Receiver Unit ID Number	Integer	Right	3	0-999	
HH_ID	Household ID Number	Integer	Right	7		
Veh_ID	Vehicle Number	Integer	Right	2		
GMT_DATE	Greenwich Mean Time Date Stamp	Integer	Right	8	MM/DD/YY	
GMT_TIME	Greenwich Mean Time Time Stamp	Integer	Right	8	HH:MM:SS (Military Time)	
LOC_DATE	Local Date Stamp	Integer	Right	8	MM/DD/YY	
LOC_TIME	Local Time Stamp	Integer	Right	8	HH:MM:SS (Military Time)	
LAT_RAW	Latitude	Float	Right	16	Decimal Degrees	XXX.XXXXXX deg
LONG_RAW	Longitude	Float	Right	16	Decimal Degrees	XXX.XXXXXX deg
ELEV_RAW	Elevation	Float	Right	16	Meters	
VELOCITY	Velocity	Float	Right	8	Meters/Second	0..514.00m/s
HEADING	Direction of Vehicle	Float	Right	6	True North	0.0..359.9 deg
HDOP	HDOP (horizontal dilution of precision)	Integer	Right	4	00.5-99.9	
SATS	Number of Satellites	Integer	Right	2	00-12	

TABLE A-10 Trip Comparison Table Format (Laredo)

Variable Name	Variable Description	Data Type	Values
Date	The assigned travel day for the household	Date	MM/DD/YY
Household ID	Used to give each household a distinct ID	Integer	1…X
Vehicle ID	Used to give each vehicle a distinct ID within each household	Integer	1…X
Trip Number	The *n*th trip for the assigned travel day	Integer	1…X
Trip Reported	Whether or not the GPS trip was reported in the CATI	Float	0 (no) or 1 (yes)
GPS Begin Time	Time of day the trip began	Time	HH:MM:SS
GPS End Time	Time of day the trip ended	Time	HH:MM:SS
GPS Total Time	The total time for the trip	Time	HH:MM:SS
GPS Activity Time	The length of time spent at the activity following the trip	Time	HH:MM:SS
GPS Distance	The measured distance of the trip (miles)	Float	0.01…X
GPS Speed	The measured average speed of the trip	Float	0.01…X
GPS Trip Purpose	The purpose of the trip	Text	HBW, HBNW, or NHB
CATI Begin Time	Time of day the trip began	Time	HH:MM:SS
CATI End Time	Time of day the trip ended	Time	HH:MM:SS
CATI Total Time	The total time for the trip	Time	HH:MM:SS
CATI Activity Time	The length of time spent at the activity following the trip	Time	HH:MM:SS
CATI Distance	The measured distance of the trip (miles)	Float	0.01…X
CATI Speed	The measured average speed of the trip	Float	0.01…X
CATI Trip Purpose	The purpose of the trip	Text	HBW, HBNW, or NHB

APPENDIX B

SAS Program for Laredo

```
options nodate notime nocenter ps=80 ls=78;
data tripreporting;
input hhvehid hhtrips purpose $ begin $ time dist dayweek $
      numper numemp income vehavail restype $ delveh reported;
cards;

<DATA>;

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      selection=stepwise
      slentry=0.001
      slstay=0.001
      details
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Laredo - Stepwise Selection Analysis of Trip Reporting Data';

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Laredo - Full Analysis of Trip Reporting Data';

run;
```

SAS Program for Tyler-Longview

```
options nodate notime nocenter ps=80 ls=78;
data tripreporting;
input hhvehid hhtrips purpose $ begin $ time dist dayweek $
      numper numemp income vehavail restype $ delveh reported;
cards;

<DATA>;

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      selection=stepwise
      slentry=0.001
      slstay=0.001
      details
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Tyler-Longview - Stepwise Selection Analysis of Trip Reporting
Data';

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Tyler-Longview - Full Analysis of Trip Reporting Data';

run;
```

