PRELIMINARY REVIEW COPY

Technical Report Documentation Page

1. Report No. FHWA/TX-06/0-5176-3	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Conversion of Volunteer-Collected GPS Diary Data into Travel Time Performance Measures: Final Report		5. Report Date February 2006 6. Performing Organization Code
7. Author(s) Sivaramakrishnan Srinivasan, Prabuddha Ghosh, Aruna Sivakumar, Aarti Kapur, Chandra R. Bhat, and Stacey Bricka		8. Performing Organization Report No. 0-5176-3
9. Performing Organization Name and Address Center for Transportation Research The University of Texas at Austin 3208 Red River, Suite 200 Austin, TX 78705-2650		10. Work Unit No. (TRAIS) 11. Contract or Grant No. 0-5176
12. Sponsoring Agency Name and Address Texas Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, TX 78763-5080		13. Type of Report and Period Covered Technical Report (9/1/04-12/31/05 14. Sponsoring Agency Code
15. Supplementary Notes Project performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.		

16. Abstract

Conventional travel-survey methodologies require the collection of detailed activity-travel information, which imposes a significant burden on respondents, thereby adversely impacting the quality and quantity of data obtained. Advances in the global positioning system (GPS) technology have provided transportation planners with an alternative and powerful tool for more accurate travel-data collection with minimal user burden. The data recorded by GPS devices, however, do not directly yield travel information; the navigational streams recorded by GPS devices have to be processed and the travel patterns derived from them. This research project developed prototype software to automate the processing of raw GPS data and to generate outputs of activity-travel patterns in the conventional travel-diary format. The software identifies trips and characterizes them by several attributes, including trip-end locations, trip purpose, time of day, distance, and speed. This final report documents the entire research performed as part of this project. Specifically, we present the conceptual overview of the software, the detailed algorithm for extracting travel diaries, the software implementation procedures, and the testing and validation of the software.

17. Key Words Household travel surveys, global positioning system (GPS), GPS-based travel surveys, GPS data recording formats, processing GPS navigational data	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161; www.ntis.gov.

19. Security Classif. (of report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of pages 66	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

CONVERSION OF VOLUNTEER-COLLECTED GPS DIARY DATA INTO TRAVEL TIME PERFORMANCE MEASURES: FINAL REPORT

Sivaramakrishnan Srinivasan
Prabuddha Ghosh
Aruna Sivakumar
Aarti Kapur
Chandra R. Bhat
Stacey Bricka

CTR Technical Report:	0-5176-3
Report Date:	February 2006
Project:	0-5176
Project Title:	Conversion of Volunteer-Collected GPS Diary Data into Travel Time Performance Measures
Sponsoring Agency:	Texas Department of Transportation
Performing Agency:	Center for Transportation Research at The University of Texas at Austin

Project performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.

Center for Transportation Research
The University of Texas at Austin
3208 Red River
Austin, TX 78705

www.utexas.edu/research/ctr

Copyright (c) 2006
Center for Transportation Research
The University of Texas at Austin

All rights reserved
Printed in the United States of America

Disclaimers

Author's Disclaimer: The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the Federal Highway Administration or the Texas Department of Transportation (TxDOT). This report does not constitute a standard, specification, or regulation.

Patent Disclaimer: There was no invention or discovery conceived or first actually reduced to practice in the course of or under this contract, including any art, method, process, machine manufacture, design or composition of matter, or any new useful improvement thereof, or any variety of plant, which is or may be patentable under the patent laws of the United States of America or any foreign country.

Engineering Disclaimer

NOT INTENDED FOR CONSTRUCTION, BIDDING, OR PERMIT PURPOSES.

Project Engineer: Chandra R. Bhat
Professional Engineer License State and Number: Texas No. 88971
P. E. Designation: Research Supervisor

Acknowledgments

The authors express appreciation to Michael Chamberlain and the rest of the project monitoring committee for their valuable input throughout the course of the project.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

For nearly fifty years, household travel surveys have been used to document the travel behavior of regional households as part of long-range transportation planning efforts. The survey data are used for general planning and policy analysis, as well as to serve as the foundation for regional travel demand models. Technology advancements have resulted in changes in household travel survey data collection procedures, the most recent being the introduction of Global Positioning Systems (GPS) to record travel patterns. The GPS technology shows promise to minimize costs, while maximizing the volume of travel data collected. However, the data recorded by GPS devices do not directly yield travel information; rather, the outputs from these devices are in the form of navigational streams that have to be processed to derive travel information. Therefore the success of this new technology as a travel survey instrument depends on the ability of the analyst to derive meaningful trip information from the navigational data streams of GPS devices.

Research project 0-5176, funded by the Texas Department of Transportation (TxDOT), focused on the development of a prototype software tool labeled the "GPS-Based Travel Diary Generator" (GPS-TDG) that automates the process of converting navigational data streams collected passively from in-vehicle GPS devices into an electronic travel diary. This derived travel diary comprises a sequence of vehicle trips identified from the GPS streams, with each trip characterized in terms of attributes such as trip-end location, trip purpose (or activity type at destination), time of day, duration, distance, and speed.

The software has been developed in the Java programming language using ArcGIS 9.0 as the platform for GIS processing. The software has been designed to operate either in a basic analysis mode or in an enhanced analysis mode. The basic mode converts the GPS data into a simple trip file that distinguishes among home-based work, home-based other, and non–home-based trips. The enhanced mode utilizes additional land-use data and pre-estimated model parameters to determine disaggregate trip purposes such as shopping and recreation.

The algorithm implemented within the GPS-TDG software is controlled by several parameters (such as the dwell-time thresholds), which can be easily modified by the analyst. Thus the software can be calibrated for any specific study region. We also provide default values

for all these parameters based on the limited testing and validation undertaken with available data.

Finally, the software is also capable of aggregating the derived trip diaries to produce inter-zonal vehicle trip tables and network performance measures (average trip speed, distance, and travel time). These measures can be generated in the overall or for specific trip purposes or for specific times-of-the-day.

This final report summarizes the research undertaken in Project 0-5176. The rest of this document is organized as follows. Chapter 2 describes the conceptual structure of the GPS-TDG system. Chapter 3 details the algorithm developed to derive travel diaries from GPS streams. Chapter 4 focuses on software implementation details. Chapter 5 presents exercises undertaken to test and validate the software. Finally, Chapter 6 summarizes the document and identifies important areas for further research and development.

CHAPTER 2. GPS-TDG: CONCEPTUAL OVERVIEW

A conceptual overview of the GPS-TDG is presented in Figure 2.1. The software is designed to operate either in a basic analysis mode or in an enhanced analysis mode. The basic mode is designed to convert the GPS data into a simple trip file that distinguishes among home-based work, home-based other, and non–home-based trips. The enhanced mode utilizes additional data to determine disaggregate trip purposes. As a result, the basic analysis mode is less demanding in terms of input data compared to the enhanced analysis mode.

Four input data components are common to both the basic and enhanced analysis modes, as indicated in Figure 2.1. In addition, the enhanced analysis mode requires three additional inputs. Section 2.1 of this report discusses all software inputs. Among these inputs, the GPS streams require preprocessing to be converted into a format that can be readily used by the software. This preprocessing step is described in Section 2.2. Section 2.3 focuses on the main processing module of the GPS-TDG software. This is the core of the software that implements the travel-diary extraction algorithm and creates the trip file. (The algorithm is discussed in Chapter 3 and the corresponding implementation details are presented in Chapter 4.) The primary output of the GPS-TDG software is the derived travel diary, which is discussed in Section 2.4. These travel-diary data can then be aggregated, using the querying module, to compute trip tables and inter-zonal highway performance measures (Section 2.5).

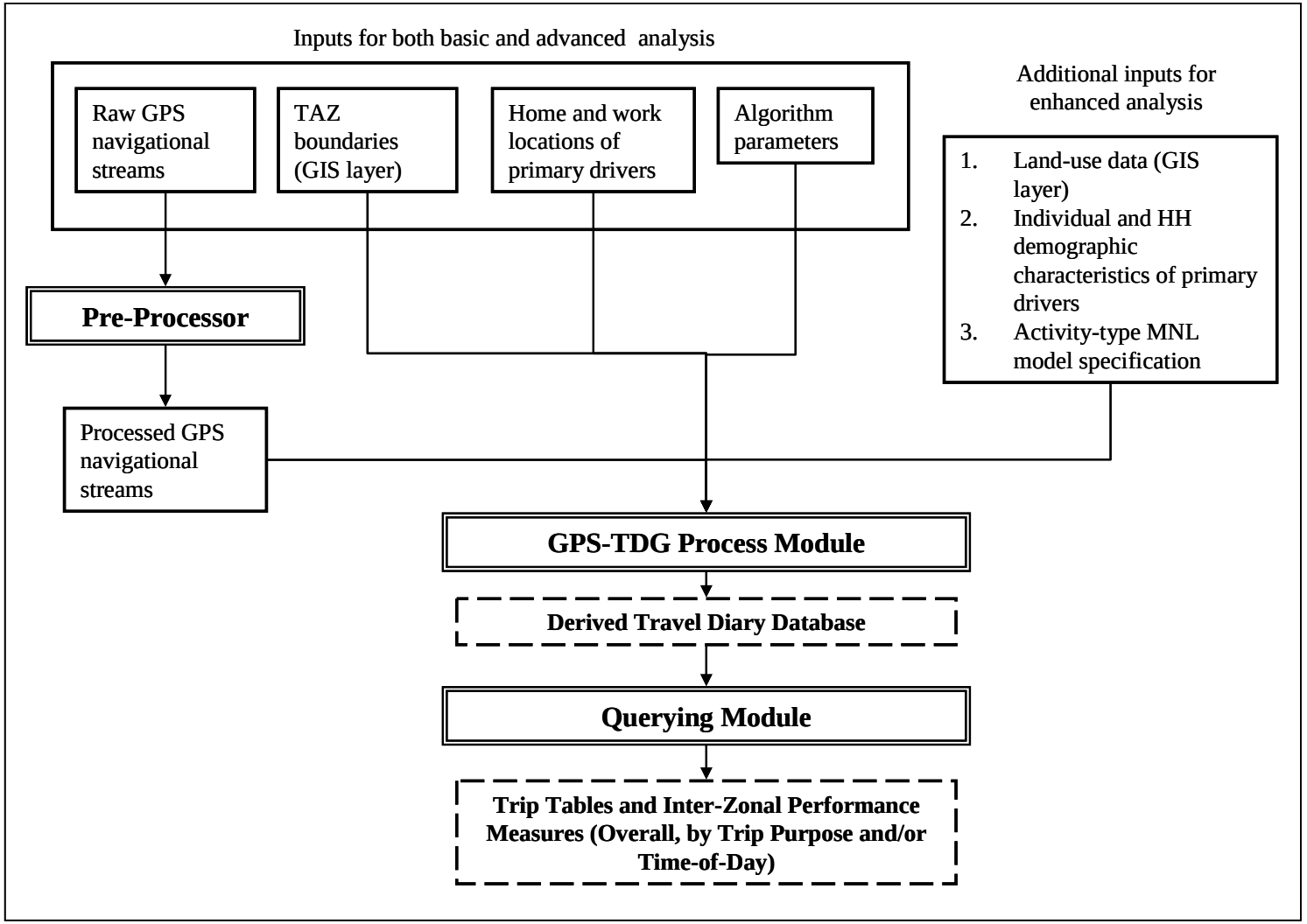

Figure 2.1 Conceptual overview of the GPS-TDG

2.1 Inputs

As discussed, four main inputs are required for both the basic and enhanced analysis modes:

1. The raw GPS navigational streams.
2. The boundaries of the traffic analysis zones (TAZs) as a geographic information system (GIS) layer.
3. The home and work (if employed) locations (as latitudes and longitudes) of the primary driver for each household vehicle. In the development of the GPS-TDG algorithms, it is assumed that each GPS-equipped vehicle is used predominantly by one person and this person is designated as the primary driver of the corresponding vehicle.
4. Parameters (such as dwell-time threshold) to control the travel-diary extraction algorithm. These parameters enable the analysts to better customize the software to different study regions and for different study objectives.