SAS Program for Combined Study Areas

```
options nodate notime nocenter ps=80 ls=78;
data tripreporting;
input hhvehid hhtrips purpose $ begin $ time dist dayweek $
      numper numemp income vehavail restype $ delveh reported;
cards;

<DATA>;

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      selection=stepwise
      slentry=0.001
      slstay=0.001
      details
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Combined - Stepwise Selection Analysis of Trip Reporting Data';

proc logistic desc;
      class purpose(ref='HBW') begin(ref=last) dayweek(ref='F')
      restype(ref='Multi')/param=ref order=data;
      model reported=hhtrips purpose begin time
      dist dayweek numper numemp income vehavail restype delveh/
      lackfit
      rsquare
      scale=none
      aggregate;
title 'Combined - Full Analysis of Trip Reporting Data';

run;
```

APPENDIX C

SAS Output for Laredo – Stepwise Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	917
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	422
2	0	495

Probability modeled is reported=1.

Analysis of Effects in Model

Effect	DF	Wald Chi-Square	Pr > ChiSq
purpose	2	45.7966	<.0001
dayweek	4	33.8268	<.0001

Analysis of Effects Not in the Model

Effect	DF	Score Chi-Square	Pr > ChiSq
hhtrips	1	9.8577	0.0017
begin	4	16.3480	0.0026
time	1	0.6880	0.4069
dist	1	1.0619	0.3028
numper	1	0.0009	0.9755
numemp	1	1.7405	0.1871
income	1	0.6523	0.4193
vehavail	1	1.1104	0.2920
restype	1	5.5925	0.0180
delveh	1	0.0737	0.7860

R-Square 0.1032 **Max-rescaled R-Square** 0.1379

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	1161.7153	905	1.2837	<.0001
Pearson	905.1626	905	1.0002	0.4922

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
13.4856	9	0.1418

Association of Predicted Probabilities and Observed Responses

Percent Concordant	61.5	Somers' D	0.318
Percent Discordant	29.7	Gamma	0.349
Percent Tied	8.8	Tau-a	0.158
Pairs	208890	c	0.659

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.7897	0.4668	35.7149	<.0001
purpose	NHB	1	-2.9224	0.4476	42.6204	<.0001
purpose	HBNW	1	-2.5240	0.4482	31.7151	<.0001
dayweek	M	1	-0.1984	0.2125	0.8713	0.3506
dayweek	W	1	-0.8671	0.2600	11.1218	0.0009
dayweek	T	1	0.0213	0.2133	0.0099	0.9206
dayweek	R	1	-0.9951	0.2333	18.1912	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
purpose NHB vs HBW	0.054	0.022	0.129
purpose HBNW vs HBW	0.080	0.033	0.193
dayweek M vs F	0.820	0.541	1.244
dayweek W vs F	0.420	0.252	0.699
dayweek T vs F	1.021	0.672	1.552
dayweek R vs F	0.370	0.234	0.584

SAS Output for Laredo – Full Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	917
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	422
2	0	495

Probability modeled is reported=1.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhtrips	1	10.4752	0.0012
purpose	2	35.9445	<.0001
begin	4	17.2346	0.0017
time	1	19.6243	<.0001
dist	1	13.0680	0.0003
dayweek	4	36.9445	<.0001
numper	1	0.0294	0.8638
numemp	1	2.8818	0.0896
income	1	1.0822	0.2982
vehavail	1	3.9834	0.0460
restype	1	4.3287	0.0375
delveh	1	2.7363	0.0981

R-Square 0.1649 **Max-rescaled R-Square** 0.2203

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	1096.3576	892	1.2291	<.0001
Pearson	906.8374	892	1.0166	0.3575

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
4.6546	8	0.7938

Association of Predicted Probabilities and Observed Responses

Percent Concordant	72.8	Somers' D	0.458
Percent Discordant	27.0	Gamma	0.460
Percent Tied	0.3	Tau-a	0.228
Pairs	208890	c	0.729