In order to run the software in the enhanced analysis mode, three additional inputs are required. These are:

1. Land-use or zoning data. These data are provided as a GIS layer in which the study area is divided into a number of land-use parcels and each parcel is assigned a specific land-use type.
2. Individual and household demographic characteristics (such as gender, employment/student status, and number of children in the household) of the primary driver associated with each GPS-equipped vehicle. These data are used by an MNL model for disaggregate activity-type determination.
3. Specifications for the multinomial logit (MNL) model for a disaggregate activity-type classification. This includes the activity-type classification, the list of explanatory variables, and the values of the model coefficients on these explanatory variables.

2.2 Preprocessor

The GPS units record vehicle movement at 1–5 second intervals. Each of these records contains the vehicle's position in terms of latitude and longitude, as well as velocity and other

position indicators. In addition, each record also contains a flag to indicate whether sufficient detail is present to consider the record "valid" for location identification purposes. In the preprocessing step the invalid records are removed and the number of invalid records removed immediately prior to each valid record is recorded. The valid records are then restructured into a format that is readily usable for trip-diary extraction processing. The preprocessing algorithm is discussed in Chapter 3.

2.3 Process Module

The process module forms the core of the GPS-TDG software and implements the travel-diary extraction algorithm. As discussed in detail in Research Report 1, operational characteristics (user-flagged versus purely passive systems), data-collection protocols (switched power versus continuous power systems), and data logging rules (such as speed-checked data recordings) define the structure of the GPS data streams recorded. This in turn impacts the processing methods required for travel-diary extraction. Our software has been developed primarily for purely passive systems (i.e., the driver does not manually "flag" the end of trips) that are switched-power (i.e., the recording of GPS data stops when the engine is powered off) and have no data logging rules (i.e., data points are recorded continuously as long as the engine is powered-on, regardless of the travel speed or other characteristics)[1].

It is also useful to note here that the objective of this software is to completely automate the travel-diary extraction procedure. However, identification of the network links actually traversed by the vehicle and hence the determination of travel route is not within the scope of this software.

2.4 Derived Travel Diary

The fundamental output of the GPS-TDG software is the derived travel diary. This output contains the trips extracted from the GPS streams for each of the equipped vehicles. Each trip is characterized by the following attributes: trip timing, duration of activity at trip destination, trip-end locations (latitude, longitude and TAZ), trip distance, average trip speed, measure of

[1] The software developed is also capable of analyzing GPS streams from continuous powered systems since it includes a routine to detect non–engine power-off stops. However, GPS streams from continuous powered systems typically tend to be very large files and the performance of this software has not been tested for these cases.

variation in speed during the trip, and activity purpose at trip-end location (aggregate classification scheme in the case of basic analysis and a disaggregate classification scheme in the case of enhanced analysis). In addition, accuracy measures are also generated to capture the impacts of signal loss or equipment malfunction on the identification and characterization of trips.

2.5 Querying Module

The querying module takes as input the derived travel diary and generates trip tables, inter-zonal performance measures (average trip lengths, distance, and speeds), and other summary measures (such as average trip lengths, distance, and speeds by trip-end activity purpose). Further, the module is designed to support performing these analyses in the overall and by trip-purpose and time-of-day.

CHAPTER 3. TRAVEL-DIARY EXTRACTION ALGORITHM

This chapter presents a detailed description of the methodology developed to derive the travel-diary data from preprocessed GPS navigational streams and other inputs. Figure 3.1 presents a conceptual overview of the travel-diary extraction algorithm. The first step of this algorithm accepts all required inputs. These inputs include the preprocessed GPS navigational streams, household and individual demographic data, a GIS layer of the TAZ boundaries, and other optional GIS land-use data. In addition, the software allows for the analyst to specify several parameters used for trip detection and characterization. A complete description of all inputs is provided in Section 3.1.

The algorithm is designed to extract the trip-diary information for one vehicle at a time. Correspondingly, all data for a single vehicle are read in Step 2 and Steps 3, 4, and 5 are performed on this data. Step 3, described in detail in Section 3.2, is trip detection, which involves scanning through the GPS stream for the vehicle under consideration until a potential trip end (or stop) is detected. In Step 4, the trip characteristics such as start and end times, distance, and average speed are computed for the potential trip detected in Step 3. A complete list of all trip attributes is computed and the corresponding methods are presented in Section 3.3. In Step 5, the reasonableness of the trip attributes determined in the previous step is examined by comparing it to user-provided thresholds. A potential trip that fails these checks (described in Section 3.4) is classified as a “false” trip. On the other hand, if the potential trip end is found to be reasonable, the trip characteristics are written out to the output file. If the end of the GPS stream has not been reached, the processing proceeds by repeating Steps 3 through 5 until the entire GPS stream has been processed. This completes the trip-diary extraction for the vehicle under consideration. The data for the next vehicle in the input are then read in and Steps 3, 4, and 5 are executed in a loop until the diary data have been generated for all vehicles.

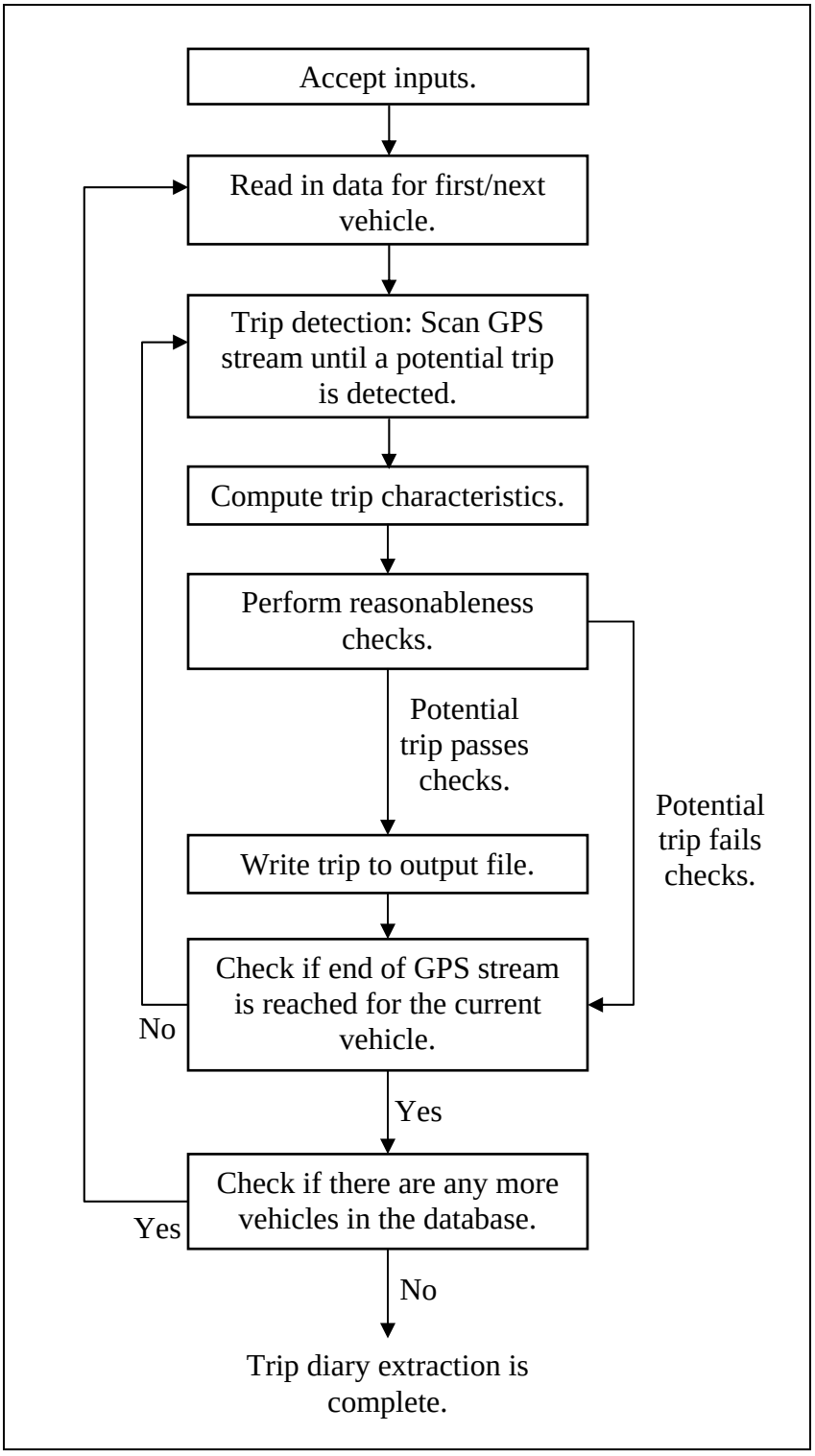

Figure 3.1 Conceptual overview of the travel-diary extraction algorithm

3.1 Step 1: Accept Inputs

The inputs to the GPS-TDG software are:

1. The preprocessed GPS navigational streams. Each record (or line) of this file contains the following information: household and vehicle identifiers, local date and time, latitude, longitude, speed, heading, and the number of invalid records immediately prior to this record in the raw file.
2. The demographics data file identifying the individual and household characteristics of the primary drivers of the GPS-equipped vehicles. If only the basic analysis is to be performed, the data required are the home and work locations. If enhanced analysis is to be performed, all demographic data required to predict activity types using the activity-type classification model are also included. Such additional demographic data required in the context of the Laredo application are employment and student status of the individual, gender, and the number of children in the household.
3. A GIS layer identifying the TAZ boundaries.
4. A GIS land-use layer (required only for enhanced analysis). Locally available parcel-level data with a rich land-use classification scheme are suitable. The Laredo application of the software uses such parcel-level zoning data.
5. Specification of the MNL model for activity-type classification (required only for enhanced analysis). This includes the activity-type classification scheme employed, the list of explanatory variables to predict the disaggregate activity purpose for each trip, and the vector of model parameters.
6. Algorithm parameters. The trip detection and characterization algorithm implemented in the GPS-TDG software are controlled by several user-specified parameters. The values of these parameters can be varied as required by the analyst to suit the requirements of different study areas and for various analysis objectives. (Default values are presented in Chapter 4.) A description of all the user-defined parameters is presented in Table 3.1. The use of these thresholds in the trip detection and characterization algorithm is discussed in the next section.

Table 3.1 List of user-specified algorithm parameters

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

3.2 Step 2: Detect Potential Trip Ends

The detection of a potential trip involves identifying a pair of GPS records within the overall navigational stream, with one record corresponding to the end of a trip and the other corresponding to the start of the next trip. The first valid GPS record in the stream corresponds to the start of the first trip. Based on prior GPS trip research, there are two types of trips to be detected: those involving an engine power-off (e.g., parking at a location for work, shopping, at home) and those in which the engine is not turned off e.g., dropping someone off at a location, going through a drive-through bank or fast food location). The detection procedure for each of these cases is discussed below. All user-specified algorithm parameters referenced henceforth are underlined.

3.2.1 The engine power-off case

In the engine power-off case, a pair of successive GPS records, one corresponding to the end of a trip and the next corresponding to the start of a subsequent trip are detected by examining the dwell times between successive valid records in the GPS stream. The dwell time is computed by subtracting the time resulting from signal loss from the total time between successive valid GPS records. Thus the procedure guards against classifying time gaps caused by extended periods of signal loss as trip ends. The procedure is as follows:

1. Compute the total time difference (*TotTimeDiff*) between successive GPS records as the difference in the time stamps of the two records.
2. Compute the signal loss time (*SigLossTime*) between the successive GPS records as the product of the number of invalid records removed (during preprocessing) immediately prior to the second GPS record and the Average Update Rate. The reader is referred to Section 4.1 for the discussion on preprocessing and the removal of invalid records.
3. Compute the Dwell Time (*DwellTime*) as the difference between *TotTimeDiff* and *SigLossTime*.
4. If the *DwellTime* between a pair of successive GPS records exceeds the Engine Power-off Dwell-Time Threshold, a potential trip end is flagged. The first of these two successive GPS records corresponds to the end of a trip and the second corresponds to the start of the next trip.

In addition, a potential trip end is also flagged at the end of the GPS stream for the vehicle and the last GPS record represents the trip end of this last trip.

3.2.2 The non–engine power-off case

Stops without engine power-off cannot be detected using the dwell time concept, because the GPS points are continually recorded, even during the period when the vehicle is stopped. Hence, we examine the speed data to identify nonmovement and a non–engine power-off trip end. Specifically, if the instantaneous speed recorded in the GPS stream is less than the Speed Threshold for Non–Engine Power-off Stop Determination continuously for a period greater than the Non–Engine Power-off Dwell-Time Threshold, a potential trip end is detected. The GPS record from which the speed continually remains below the threshold value is taken to represent the end of a trip. The first subsequent GPS record with a speed above the threshold value represents the start of the next trip.

3.3 Step 3: Compute Trip Attributes

Once a potential trip end is detected using the procedures described in the previous section, several trip attributes are computed. These include (1) trip start and end times, (2) trip-end locations (TAZs), (3) trip length, (4) trip speed, (5) activity type at trip destination, and (6) accuracy measure. The methods for computing each of these attributes are discussed in the paragraphs that follow. The reader will note that, from the above attributes, it is straightforward to derive additional attributes such as trip duration, activity duration at trip end, and trip purpose.

3.3.1 Trip start and end times

The start time of a trip is computed as the time stamp on the first GPS navigational stream record corresponding to a trip and the end time of a trip is computed as the time stamp on the last GPS navigational stream record corresponding to the same trip. The trip duration can then be computed as the difference between the start and end times of the trip.

The duration of activity at a trip end can be computed as the difference between the end time of a trip and the start time of the subsequent trip. The reader will note that the activity duration at the origin of the first trip and at the destination of the last trip cannot be determined.

3.3.2 Trip-end locations

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

3.3.3 Trip length

The trip length (or distance) is determined using the point-to-point sum of distances approach (the PP approach; see Research Report 1 for a detailed discussion on alternate methods). Broadly, this method involves computing the distance between successive pairs of recorded locations, the two points in each pair being spaced apart by at least the Trip Distance Computation Time Threshold. The reader will note that, by computing the distances between GPS points that are 5 or 10 seconds (the Trip Distance Computation Time Threshold) apart rather than 1 second apart, it is possible to minimize the overestimation of the trip distance resulting from the positional errors associated with the GPS devices (see detailed discussion in Research Report 1). These distances are then summed over the entire trip to obtain the trip length. The summing of the distances is not performed over segments of the trip where the speed is less than the Speed Threshold for Trip Distance Computation (i.e., short stretches when the vehicle is not moving).

Further, two different methods are used to compute the distance between pairs of locations in the PP approach. In the first method the latitude and longitude information for the two points are used to determine the distance using the following formula:

$$D = \left([b_1 - a_1]^2 + [b_2 - a_2]^2 \right) * 69.1$$

where,

a_1 and a_2 represent, respectively, the latitude and longitude of the first point (in decimal degrees),

b_1 and b_2 represent, respectively, the latitude and longitude of the second point (in decimal degrees).

In the second method the distance is computed as the product of the recorded instantaneous speed and the time between the pair of GPS points. Thus the software provides two estimates of the trip distance, one based on position information and the other based on speed information collected by the GPS devices. If actual trip distance measurements were also available (e.g., from odometer readings) along with GPS streams, these two methods can be validated to determine the better method.

3.3.4 Trip speed

The instantaneous speeds are averaged over all the GPS records corresponding to a trip to compute the average trip speed. The standard deviation of the instantaneous speed measurements is also computed to provide a measure of variation in speed along the trip length. A second approach is to calculate instantaneous speeds as the ratio of the distance between successive GPS records to the difference between the time stamps of these records.

3.3.5 Activity type at trip destination

The activity type undertaken by the driver of the vehicle at the trip-end location is determined next. This varies according to whether or not the user is in the basic or enhanced mode of the software.

In the basic analysis mode, the trip-end activities are classified into one of the following three aggregate types: home, work, and other. As already discussed in Section 3.1, the locations of home and work in terms of latitude and longitude are provided as inputs to the algorithm. A trip-end activity is classified as "home" if the distance of the trip-end location from home is less than the Home Location Distance Threshold. (This is necessary to account for the difference between where the vehicle is parked and where the home is located.) A trip-end activity is classified as "work" if (1) the distance of the trip-end location from work is less than the Work Location Distance Threshold and (2) the activity duration at the trip end is greater than the Work Duration Threshold. (Again, it is necessary to consider vehicles being parked in garages and lots

off-site from the main work location.) Trip ends that are not classified as home or work are classified as "other" by default.

In the enhanced analysis mode, the trip-end activities classified as "other" are further classified into disaggregate types if land-use data and the explanatory variables and parameters of the multinomial logit model for the activity-type classification are also provided as inputs. For ease in presentation, the probabilistic procedure for disaggregate activity-type determination is described here in the context of three activity types: shopping, leisure, and serve-passenger. This procedure can be extended in a straightforward manner to any number of activity types.

1. Let X_S, X_L, and X_{SP} be the values of the explanatory variables corresponding to shopping, leisure, and serve-passenger, respectively.
2. Let β be the vector of coefficients (i.e., model parameters) estimated for the above explanatory variables.
3. Compute the probability that the current trip-end activity type is shopping, as
$$\Pr(Shop) = \frac{\exp(\beta X_S)}{\exp(\beta X_S) + \exp(\beta X_L) + \exp(\beta X_{SP})}.$$
4. Similarly, compute the probability that the current trip-end activity type is leisure and serve-passenger, respectively, as
$$\Pr(Leisure) = \frac{\exp(\beta X_L)}{\exp(\beta X_S) + \exp(\beta X_L) + \exp(\beta X_{SP})}$$
$$\Pr(ServePax) = \frac{\exp(\beta X_{SP})}{\exp(\beta X_S) + \exp(\beta X_L) + \exp(\beta X_{SP})}$$
5. Draw a random number, U, from the uniform [0,1] distribution.
6. The assignment of the trip-end activity type is accomplished as follows:
    a. If $0 \le U \le \Pr(Shop)$, assign shopping as the activity type.
    b. If $\Pr(Shop) \le U \le \left(\Pr(Shop) + \Pr(Leisure)\right)$, assign leisure as the activity type.
    c. If $\left(\Pr(Shop) + \Pr(Leisure)\right) \le U \le 1$, assign serve-passenger as the activity type.

The activity types at the origin and the destination of the trip can then be used to determine the trip purpose. For example, if the activity types were determined using the basic

analysis mode, the following procedure provides the three commonly used trip-purpose categories:

1. If the activity type at one of the trip ends (i.e., origin or destination) is "home" and the activity type at the other trip end is "work," then the trip purpose is "home-based work" (HBW).
2. If the activity type at one of the trip ends (i.e., origin or destination) is "home" and the activity type at the other trip end is not "work," then the trip purpose is "home-based non-work" (HBNW).
3. If the activity type at both trip ends is not "home," then the trip purpose is "non-home based" (NHB).

With the computation of disaggregate activity purposes at trip-ends, it is also possible to further disaggregate home-based non-work and non–home-based trips into finer categories. The above procedure can be extended in this regard and the output file containing the derived trip diaries can be easily manipulated within any spreadsheet software to compute the disaggregate trip purposes of interest.

3.3.6 Accuracy measure

An accuracy measure is computed for each trip detected. This measure, called *NRecRatio,* is computed as follows:

$$NRecRatio = \frac{N_{Valid}}{N_{Valid} + N_{Invalid}}$$

where N_{Valid} is the number of valid GPS points for the trip and $N_{Invalid}$ is the number of invalid GPS points for the trip. As discussed in Section 2.2, the preprocessor removes the invalid records from the raw GPS stream and records the number of such invalid records removed immediately prior to each valid GPS record. Thus *NRecRatio* is a measure of the extent of missing or invalid GPS data for a trip. Smaller values of this measure indicate that a significant fraction of the complete GPS records corresponding to this trip were invalid and hence the trip attributes computed are less accurate than those records with higher values.