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.2785	0.6130	4.3498	0.0370
hhtrips		1	-0.0425	0.0131	10.4752	0.0012
purpose	NHB	1	-2.7475	0.4668	34.6411	<.0001
purpose	HBNW	1	-2.4076	0.4638	26.9467	<.0001
begin	4a-8a	1	0.8913	0.3135	8.0848	0.0045
begin	8a-12p	1	1.0207	0.2872	12.6275	0.0004
begin	12p-4p	1	0.4350	0.2671	2.6530	0.1034
begin	4p-8p	1	0.7539	0.2681	7.9077	0.0049
time		1	0.1073	0.0242	19.6243	<.0001
dist		1	-0.1404	0.0388	13.0680	0.0003
dayweek	M	1	-0.1387	0.2320	0.3577	0.5498
dayweek	W	1	-1.0138	0.2774	13.3538	0.0003
dayweek	T	1	0.00861	0.2376	0.0013	0.9711
dayweek	R	1	-1.1253	0.2490	20.4289	<.0001
numper		1	0.0105	0.0610	0.0294	0.8638
numemp		1	-0.1969	0.1160	2.8818	0.0896
income		1	3.193E-6	3.069E-6	1.0822	0.2982
vehavail		1	0.2094	0.1049	3.9834	0.0460
restype	Single	1	0.6246	0.3002	4.3287	0.0375
delveh		1	0.3042	0.1839	2.7363	0.0981

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
hhtrips		0.958	0.934	0.983
purpose	NHB vs HBW	0.064	0.026	0.160
purpose	HBNW vs HBW	0.090	0.036	0.223
begin	4a-8a vs 8p-12a	2.438	1.319	4.508
begin	8a-12p vs 8p-12a	2.775	1.580	4.872
begin	12p-4p vs 8p-12a	1.545	0.915	2.608
begin	4p-8p vs 8p-12a	2.125	1.257	3.594
time		1.113	1.062	1.167
dist		0.869	0.805	0.938
dayweek	M vs F	0.870	0.552	1.372
dayweek	W vs F	0.363	0.211	0.625
dayweek	T vs F	1.009	0.633	1.607
dayweek	R vs F	0.325	0.199	0.529
numper		1.011	0.897	1.139
numemp		0.821	0.654	1.031
income		1.000	1.000	1.000
vehavail		1.233	1.004	1.514
restype	Single vs Multi	1.867	1.037	3.363
delveh		1.356	0.945	1.944

SAS Output for Tyler-Longview – Stepwise Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	1878
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	1289
2	0	589

Probability modeled is reported=1.

Analysis of Effects in Model

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhtrips	1	100.4091	<.0001
purpose	2	70.0652	<.0001
time	1	12.3329	0.0004
numper	1	10.8919	0.0010

Analysis of Effects Not in the Model

Effect	DF	Score Chi-Square	Pr > ChiSq
begin	5	9.2073	0.1011
dist	1	3.0583	0.0803
dayweek	4	9.2929	0.0542
numemp	1	6.4886	0.0109
income	1	7.1012	0.0077
vehavail	1	2.1933	0.1386
restype	1	0.1764	0.6745
delveh	1	2.4860	0.1149

R-Square 0.1291 **Max-rescaled R-Square** 0.1814

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	2076.5382	1872	1.1093	0.0006
Pearson	1891.1464	1872	1.0102	0.3734

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
14.1433	8	0.0781

Association of Predicted Probabilities and Observed Responses

Percent Concordant	71.9	Somers' D	0.442
Percent Discordant	27.7	Gamma	0.444
Percent Tied	0.3	Tau-a	0.190
Pairs	759221	c	0.721

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.0168	0.2594	60.4423	<.0001
hhtrips		1	-0.0641	0.00640	100.4091	<.0001
purpose	HBNW	1	-0.5137	0.2290	5.0317	0.0249
purpose	NHB	1	-1.3130	0.2221	34.9411	<.0001
time		1	0.0274	0.00781	12.3329	0.0004
numper		1	0.1661	0.0503	10.8919	0.0010