3.4 Step 4: Reasonableness Checks

One could introduce reasonableness checks on each attribute and also could examine combinations of attributes (e.g., trip timing and purpose) to ensure that the trip characteristics predicted are intuitive. Based on examinations of the trips detected and characterized using the procedures described in Sections 3.2 and 3.3, we have identified two checks. The first check ensures that the trip duration is of at least a certain minimum value by comparing the computed duration against the Minimum Trip Duration Threshold. The second check ensures that the trip length (distance) is of at least a certain minimum value by comparing the computed trip length against the Minimum Trip Length Threshold. Potential trips identified using the procedures described in Section 3.2, which have trip durations lower than the Minimum Trip Duration Threshold or trip lengths lower than the Minimum Trip Length Threshold, are classified as false trips.

CHAPTER 4. SOFTWARE IMPLEMENTATION DETAILS

This chapter discusses the structure of the GPS-TDG software conceptualized in Chapter 2. The prototype software has been developed using the Object-Oriented Programming Paradigm in the Java programming language. ArcGIS 9.0 is used as the underlying GIS platform. The querying module is implemented using the ODBC functionality provided with Microsoft Office.

The software comprises four main components. These are (1) the preprocessor, which converts the raw GPS navigational streams into the desired input format; (2) the process module, which implements the travel diary extraction algorithm; (3) the querying module, which allows the user to perform aggregation-analyses on the generated travel diaries; and (4) a graphical user interface, which ties the three preceding components together and allows the analyst to interact with the software with ease. Each of these components is discussed in a separate section below.

4.1 The Preprocessor

The preprocessor (implemented as a stream processor) takes as inputs the raw GPS streams and produces as outputs "cleaned" GPS files (i.e., invalid records removed) restructured for subsequent input to the main process module of the GPS-TDG software. The raw GPS navigational streams represent the GPS data largely in the same format in which they have actually been recorded by the in-vehicle devices. It is assumed that the stream from each vehicle is saved as a separate comma-separated file (i.e., one file per surveyed vehicle) with each line comprising a sequence of data elements in a predefined format (this sequence is also called a sentence). The sentence structures of the input GPS streams that the preprocessor can handle are presented in Section 4.1.1. Section 4.1.2 presents the structure of the preprocessor outputs and the implementation of the preprocessing methodology is described in Section 4.1.3.

4.1.1 Raw GPS sentence structures

The preprocessor is capable of handling sentences that are either in the "GPRMC" or in the "GeoLogger" format. The raw GPS files are to be named with a ".dlv" extension if they are in the GeoLogger format and with a ".gprmc" extension if they are in the GPRMC format.

The GeoLogger Sentence Format

Example:

5,273,1005361,1,03/25/02,13:13:24,03/25/02,07:13:24,27.52020,99.46400,00134,0.00,000,01.8,00

Generic Structure:

RecType, GPSID, HHID, VehID, UTC_Date, UTC_Time, Loc_Date, Loc_Time, Lat_Raw, Long_Raw, Elev_Raw, Speed, Heading, HDOP, SATS

1. RecType takes the value 5 indicating a GPS record.
2. GPSID is the identifier for the GPS device used.
3. HHID is the identifier for the household.
4. VehID is the identifier for the GPS-equipped vehicles in a household.
5. UTC_Date is the UTD date in the MM/DD/YY format.
6. UTC_Time is the UTC military time in the HH:MM:SS format.
7. Loc_Date is the local date in the MM/DD/YY format.
8. Loc_Time is the local military time in the HH:MM:SS format.
9. Lat_Raw is the instantaneous latitude position recording in decimal degrees.
10. Long_Raw is the instantaneous longitude position recording in decimal degrees.
11. Elev_Raw is the instantaneous elevation (in meters).
12. Speed is the instantaneous speed recoding in meters per second (0 to 514 mps).
13. Heading is the direction of movement in degrees (0 to 359.9 degrees).
14. HDOP is the horizontal dilution of precision (0.5 to 99.9).
15. SATS is the number of satellites in view (0 to 12).

The GPRMC Sentence Format

Example:

$GPRMC,040113,A,3653.10,S,17437.47,E,000.6,074.4,190903,,*03

Generic Structure:

"$GPRMC", Time, Status, Latitude, Lat_Hem, Longitude, Long_Hem, Speed, Heading, Date, Blank, CheckSum

1. "$GPRMC" is a record-type identifier.
2. Time is the UTC time in HHMMSS.SSS format.
3. Status field takes the value "A" if the record is valid or "V" otherwise.
4. Latitude is the instantaneous latitude position recording (ddmm.mmmm format).
5. Lat_Hem represents the latitude hemisphere (N = North and S = South).
6. Longitude is the instantaneous longitude position recording (dddmm.mmmm format).
7. Long_Hem represents the longitude hemisphere (E= East and W = West).
8. Speed is the instantaneous speed recording in knots (values between 0 and 999.9).
9. Heading is the direction of movement in degrees (0 to 359.9 degrees).
10. Date is the UTC date in the DDMMYY format.
11. Blank is a blank entry.
12. Checksum (the *03 entry in the example).

If data are recorded in the GPRMC format, identifiers for the GPS unit, the household, and the vehicle have to be manually added to files containing the GPS streams. This is accomplished by adding a single line header with the identifiers to each raw GPS stream file. The identifiers are in the following sequence: GPSID, HHID, VehID.

4.1.2 Output format

The preprocessed GPS streams generated serve as subsequent inputs to the main trip-detection component of the GPS-TDG software. These files are written out in the comma-separated format with a ".cln" extension. One file corresponds to each vehicle in the survey. The structure of each row of data for these files is discussed below.

Example:

GREC,273,1005361,1,1017058405000,27.52020,-99.46402,0.00,000,1

Generic Structure:

“GREC”, GPSID, HHID, VehID, Time, Latitude, Longitude, Speed, Heading, NumInvRecs

1. “GREC” is a record-type identifier indicating that this is a record in a cleaned GPS stream file.
2. GPSID is the identifier for the GPS device used.
3. HHID is the identifier for the household.
4. VehID is the identifier for the GPS-equipped vehicles in a household.
5. Time is the instantaneous time in milliseconds from 1970 (corrected for DST).
6. Latitude is the instantaneous latitude position recording in decimal degrees.
7. Longitude is the instantaneous longitude position recording in decimal degrees.
8. Speed is the instantaneous speed recording in meters per second (0 to 514 mps).
9. Heading is the direction of movement in degrees (0 to 359.9 degrees).
10. NumInvRecs is the number of invalid records immediately prior to this GPS record that were removed during preprocessing.

Each of the preprocessed GPS streams file also contains a header record and a trailer record. The header record provides the parameter values used in preprocessing. For example, the header row of preprocessed GPS files created from raw GeoLogger files could be:

HREC,HDOP=5.0,Min Satellites=3

This indicates that GPS records with an HDOP value greater than 5 or number of satellites less than 3 were considered invalid and removed.

In the case that the raw GPS streams have the GPRMC format, the A/V flag of the GPS sentences are used in preprocessing. Therefore the header row of a corresponding cleaned file does not have any parameter values and simply reads HREC, GPRMC.

The trailer record is identified by the "TREC" field. This is followed by the total number of GPS records in the input file and the total number of GPS records designated as invalid by the preprocessor.

4.1.3 Implementation

The preprocessor is implemented as a stream processor. Its main steps are:

1. Determine the type of input GPS stream file being preprocessed (i.e., ".gprmc" or ".dlv").
2. If the input has a ".gprmc" format, go to Step 12.
3. Accept inputs for *max_HDOP* (maximum dilution in precision) and *min_NumSat* (minimum number of satellites) parameters.
4. Open output file (with a .dlv.cln extension) and write the header (i.e., HREC,HDOP= *max_HDOP*,Min Satellites= *min_NumSat*).
5. Set *Num_Skipped_Rec* = 0 (this variable keeps count of number of invalid records removed prior to each valid record).
6. Repeat Steps 7-9 to the end of input GPS stream file.
7. Read in the first or next sentence from the input GPS stream
8. If the number of satellites recorded in the sentence is greater than *min_NumSat* and the HDOP value recorded is less than *max_HDOP,*
    a. Create a new row in the output file and write out "$GREC" and subsequently copy the GPSID, HHID, VehID, Time, Latitude, Longitude, Speed, Heading values from the input GPS sentence separated by commas.
    b. Add *Num_Skipped_Rec* to the end of this record.
    c. Set *Num_Skipped_Rec* = 0.
9. Else (i.e., condition specified in step 8 is not satisfied) increment *Num_Skipped_Rec* by 1
10. Write the trailer for the .dlv.cln file.
11. Go to Step 19.
12. Open output file (with a .gprmc.cln extension) and write the header (i.e., HREC, GPRMC).

13. Set *Num_Skipped_Rec* = 0.
14. Repeat Steps 15-17 to the end of input GPS stream file.
15. Read in the first or next sentence from the input GPS stream.
16. If the "A/V" entry recorded in the input stream is "A,"
    a. Create a new row in the output file and write out "$GREC" and subsequently copy the GPSID, HHID, VehID, Time, Latitude (converted to decimal degrees), Longitude (converted to decimal degrees), Speed (converted to miles per hour), Heading values from the input GPS sentence separated by commas.
    b. Add *Num_Skipped_Rec* to the end of this record.
    c. Set *Num_Skipped_Rec* = 0.
17. Else (i.e., condition specified in step 16 is not satisfied) increment *Num_Skipped_Rec* by 1.
18. Write the trailer for the .gprmc.cln file.
19. Stop.

4.2 The Process Module

The process module is the core component of the GPS-TDG software that implements the trip-diary extraction algorithm described in the previous chapter. The structure of the various inputs to this module is discussed in Section 4.2.1. Section 4.2.2 describes the derived trip file generated as output by the software. The implementation details are presented in Section 4.2.3.

4.2.1 Input formats

The process module of the GPS-TDG takes as inputs (1) spatial (i.e., GIS-based) data, (2) nonspatial data, and (3) parameters. The spatial data inputs are the TAZ-boundaries GIS layer and the land-use GIS layer. The input formats of these data items are discussed in Section 4.2.1.1. The nonspatial data inputs are the preprocessed GPS streams and the individual and households characteristics. The structure of the former has been discussed in Section 4.1.2. The latter data are provided as two input files, namely, the demographics file and the relationships file. These are discussed in Section 4.2.1.2. In addition to providing data inputs, the analyst also

provides the parameters to be used in the algorithm and the specification of the MNL model for activity-type determination (if the software is run in the enhanced analysis mode). These parameter inputs are discussed in Section 4.2.1.3.

4.2.1.1 Spatial data inputs

The spatial data inputs are the TAZ-boundaries GIS layer and the land-use GIS layer. The TAZ-boundaries GIS layer is to be provided as an ESRI shape file. The software is capable of handling any co-ordinate system as along as the TAZ Ids are provided in the 7th column called "TAZ" in the underlying database. The GIS land-use layer is also to be provided in the ESRI shape file format. Again, the software is capable of handling any co-ordinate system; however, the land-use variable must be provided in the 4th column called "P_LANDUSE" in the underlying database. Further, the following zonal land-use classification scheme must be used: "LDR," "MDR," "HDR," "HEAVY INDUSTRIAL," "INSTITUTIONAL," "RETAIL/OFFICE," "AG," "WATER," "PARK/REC. OPEN SPACE," and "WAREHOUSE/LI"

4.2.1.2 Household and person characteristics

Data on individual and household characteristics of the survey respondents are provided as two files: (1) the Demographics File, and (2) the Relationship File. The first file includes the household and work locations of the survey respondents and other characteristics that are required for activity type determination in the enhanced mode. The second file associates or links each vehicle equipped with the GPS device with one member of the household called the primary driver. The structure of these files is presented below.

<u>The Demographics File</u>

This is a comma-separated file with a ".demo" extension. This file comprises two types of records: the household record and the person record. The surveyed households are represented sequentially in this file. Each household is represented by one household record followed immediately by one or more person record. Every person in the household who has been identified as a primary driver of a GPS-equipped vehicle should be represented by a person

record. Individuals who are not primary drivers (e.g., adults without a driver's license and children) need not be included in this file.

Example (representation for a household with one primary driver):

DREC,1,1001373,-99.470800, 27.469860,2

DREC,2,1001373, 1,1,-99.479140, 27.571130,1,1

The generic structure of the household record is as follows:

"DREC," RecType, HHID, HomeLocLat, HomeLocLong, NumChild

1. "DREC" is a record-type identifier indicating that this is a record in a demographics file.
2. RecType is the identifier used to distinguish between household and person records in the location and demographics file. Household records have an entry 1.
3. HHID is the identifier for the household.
4. HomeLocLat is the latitude of the home location in decimal degrees.
5. HomeLocLong is the longitude of the home location in decimal degrees.
6. NumChild is the number of children (age < 16) in the household. This is required only for enhanced analysis. If the software is to be run only in the basic mode, this user can specify a value of zero for this field.

The person record has the following generic format:

"DREC," RecType, HHID, PersID, EmpFlag, WorkLocLat, WorkLocLong, Student, Male

1. "DREC" is a record-type identifier indicating that this is a record in the location and demographics file.
2. RecType is the identifier used to distinguish between household and person records in the location and demographics file. Person records have an entry 2.
3. HHID is the identifier for the household.
4. PersID is the identifier for a person (primary driver) in the household.

5. EmpFlag is a binary flag used to indicate the employment status of the individual (1 if employed).
6. WorkLocLat is the latitude of the work location in decimal degrees (provide a value of zero if the person is not employed).
7. WorkLocLong is the longitude of the work location in decimal degrees (provide a value of zero if the person is not employed).
8. Student is a binary flag used to indicate the student status of the individual (1 if a student). This is required only for enhanced analysis. If the software is to be run only in the basic mode, this user can specify a value of zero for this field.
9. Male is a binary flag used to indicate the gender of the individual (1 if male). This is required only for enhanced analysis. If the software is to be run only in the basic mode, this user can specify a value of zero for this field.

The Relationship File

The relationship file (also called the "link" file) is used to specify an association between each vehicle equipped with a GPS device and a member of the household who is the "primary driver" of this vehicle. Hence, this file provides a link between the files containing the GPS streams and the person-level information contained in the demographics file. There is one record in this file for each GPS-equipped vehicle in the survey. This is a comma-separated file with a ".link" extension.

Example:

LREC,040802_1001373_V1.dlv,1001373,1,1

Generic Format:

"LREC," GPSFilename, HHID, VehID, PersID

1. "LREC" is a record-type identifier indicating that this is a record in the link file.
2. GPSFilename is the name of file containing the raw GPS stream corresponding to the vehicle VehID.

3. HHID is the identifier for the household.
4. VehID the identifier for the GPS-equipped vehicles in a household.
5. PersID is the identifier for the primary driver in the household associated with the vehicle VehID.

4.2.1.3 Parameter inputs

GPS-TDG takes two types of parameters as inputs: (1) specification of the MNL model for activity-type classification, and (2) algorithm parameters. These are discussed in further detail below.

MNL Model Specification

The disaggregate activity-type classification and the parameters of the MNL model (i.e., coefficients on the explanatory variables) are specified using a comma-separated file. The first row of input defines the structure of this file. Each subsequent row of this file comprises the set of parameters defining the "utility function" for one activity purpose. The parameters are input in the sequence defined in the first line. The analyst can classify the activities into any number of disaggregate purposes and provide the appropriate number of rows of input. In this version of the software, the set of explanatory variables that can be used are predetermined. These variables and the specification structure are discussed below:

Example (specification of parameters for one of the activity purpose types)

MREC,Serve_passenger_at_otherlocation,0.0,0.0,0.0,0.9835,0.0,0.0,0.0,0.0384,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.1509,5.8298

The first entry in this record is "MREC," which identifies this record as containing model parameters. The second field has the name of the activity type. The subsequent records represent the coefficients on the explanatory variables corresponding to this activity type. The sequence of explanatory variables and their descriptions are provided in the table below:

Table 4.1 Input structure for the MNL model for activity-type classification

Column	Variable Name	Description
3	Industrial_or_warehouse	1 if the trip end falls in a zone with industrial or warehouse land use, 0 otherwise
4	Institutional	1 if the trip end falls in a zone with institutional land use, 0 otherwise
5	Retail/office_space	1 if the trip end falls in a zone with retail or office land use, 0 otherwise
6	LDR	1 if the trip end falls in a zone with LDR land use, 0 otherwise
7	MDR_or_HDR	1 if the trip end falls in a zone with MDR or HDR land use, 0 otherwise
8	No_land_use	1 if the trip end falls in a zone with blank or no land use, 0 otherwise
9	Trip_duration	Duration of the trip in seconds
10	Activity_duration	Duration of the activity at trip end in seconds
11	7_AM_9_AM	1 if trip ends between 7 AM and 9 AM, 0 otherwise
12	9_AM_2_PM	1 if trip ends between 9 AM and 2 PM, 0 otherwise
13	2_PM_6_PM	1 if trip ends between 2 PM and 6 PM, 0 otherwise
14	after_6_PM	1 if trip ends after 6 PM, 0 otherwise
15	Trip_origin_is_home	1 if the trip originates at home, 0 otherwise
16	Trip_origin_is_work	1 if the trip originates at work, 0 otherwise
17	Employed	1 if the primary driver is employed, 0 otherwise
18	Student	1 if the primary driver is a student, 0 otherwise
19	Male	1 if the primary driver is male, 0 otherwise
20	Number_of_children_in_HH	Number of children in the household
21	Constant	Alternative-specific constant

The default MNL model parameters provided with the software were estimated using data from Laredo employing a nine-way classification scheme for the activity types. The complete empirical specification of this model in the input format is presented in the Users Manual accompanying the software. The reader is referred to Report 2 for the model estimation details.

Algorithm Parameters

The trip detection and characterization algorithm implemented in the GPS-TDG software is controlled by several user-specified parameters. The values of these parameters can be varied as required by the analyst to suit the requirements of different study areas and for various analysis objectives. A description of all the user-defined parameters is presented in Table 4.2 below. The prescribed default values are also provided.

Table 4.2 Algorithm parameters and default values

Parameter Name	Flag	Description	Default Flag/Value
Debug Options Flag	-debug or -nodebug[1]	Determines whether the software is run in the debug mode (-debug) or not (-nodebug). The analyst will only run the software in the –nodebug mode. The –debug option was developed for software testing purposes	-nodebug
Non–Engine Power-off Flag	-nepo or -nonepo[1]	Determines whether the algorithm to determine non–engine power-off stops is executed (-nepo) or not (-nonepo).	-nepo
Engine Power-off Dwell Time Threshold	-t	Minimum dwell-time gap between successive valid GPS records for signaling an engine power-off stop (seconds)	30 secs
Non–Engine Power-off Dwell-Time Threshold	-t2	Minimum duration for which the speed should be less than the Speed Threshold for signaling a non–engine power-off stop	60 secs
Speed Threshold for Non– Engine Power-off Stop Determination	-s2	Value of instantaneous speed below which the vehicle is assumed to be at rest (used in determination of non–engine power-off stops)	0.1 mps
Speed Threshold for Trip Distance Computation	-s	If the value of instantaneous speed is below this threshold, this GPS point is not used in distance computation	0.5 mps
Time Threshold for Trip Distance Computation	-i	Minimum time between pairs of GPS points for which the distances are computed and summed to determine the trip length	1 secs
Home Location Distance Threshold	-h	Maximum distance between a trip-end location and home location for classifying the trip-end activity purpose as "Home"	0.3 miles
Work Location Distance Threshold	-w	Maximum distance between a trip-end location and work location of primary driver for classifying the trip-end activity purpose as "Work"	0.5 miles
Work Duration Threshold	-d	Minimum activity duration at a trip end for classifying a trip-end activity as "Work"	3600 sec

Average Update Rate	-u	Average duration between successive navigational points recorded by the GPS instrument used in the survey	1/sec
Minimum Trip Duration Threshold	-mt	Minimum trip duration for a potential trip to be classified as a real trip	120 secs
Minimum Trip Length Threshold	-ml	Minimum trip length (distance) for a potential trip to be classified as a real trip	1 miles

[1] there is no value following this flag

4.2.2 Output format

The derived travel diary constitutes the fundamental output from the GPSTDG software. The file containing these derived travel patterns is written out in two formats. The first file with a ".trp" extension is a compact comma-separated file with headers. The second file is a flat file (with a ".csv" extension). This file structure is useful for display and processing queries. Further, the file format can also be opened in EXCEL or other spreadsheet applications with ease for further analysis.

Format of the ".trp" file

The overall format of the ".trp" file is as follows:

"HREC", WorkThresh, HomeThresh, SpeedThresh, DwellTime, MinWorkDur

"VH", HHVehID, DriverID

"TR", OriginalTripNo, TripNo, TripDuration, StartTAZ, EndTAZ, StartLU, EndLU, StartActivity, EndActivity, StartLat, StartLong, EndLat, EndLong, StartDate, EndDate, TripLength1, AvSpeed1, VarSpeed1, TripLength2, AvSpeed2, VarSpeed2, StartTOD, EndTOD, NRecRatio

"VT", HHVehID, NumTrips

….