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
hhtrips	0.938	0.926	0.950
purpose HBNW vs HBW	0.598	0.382	0.937
purpose NHB vs HBW	0.269	0.174	0.416
time	1.028	1.012	1.044
numper	1.181	1.070	1.303

SAS Output for Tyler-Longview – Full Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	1878
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	1289
2	0	589

Probability modeled is reported=1.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhtrips	1	98.3472	<.0001
purpose	2	68.2187	<.0001
begin	5	7.6636	0.1758
time	1	7.2999	0.0069
dist	1	3.1256	0.0771
dayweek	4	8.6809	0.0696
numper	1	20.0804	<.0001
numemp	1	14.4717	0.0001
income	1	11.6334	0.0006
vehavail	1	5.1774	0.0229
restype	1	2.4569	0.1170
delveh	1	1.5525	0.2128

R-Square 0.1509 **Max-rescaled R-Square** 0.2120

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	2028.9880	1857	1.0926	0.0030
Pearson	1931.0966	1857	1.0399	0.1130

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
15.9576	8	0.0430

Association of Predicted Probabilities and Observed Responses

Percent Concordant	73.9	Somers' D	0.480
Percent Discordant	25.8	Gamma	0.482
Percent Tied	0.3	Tau-a	0.207
Pairs	759221	c	0.740

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.9613	0.4171	22.1080	<.0001
hhtrips		1	-0.0760	0.00767	98.3472	<.0001
purpose	HBNW	1	-0.4999	0.2370	4.4473	0.0350
purpose	NHB	1	-1.3192	0.2297	32.9854	<.0001
begin	12a-4a	1	9.3657	493.1	0.0004	0.9848
begin	4a-8a	1	0.6731	0.2785	5.8410	0.0157
begin	8a-12p	1	0.2093	0.2452	0.7287	0.3933
begin	12p-4p	1	0.2127	0.2390	0.7919	0.3735
begin	4p-8p	1	0.3051	0.2406	1.6087	0.2047
time		1	0.0570	0.0211	7.2999	0.0069
dist		1	-0.0446	0.0252	3.1256	0.0771
dayweek	T	1	-0.2146	0.1784	1.4469	0.2290
dayweek	W	1	-0.3251	0.1832	3.1481	0.0760
dayweek	R	1	-0.4550	0.1616	7.9293	0.0049
dayweek	M	1	-0.3252	0.1864	3.0437	0.0811
numper		1	0.2634	0.0588	20.0804	<.0001
numemp		1	-0.2980	0.0783	14.4717	0.0001
income		1	6.829E-6	2.002E-6	11.6334	0.0006
vehavail		1	0.1729	0.0760	5.1774	0.0229
restype	Single	1	-0.4171	0.2661	2.4569	0.1170
delveh		1	0.2038	0.1635	1.5525	0.2128

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
hhtrips		0.927	0.913	0.941
purpose	HBNW vs HBW	0.607	0.381	0.965
purpose	NHB vs HBW	0.267	0.170	0.419
begin	12a-4a vs 8p-12a	>999.999	<0.001	>999.999
begin	4a-8a vs 8p-12a	1.960	1.136	3.383
begin	8a-12p vs 8p-12a	1.233	0.762	1.994
begin	12p-4p vs 8p-12a	1.237	0.774	1.976
begin	4p-8p vs 8p-12a	1.357	0.847	2.174
time		1.059	1.016	1.103
dist		0.956	0.910	1.005
dayweek	T vs F	0.807	0.569	1.145
dayweek	W vs F	0.722	0.505	1.035
dayweek	R vs F	0.634	0.462	0.871
dayweek	M vs F	0.722	0.501	1.041
numper		1.301	1.160	1.460
numemp		0.742	0.637	0.865
income		1.000	1.000	1.000
vehavail		1.189	1.024	1.380
restype	Single vs Multi	0.659	0.391	1.110
delveh		1.226	0.890	1.689

SAS Output for Combined Study Areas – Stepwise Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	2795
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	1711
2	0	1084

Probability modeled is reported=1.