TREC, TripCount, VehCount

The first record in this file is a header characterized by the entry “HREC” in its first field. This header record sequentially includes the values of all the threshold parameters used in the travel-diary extraction algorithm. The complete format of this header is as follows:

“HREC,” WorkThresh, HomeThresh, SpeedThresh, DwellTime, MinWorkDur

1. “HREC” is a record-type identifier indicating that this is a header record in the output file.
2. WorkThresh is the parameter defining the maximum distance between a trip-end location and work location of primary driver for classifying the trip-end activity purpose as "Work."
3. HomeThresh is the parameter defining maximum distance between a trip-end location and home location for classifying the trip-end activity purpose as "Home."
4. SpeedThresh is the parameter for trip distance computation (if the value of instantaneous speed is below this threshold, this GPS point is not used in distance computation).
5. DwellTime is the parameter defining the minimum dwell-time gap between successive valid GPS records for signaling an engine power-off stop (seconds).
6. MinWorkDur is the parameter defining the minimum activity duration at a trip end for classifying a trip-end activity as "Work."

The very last record in the output file is a trailer record with the following format:

“TREC”, TripCount, VehCount

where

1. “TREC” is a record-type identifier indicating that this is a trailer record in the output file.
2. TripCount is the total number of trips derived from GPS streams contained in the output file.
3. VehCount is the total number of vehicles in the output file.

The rest of the output file is comprised of three types of records: the vehicle header records, the vehicle trailer records, and the trip records. The vehicle header and vehicle trailer records, respectively, are included immediately before and after the trip records corresponding to

a particular vehicle and thus help to distinguish the trips of one vehicle from another. The vehicle header record has the following format:

“VH”, HHVehID, DriverID

where

1. “VH” is a record type identifier indicating that this is a vehicle header record.
2. HHVehID the identifier for the GPS-equipped vehicles in a household. This is created by appending the vehicle ID to the household ID.
3. DriverID is the identifier for the primary driver in the household associated with the vehicle HHVehID. This is created by appending the person ID to the household ID.

The trailer record has the following format:

“VT”, HHVehID, NumTrips

where

1. “VT” is a record type identifier indicating that this is a vehicle trailer record.
2. HHVehID the identifier for the GPS-equipped vehicles in a household. This is created by appending the vehicle ID to the household ID.
3. NumTrips is the total number of trips detected for this vehicle. This is equal to the number of trip records included between the header and the trailer records of a particular vehicle.

Each trip record in this file contains the characteristics of a single trip in the following format:

“TR”, OriginalTripNo, TripNo, TripDuration, StartTAZ, EndTAZ, StartLU, EndLU, StartActivity, EndActivity, StartLat, StartLong, EndLat, EndLong, StartDate, EndDate, TripLength1, AvSpeed1, VarSpeed1, TripLength2, AvSpeed2, VarSpeed2, StartTOD, EndTOD, NRecRatio

where

1. TR is a record-type identifier indicating that this is a trip record.

2. OriginalTripNo is the trip number assigned originally to all potential trips. These trips are subsequently classified as either true trips and retained or as false trips and discarded. Further, a single potential trip could also be broken into two or more trips if non–engine power-off stops are detected.
3. TripNo is a trip identifier. The trips are sequentially numbered over the final set of "true" trips.
4. TripDuration is the duration of trip is seconds.
5. StartTAZ is the TAZ identifier for the trip-start location.
6. EndTAZ is the TAZ identifier for the trip-end location.
7. StartLU is the land use at the trip-start location.
8. EndLU is the land use at the trip-end location.
9. StartActivity is the activity type pursued by the driver of the vehicle at the origin of the trip (an aggregate classification scheme is used in the basic analysis and a disaggregate analysis scheme is used in the enhanced analysis).
10. EndActivity is the activity type pursued by the driver of the vehicle at the destination of the trip (the classification scheme is the same as that for StartActType).
11. StartLat is the latitude of the trip-start location.
12. StartLong is the longitude of the trip-start location.
13. EndLat is the latitude of the trip-end location.
14. EndLong is the longitude of the trip-end location.
15. StartDate is the date and time of the start of the trip.
16. EndDate is the date and time of the end of the trip.
17. TripLength1 is the trip length (in miles) computed using the instantaneous position information.
18. AvSpeed1 is the average speed of the trip (miles per hour) computed (speed computed as ratio of distance between successive GPS records to the time stamp between successive records).

19. VarSpeed1 is the variance in the speed over the length of the trip (speed computed as ratio of distance between successive GPS records to the time stamp between successive records).
20. TripLength2 is the trip length (in miles) computed using the instantaneous speed information.
21. AvSpeed2 is the average speed of the trip (miles per hour) computed using instantaneous speed data.
22. VarSpeed2 is the variance is the speed over the length of the trip.
23. StartTOD is the time of day of start of trip (in the hhmm format, for example, 1527 refers to 3:27 p.m.).
24. EndTOD is the time of day of end of trip (in the hhmm format).
25. NRecRatio is an accuracy measure computed as the ratio of the number of valid GPS records corresponding to this trip to the total number of records for this trip.

Format of the ".csv" file

The comma-separated file (or the flat file) has all the information as the trip records in the ".trp" file. The first field of a record in this file has the entry "TR" indicating a trip record. The next two fields are, respectively, the vehicle and driver identifiers (i.e., the HHVehID and HHPersID variables in the ".trp" file). The next twenty-four columns correspond respectively to the fields 2-25 in the trip records of the ".trp" file.

4.2.3 Implementation

This section of the chapter discusses the implementation details of the processes module. Overall, the software was developed using the Object Oriented Programming Paradigm. Section 4.2.3.1 identifies the classes and the relationships among the classes. Section 4.2.3.2 discusses the process flow logic for trip detection and characterization.

4.2.3.1 Classes and relationships

The object-oriented design of the software system resulted in the identification of the following classes for the process module: (1) Household, (2) Person, (3) Vehicle, (4) LinkRec, (5) Trip, (6) Location, (7) GPSRec, (8) ShapeFileModel, (9) MNLModel, (10) MNLSample, and (11) MNLResultMatrix. The attributes and methods encapsulated within each of these classes and the relationships among these classes are briefly discussed next.

The household class contains the location and demographic information of the household whose GPS records are currently being processed by the software. Similarly, the person class contains the work location and other demographic information of the primary driver of the vehicle whose GPS records are currently being processed.

The vehicle class contains a pointer to the preprocessed GPS file corresponding to the vehicle currently being processed. The vehicle class also encapsulates the method for trip detection. Specifically, this method implements the algorithm discussed in Section 3.2 to detect potential trips from a GPS navigational stream. Finally, the vehicle class also contains the method to write out the set of trips detected to the output file.

The data in the LinkRec class are associated with those in the relationship file. As a first step to the processing of data from any vehicle, a LinkRec object is created, which is then used to fetch appropriate data to the create household, person, and vehicle objects.

The trip class contains all data required to characterize a single trip. An instance of a trip class is created each time a potential trip is detected by the method included in the vehicle class. The methods to compute the various attributes of a trip (as discussed in Section 3.3) are encapsulated within this class. Thus, when provided with a stream of GPS records corresponding to a single trip, the trip class can completely characterize this trip.

The trip class also contains two instances of the location class, one corresponding to the start location of the trip and the other corresponding to the end location of the trip. The advantage of constructing a separate class for the locations is in the encapsulation of the GIS overlay procedures within this class. Specifically, the location class contains the procedure to determine the attributes of the zone within which the current location lies when provided with the underlying GIS layer. The Activity Purpose at the trip start or end is also encapsulated within the location at trip start and end.

The GPSRec class represents a single record of the GPSStream. An instance of this class is created to store each record of the GPS stream corresponding to the trip currently being extracted.

The ShapeFileModel is a class used to represent the GIS shape files in the software. When the program is run, two instances of the ShapeFileModel are created: one for the land-use input and the other for the TAZ-boundaries input. This class stores all the information from the corresponding shape files and provides a standard format for doing spatial queries. Specifically, the ShapeFileModel has procedures for determining the TAZ ID or the land-use classification, given the latitude and longitude of a point location. Since these procedures are invoked once corresponding to each trip-end detected, the ShapeFileModel class provides an efficient way of accessing the backend services provided by ArcGIS 9.0.

The last three classes to be discussed here all relate to the disaggregate activity-type determination in the enhanced version of the software.

The MNLModel class is used to internally represent the empirical specification of the MNL model. This class contains members representing each of the explanatory variables in the model and their coefficients. The contents of the comma-separated input file used to specify the activity-type classification and the coefficients on the explanatory variables are loaded into this class.

The MNLSample class is used to represent input data for any instance of application of the MNL model. This class contains members representing each of the explanatory variables used in the MNL model. When the trip-end activity type is to be determined, an instance of this class is created with the values of explanatory variables specific to the trip-end under consideration.

The MNLResultMatrix is a helper class that uses the empirical specification from the MNLModel class and the values of the explanatory factors from the MNLSample class to determine the disaggregate activity type at the trip-end under consideration. The activity-type determination is accomplished by two methods of this class, which are invoked sequentially. First, the “populate” method is invoked; it computes and stores the probabilities for each activity type. Next, the “doRandomDraw” method is invoked; it makes makes a deterministic assignment

of the activity type based on the computed choice probabilities and a uniform random number draw (see discussion in Section 3.3.5).

4.2.3.2 Process flow

The main steps in the processing logic implemented within the GPS-TDG software to extract the travel-diary data from the GPS navigational streams and other data are as follows:

1. Set the algorithm parameters based on user-specified inputs.
2. Load the relationships file and the demographics file into main memory as hashtables for easy access.
3. Load the TAZ file into an instance of the ShapeFileModel in memory.
4. If analysis is being performed in the enhanced mode, load the land-use file into an instance of the ShapeFileModel and the MNL parameters file into an instance of the MNLModel
5. Open the output file and write the header record.
6. Perform the following steps looping over the elements in the hashtable containing the relationship file data:
    a. Read in the first or next record from the relationship file hashtable and create an instance of the LinkRec class.
    b. If end of file is reached, proceed to step 7.
    c. Create a vehicle object for the vehicle contained in the LinkRec and assign the appropriate GPS file containing the navigational streams for this vehicle.
    d. Using the person-to-vehicle association provided in the LinkRec object, fetch the data for the appropriate person (primary driver) and household from the locations and demographics hashtable and create instances of person and household classes.
    e. Write the vehicle header to the output trip file.
    f. Perform the following steps looping over the GPS navigational stream records corresponding to the current vehicle:

i. Read the first or next record from the GPS navigational stream and create an instance of the GPSRec class and add this GPSRec object to a buffer in the vehicle class.

ii. If the end of the GPS navigational stream has been reached proceed to Step xii.

iii. Invoke the engine power-off trip detection procedure embedded within the vehicle class.

iv. If a potential engine power-off trip is not detected, revert to Step i.

v. If a potential engine power-off trip is detected, create a trip object and pass the buffer of GPS records corresponding to this trip to this object.

vi. The trip object computes all the characteristics of the current trip.

vii. If analysis is being performed in the basic mode go to Step x.

viii. Create an instance of MNLSample by combining the required information from the Trip and Vehicle currently working on.

ix. Create an instance of MNLResultMatrix, pass the MNLModel and the MNLSample, and invoke the required procedures to determine the disaggregate activity type at trip-end.

x. Add the newly created trip to the buffer of trips contained in the vehicle class.

xi. If the flag has been set for detecting non–engine power-off trips also, perform the following steps:

    A. Loop on the buffer of GPSRecs for the current trip (from Step v) and invoke the non–engine power-off trip detection procedure embedded within the vehicle class.

    B. If a potential non–engine power-off trip is not detected, go to Step A.

    C. If a potential non–engine power-off trip is detected, create a trip object and pass the buffer of GPS records corresponding to this trip

to this object. The subsequent steps are similar to those that follow the detection of an engine power-off trip (i.e., Steps vi-x).

D. On exiting from the loop defined in Step 1, pass the last portion of the buffer of GPS records to a trip object. Follow procedure similar to Steps vi-x to complete characterizing this trip.

xii. Flush the buffer of GPS records contained in the vehicle class and revert to Step i.

xiii. Create a trip object and pass the buffer of GPS records corresponding to the last trip.

xiv. The trip object computes all the characteristics of the last trip.

xv. Perform Steps viii and ix if enhanced analysis is being undertaken.

xvi. Add the newly created trip to the buffer of trips contained in the vehicle class.

g. Perform reasonableness checks based on minimum time and minimum distance on each trip contained in the vehicle object and write out only the reasonable trips to the output file.

h. Write out the vehicle trailer record to the output file and revert to Step a.

7. Extraction of the trips from all GPS-equipped vehicles is completed. Write the trailer record, close the output trip file, and quit.

4.3 The Querying Module

The querying module facilitates the analyst to aggregate the derived trip file and generate trip tables, inter-zonal performance measures (average trip length, duration, and speed), and other summary measures. This module has been developed with the intent of providing the analyst the convenience of performing some of the common aggregation-analyses from within the software. It is important to note that the process module of the software generates the trip file in a comma-separated format that can be easily loaded into any spreadsheet program for further analysis.

4.3.1 Queries supported

The data elements generated as query outputs are the number of trips, average trip length (miles), average trip duration (seconds), and average trip speed (miles per hour).

The aggregations can be performed, in the overall, or based on one or more of the following four criteria: (1) Trip-start location (TAZ), (2) Trip-end location (TAZ), (3) Activity type at trip-start location, and (4) Activity type at trip-end location. If the aggregation is performed over the entire file, the overall summary measures are generated. To generate trip-tables and inter-zonal performance measures, the analyst should aggregate based on both the trip start and end locations.

The analyst can also perform the above-described aggregations over a subset of trips satisfying particular criteria. The selection (or filtering) can be done based on one or more of the following criteria: (1) Trip-start location (TAZ), (2) Trip-end location (TAZ), (3) Activity type at trip-start location, (4) Activity-type at trip-end location, (5) Start time of the trip, and (6) End time of the trip. Further, the analyst is also provided the option of selecting a subset of trips based on trip purpose (i.e., home-based work, home-based non-work, and non–home-based trips). Thus it is possible to obtain, for example, the inter-zonal performance measures for specific purposes or for specific times of the day.

4.3.2 Implementation

The query functionality has been implemented using the ODBC functionality provided with Microsoft Office. Specifically, we use the Microsoft Text Driver to create a database out of the comma-separated trip file generated by the process module.

The two main classes used for implementing the queries are the CSVConnector and the Query. The CSVConnector class is used to set up the database and allows the software to connect to the database. The query class has member functions to set up the query, define the selection and aggregation criteria, and execute the query.

4.4 The Graphical User Interface (GUI)

The GPS-TDG provides a swing-based graphical interface for the user to interact with the software. This GUI allows the user to specify the inputs, run the preprocessor and process modules, view the generated trip file in a spreadsheet format, perform aggregations of the trip file using the querying module, and view the query results. This section describes the components and the implementation details of the GUI.

4.4.1 GUI components

The main window of the program, which remains on screen as long as the software is being used, comprises: (1) the Menu Bar, (2) the Command Area, and (3) the Data Area.

4.4.1.1 The Menu Bar

The Menu Bar comprises two items: (1) Tools and (2) Help. The tools menu allows the user to preprocess the raw GPS streams and convert them into a standard format for further use by the GPS-TDG software. The software is capable of taking as inputs the raw GPS streams that are in either the GPRMC or in the GeoLogger formats. Correspondingly, the user chooses either *Preprocess GPRMC* or *Preprocess GEO* from the drop-down Tools menu. The Help Menu has two items. Choosing the *About* menu item displays the software version and author information. Choosing the *Online Help* item opens the web browser to the help web page.

4.4.1.2 The Command Area

The Command Area provides the fundamental interface for the user to run the software to derive the trips from preprocessed GPS streams. In addition, the Command Area also provides the user with tools to query the derived trip file in many ways. The Command Area is organized into three panes: (1) the Basic Pane, (2) the Enhanced Pane, and (3) the Query Pane.

The Basic Pane of the Command Area allows the user to specify inputs, run the trip detection in the basic mode, and view the derived trip file. The Basic Pane comprises several buttons to specify the input data files, a text box to specify optional parameters to control the trip detection algorithm, and a button to run the trip detection algorithm and view the generated trip file.

The Enhanced Pane of the Command Area allows the user to specify inputs, run the trip detection in the enhanced mode, and view the derived trip file. This pane has all the elements of the basic pane, and two additional buttons to specify the additional inputs required for the enhanced analysis.

The Query Pane allows the user to aggregate and analyze the derived trip file in many ways. The aggregation and selection criteria (discussed in Section 4.3.1) are specified by the analyst using check boxes and text boxes.

4.4.1.3 The Data Area

The Data Area is used to display the outputs generated (in a spreadsheet format) from the process module and the querying module. The derived trips (outputs from the process module) are displayed in the conventional trip-diary format with each row of display corresponding to a single vehicle trip. The aggregate measures (outputs from the querying module) displayed are the number of trips, average trip length (miles), average trip duration (seconds), and average trip speed (miles per hour). The number of rows of data depends upon the aggregation criteria employed.

4.4.2 GUI implementation

The implementation of the graphical user interface is schematically represented in Figure 4.1. The DiaryGUI is the main class and forms the entry point to the software's GUI. This in turn comprises the following three elements (corresponding to the description in Section 4.4.1): (1) the MainMenuBar, (2) the DataArea, and (3) the LeftPanel. Implementation details of each of these components are discussed further below.

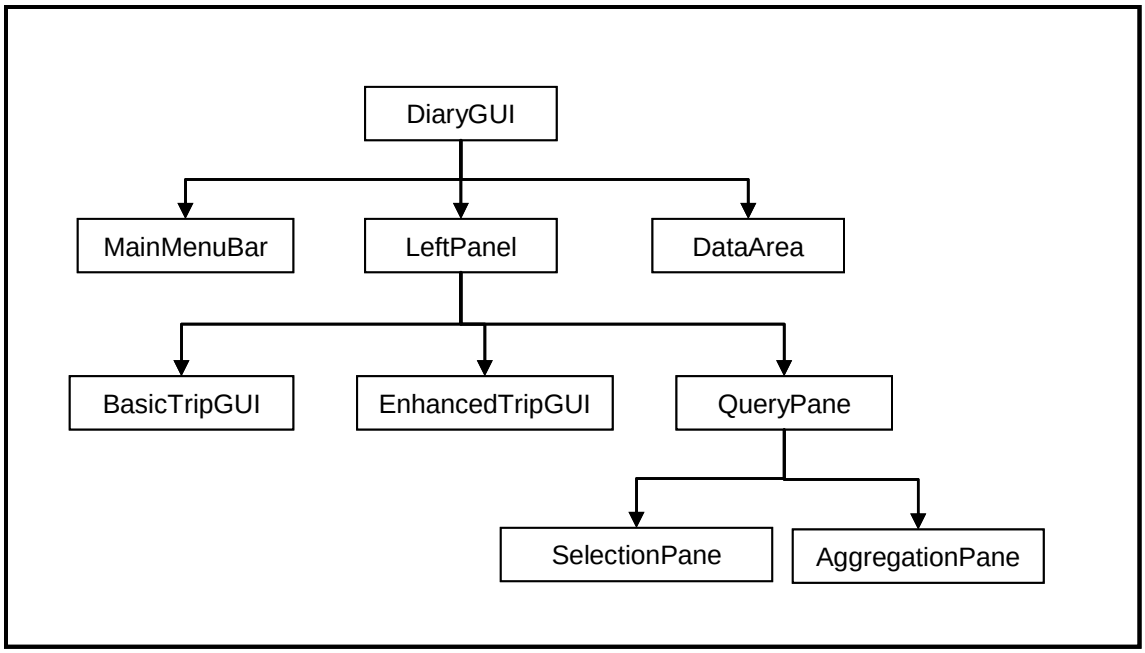

Figure 4.1 Schematic representation of the GUI implementation

The MainMenuBar represents the Menu Bar in the software. As already discussed, this has two menu items. These items have been implemented using standard Swing JMenus and JMenuItems in a JMenuBar.

The DataArea class represents the spreadsheet in the main window where all outputs generated are displayed. This is implemented using a standard Swing JScrollablePane containing a Swing JTable. The structure of the display table is determined based on the outputs being displayed (i.e, whether the outputs are the derived trips or query results).

The LeftPanel class represents the Command Area of the software. This class is implemented using a swing JTabbedPane that has three panels—the BasicTripGUI, the EnhancedTripGUI, and the QueryPane.

The BasicTripGUI class represents the Basic Pane and the EnhancedTripGUI class represents the Enhanced Pane of the Command Area (as discussed in Section 4.4.1.2). The latter class is derived from BasicTripGUI and has extra members corresponding to the additional inputs required for Enhanced analysis. The QueryPane class represents the third component of the Command Area, i.e., the query pane. The QueryPane, in turn contains two classes, the SelectionPane and the AggregationPane. These are used, respectively, for setting the selection

(or filtering) and aggregation criteria. The QueryPane has been implemented using elements such as Swing JCheckBox, JRadioButton, and JTextArea.

CHAPTER 5. TESTING AND VALIDATION

This chapter describes exercises undertaken to test and validate the GPSTDG software and thereby evaluate the ability of GPS-TDG in automating trip-diary extraction. Broadly, we examined the GPS-TDG outputs for reasonableness and also compared the derived trip diaries with corresponding data from self-reported travel surveys and trips derived by GeoStats using their own procedures. This chapter is organized as follows: Section 1 describes the data used in validation. Section 2 presents the validation framework and describes the measures used for comparative analysis. Section 3 discusses statistical comparison results.