Analysis of Effects in Model

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhtrips	1	74.9353	<.0001
purpose	2	104.8930	<.0001
begin	5	24.5115	0.0002
dayweek	4	25.8491	<.0001
numemp	1	25.4067	<.0001
vehavail	1	13.8755	0.0002

Analysis of Effects Not in the Model

Effect	DF	Score Chi-Square	Pr > ChiSq
time	1	8.1098	0.0044
dist	1	1.9122	0.1667
numper	1	0.0146	0.9037
income	1	7.9735	0.0047
restype	1	3.6686	0.0554
delveh	1	1.1859	0.2762

R-Square 0.1150 **Max-rescaled R-Square** 0.1560

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	3387.6834	2775	1.2208	<.0001
Pearson	2801.8475	2775	1.0097	0.3564

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
12.2749	8	0.1394

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.6	Somers' D	0.396
Percent Discordant	30.0	Gamma	0.397
Percent Tied	0.4	Tau-a	0.188
Pairs	1854724	c	0.698

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	2.2888	0.2763	68.6032	<.0001
hhtrips		1	-0.0469	0.00541	74.9353	<.0001
purpose	NHB	1	-1.7776	0.1993	79.5689	<.0001
purpose	HBNW	1	-1.1910	0.2011	35.0709	<.0001
begin	12a-4a	1	9.2727	280.9	0.0011	0.9737
begin	4a-8a	1	0.8630	0.1900	20.6282	<.0001
begin	8a-12p	1	0.6505	0.1708	14.5046	0.0001
begin	12p-4p	1	0.4614	0.1636	7.9514	0.0048
begin	4p-8p	1	0.6214	0.1654	14.1125	0.0002
dayweek	T	1	-0.1269	0.1304	0.9478	0.3303
dayweek	W	1	-0.5374	0.1386	15.0355	0.0001
dayweek	R	1	-0.5193	0.1266	16.8318	<.0001
dayweek	M	1	-0.2286	0.1316	3.0172	0.0824
numemp		1	-0.2618	0.0519	25.4067	<.0001
vehavail		1	0.1998	0.0536	13.8755	0.0002

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
hhtrips		0.954	0.944	0.964
purpose	NHB vs HBW	0.169	0.114	0.250
purpose	HBNW vs HBW	0.304	0.205	0.451
begin	12a-4a vs 8p-12a	>999.999	<0.001	>999.999
begin	4a-8a vs 8p-12a	2.370	1.633	3.440
begin	8a-12p vs 8p-12a	1.917	1.371	2.679
begin	12p-4p vs 8p-12a	1.586	1.151	2.186
begin	4p-8p vs 8p-12a	1.862	1.346	2.574
dayweek	T vs F	0.881	0.682	1.137
dayweek	W vs F	0.584	0.445	0.767
dayweek	R vs F	0.595	0.464	0.762
dayweek	M vs F	0.796	0.615	1.030
numemp		0.770	0.695	0.852
vehavail		1.221	1.099	1.357

SAS Output for Combined Study Areas – Full Model

Model Information

Data Set	WORK.TRIPREPORTING
Response Variable	reported
Number of Response Levels	2
Number of Observations	2795
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	reported	Total Frequency
1	1	1711
2	0	1084

Probability modeled is reported=1.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
hhtrips	1	73.5395	<.0001
purpose	2	105.1985	<.0001
begin	5	21.6579	0.0006
time	1	16.7501	<.0001
dist	1	11.4362	0.0007
dayweek	4	22.0323	0.0002
numper	1	0.0001	0.9917
numemp	1	26.0137	<.0001
income	1	5.7478	0.0165
vehavail	1	11.5297	0.0007
restype	1	2.0359	0.1536
delveh	1	1.1082	0.2925

R-Square 0.1245 **Max-rescaled R-Square** 0.1690

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	3357.3230	2769	1.2125	<.0001
Pearson	2803.2183	2769	1.0124	0.3203

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.6581	8	0.5739

Association of Predicted Probabilities and Observed Responses

Percent Concordant	70.6	Somers' D	0.414
Percent Discordant	29.2	Gamma	0.415
Percent Tied	0.3	Tau-a	0.197
Pairs	1854724	c	0.707

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.7634	0.3069	33.0054	<.0001
hhtrips		1	-0.0483	0.00564	73.5395	<.0001
purpose	NHB	1	-1.7861	0.2011	78.9220	<.0001
purpose	HBNW	1	-1.1890	0.2031	34.2770	<.0001
begin	12a-4a	1	8.9385	280.9	0.0010	0.9746
begin	4a-8a	1	0.8137	0.1921	17.9388	<.0001
begin	8a-12p	1	0.6064	0.1729	12.2997	0.0005
begin	12p-4p	1	0.4068	0.1651	6.0717	0.0137
begin	4p-8p	1	0.5608	0.1667	11.3113	0.0008
time		1	0.0557	0.0136	16.7501	<.0001
dist		1	-0.0507	0.0150	11.4362	0.0007
dayweek	T	1	-0.1517	0.1319	1.3239	0.2499
dayweek	W	1	-0.5017	0.1423	12.4291	0.0004
dayweek	R	1	-0.5105	0.1276	16.0170	<.0001
dayweek	M	1	-0.2517	0.1333	3.5674	0.0589
numper		1	0.000388	0.0375	0.0001	0.9917
numemp		1	-0.3083	0.0605	26.0137	<.0001
income		1	3.769E-6	1.572E-6	5.7478	0.0165
vehavail		1	0.1867	0.0550	11.5297	0.0007
restype	Single	1	0.2518	0.1765	2.0359	0.1536
delveh		1	0.1141	0.1084	1.1082	0.2925

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
hhtrips		0.953	0.942	0.963
purpose	NHB vs HBW	0.168	0.113	0.249
purpose	HBNW vs HBW	0.305	0.205	0.453
begin	12a-4a vs 8p-12a	>999.999	<0.001	>999.999
begin	4a-8a vs 8p-12a	2.256	1.548	3.288
begin	8a-12p vs 8p-12a	1.834	1.307	2.574
begin	12p-4p vs 8p-12a	1.502	1.087	2.076
begin	4p-8p vs 8p-12a	1.752	1.264	2.429
time		1.057	1.029	1.086
dist		0.951	0.923	0.979
dayweek	T vs F	0.859	0.664	1.113
dayweek	W vs F	0.606	0.458	0.800
dayweek	R vs F	0.600	0.467	0.771
dayweek	M vs F	0.777	0.599	1.010
numper		1.000	0.929	1.077
numemp		0.735	0.653	0.827
income		1.000	1.000	1.000
vehavail		1.205	1.082	1.342
restype	Single vs Multi	1.286	0.910	1.818
delveh		1.121	0.906	1.386

VITA

Timothy Lee Forrest was born in Peoria, Illinois, on April 13, 1978. He currently resides at 5809 Canterbury Dr., Bryan, TX, 77802. He received his undergraduate degree (B.S.) in rangeland ecology and management from Texas A&M University in 2001. He is currently a Research Associate in the Transportation Planning program of the Texas Transportation Institute, located in on the Texas A&M University main campus in College Station, TX.

His professional and research interests cover a wide variety of topics, ranging from the analysis of new methods of travel surveying, to the application of portable emissions measurement systems for monitoring light- and heavy-duty diesel and gasoline vehicles. In addition to his first publication being currently in press, his work has also been presented at multiple conferences within the transportation planning field.