5.1 Data

Data from the recently conducted household travel surveys from Laredo and Tyler-Longview were used for development, testing, and refinement of the GPS-TDG software. The self-reported survey component of both these surveys was administered using the traditional CATI procedures. In each case, a subset of households was recruited for the GPS-based component of the survey. Thus these surveys provide both passively recorded and self-reported travel data for several households. The GPS and reported travel data used for testing were drawn from these households. Specifically, we identified 45 vehicles (38 households) from Laredo and 92 vehicles (86 households) from Tyler-Longview, which provided all required information for our analysis. It is useful to note here that all these vehicles were switched-powered systems. As already discussed before, GPS-TDG is capable of handling continuous powered systems also. However, an examination of our prototype software's performance in this regard is identified as an area for future research.

In addition to the above-described GPS and self-reported travel data, we were also provided with files (one for each of the two surveys) containing trip-ends identified from the GPS streams using methods developed by GeoStats. These files provide only the start and end times of trips and no other trip-related attributes (such as purpose, distance, and trip-end location). The procedures used to identify the trips are proprietary and hence are not available to us. We compared the number of trips and trip-timings from GPS-TDG to those from the GeoStats procedures as another means of validation.

The procedure for assembling the data for performing the validation analysis involved the following steps. (These steps were performed for each of the Laredo and Tyler-Longview surveys.) First, the inputs to the GPS-TDG software were assembled from the raw GPS streams, the demographic information in the reported travel surveys, and the GIS TAZ-boundaries files. Next, the software was run to generate the derived trip files. Third, the self-reported travel information was restructured into a trip-file format that is consistent with the GPS-TDG output. Finally, GeoStats-identified trip-end information for the corresponding vehicles were extracted. All these data form inputs for computing validation statistics are identified in the next section.

5.2 Validation Framework

The framework developed for validation is described in this section. Broadly, validation was undertaken at two levels: (1) manual, ad-hoc comparisons and (2) statistical comparisons. Each of these is discussed further here.

5.2.1 Manual, ad-hoc comparisons

The ad-hoc comparisons involve detailed, manual checks of GPS-TDG outputs against reported trips in surveys and trip-ends provided by GeoStats. These analyses are performed at the trip-level. In the comparisons with the GeoStats results, the presence of a corresponding trip and its timing are examined. In the comparisons with the reported survey data, additional attributes such as trip-end purpose are also examined. The software performance at a disaggregate-level for varying input and travel types can be evaluated by undertaking such comparisons for vehicles from a wide variety of households.

5.2.2 Statistical comparisons

The next type of validation focused on examining performance at an aggregate level. Broadly, this involves comparing the derived trip attributes (such as number of trips and trip duration) for a set of vehicles with the corresponding observed or GeoStats-provided values. Two statistical measures are used for these comparisons—(1) the Mean Absolute Percentage Error (MAPE) and (2) the Root Mean Squared Error (RMSE), both of which are discussed below.

The Mean Absolute Percentage Error (MAPE) measure:

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\frac{|GPSTDG_n - TRUE_n|}{TRUE_n}$$

where,

$GPSTDG_n$ is the attribute derived by the GPS-TDG software for any vehicle n (e.g., number of trips or average trip length).

$TRUE_n$ is the corresponding attribute from either the self-reported survey or provided by GeoStats.

N is the number of vehicles in the validation sample.

The Root Mean Squared Error (RMSE) measure:

$$RMSE = \sqrt{\sum_{n=1}^{N}\frac{(GPSTDG_n - TRUE_n)^2}{N}}$$

where,

$GPSTDG_n$ is the attribute derived by the GPS-TDG software for any vehicle n (e.g. number of trips or average trip length).

$TRUE_n$ is the corresponding attribute from either the self-reported survey or provided by GeoStats.

N is the number of vehicles in the validation sample.

5.3 Results

In this section, we present results of the statistical validation of the GPS-TDG software. Several ad hoc checks were also undertaken manually (as discussed in 5.2.1) during the course of software development and refinement. As these are case specific, we are not presenting the details of these checks here. Results of the ad hoc checks, as well as the statistical tests, were used to determine our default algorithm parameters.

Prior to the discussion of statistical comparison results, it is useful to note that neither the self-reported travel surveys nor the GeoStats data can be completely construed as the "true" travel patterns (yet, we are comparing our software results to these data for some kind of limited

validation). Further, there is evidence of significant under-reporting in conventional travel surveys. Therefore, the validity of trips detected by the software but not reported cannot be inferred. The difference between the derived trips and reported trips seems to fall largely in this category (i.e., detected from GPS streams but not reported). The trips identified by GPS-TDG were, in general, found to match the trip-end data provided by GeoStats. However, it should be noted that the GeoStats trip-ends were also derived from GPS streams based on several assumptions and provide only the trip timings for comparison. In summary, additional data and research is required for a rigorous validation of algorithms for extracting travel information from GPS streams (see Chapter 6 for further discussions). In this context, it is useful to point out that GPS-TDG is a useful tool for conducting such rigorous tests. This is because the algorithm parameters can be varied using the GUI and the derived trip-diaries can be aggregated to generate summary measures using the GPS-TDG's querying capabilities.

The validation measures computed for Laredo and Tyler-Longview data are presented in Tables 5.1 and 5.2, respectively (computed using our prescribed default values of the algorithm parameters). Overall, GPS-TDG identified 262 trips for the 45 vehicles examined for Laredo. GeoStats provided 305 trips for the same set of vehicles and the primary drivers of these 45 vehicles reported undertaking 215 vehicle trips in the survey. For the Tyler-Longview data, GPS-TDG identified 545 trips for the 92 vehicles examined. GeoStats provided 582 trips for the same set of vehicles. The primary drivers of these 92 vehicles reported undertaking 534 vehicle trips in the survey.

The MAPE and RMSE measures were computed for two attributes, namely, the number of trips per vehicle and the average trip duration per vehicle. In general, values in these tables indicate a closer match of GPS-TDG trip counts and durations to the GeoStats derived trip attributes compared to the corresponding reported travel attributes.

Table 5.1 Validation measures for Laredo data

GPSTDG validated against	Number of Trips per Vehicle		Average Trip Duration per Vehicle (mins)	
	MAPE	RMSE	MAPE	RMSE
GeoStats data	28.5	2.7	32.1	3.83
Reported survey data	61.1	3.6	45.5	15.44

Table 5.2 Validation measures for Tyler-Longview data

GPSTDG validated against	Number of Trips per Vehicle		Average Trip Duration per Vehicle (mins)	
	MAPE	RMSE	MAPE	RMSE
GeoStats data	22.5	2.1	30.2	3.11
Reported survey data	53.2	3.2	41.78	13.61

Conceptually, it is straightforward to extend the computation of such measures to other trip attributes such as activity duration. Further, this analysis can also be performed at any desired level of aggregation. The conduct of such an extensive validation exercise is prescribed as an important next step to this research.

CHAPTER 6. SUMMARY

The use of GPS technologies is being increasingly sought after in order to increase the accuracy and completeness of travel surveys. The passive nature of data collection is extremely beneficial in reducing respondent burden and enhancing the quality of data. However, the data is collected in the form of navigational stream, which have to be processed to derive travel patterns in the trip-diary format. Consequently, the use of passive GPS technology in travel surveys shifts considerable burden from the respondent to the analyst. Therefore the success of GPS technology as a survey instrument depends on the ability of the analyst to derive the activity-travel information from the GPS streams.

Research Project 0-5176 funded by TxDOT focused on the development of a prototype software called the "GPS-Based Travel Diary Generator" (GPS-TDG) that automates the process of converting navigational data streams collected passively from in-vehicle GPS devices into an electronic travel diary. This derived travel diary comprises a sequence of vehicle trips identified from the GPS streams, with each trip characterized in terms of attributes such as trip-end location, trip purpose (or activity type at destination), time of day, duration, distance, and speed. The determination of the travel route is not within the scope of this project.

The software has been developed in the Java programming language using ArcGIS 9.0 as the platform for GIS processing. The software has been designed to operate either in a basic analysis mode or in an enhanced analysis mode. The basic mode converts the GPS data into a simple trip file that distinguishes among home-based work, home-based other, and non–home-based trips. The enhanced mode utilizes additional land-use data and pre-estimated model parameters to derive more refined trip purpose classification taxonomy.

The algorithm implemented within the GPS-TDG software is controlled by several parameters (such as the dwell-time thresholds), which can be easily modified by the analyst. Thus the software can be calibrated for any specific study region. Also provided are default values for all these parameters based on the testing and validation undertaken with available data.

Finally, the software is also capable of aggregating the derived trip diaries to produce inter-zonal vehicle trip tables and network performance measures (average trip speed, distance,

and travel time). These measures can be generated in the overall or for specific trip purposes or for specific times of the day.

Overall, the research project resulted in the development of flexible, user-friendly prototype software for processing GPS navigational streams. The research team prescribes (1) calibration and validation and (2) performance optimization as the most important and immediate next steps toward enhancing and refining this software for deployment in future TxDOT travel surveys. The issues related to these tasks are discussed further here.

Calibration and Validation

Calibrating and validating algorithms for extracting travel diaries from GPS navigational streams fundamentally require data on the "true" travel of the equipped vehicles. These "true" data are often available in the form of self-reported travel patterns of the household members from which the travel characteristics of the equipped vehicles may be inferred (assuming that information on the vehicle used for each person-trip is also recorded in the travel survey). The shortcoming of these data as a benchmark is that the validity of trips detected from the GPS streams but not reported cannot be ascertained. Therefore, rigorous validation requires CATI data where the extent of under-reporting is known to be minimal. The Laredo and Tyler-Longview data examined in this research appear to have significant under-reporting thereby limiting our validation efforts. The ability to match all (or most of) detected trips to corresponding reported trips is a necessary precursor to validating other trip attributes (such as purpose, trip duration, trip-end location, and activity type). Once the above is accomplished, the validation framework presented in Chapter 5 can be effectively used to quantify the accuracy of the algorithm in determining each of the derived attributes such as timing, duration, speed, and purpose.

The validation exercise will also substantially benefit with additional data collection. Specifically, well-designed test runs can be performed aimed at (1) fine-tuning procedures for handling signal-loss situations because of travel through urban canyons, (2) identifying GPS stream patterns that may help distinguish between short duration stops without engine off and signal delay, (3) developing algorithms for determining the trip timing more accurately (accounting for signal acquisition times), and (4) evaluating trip-distance and trip-speed

computation procedures using odometer readings and self-recorded times (note that trip distances and speeds are not collected in travel surveys).

Performance Optimization

The second major task is to optimize the software for performance both in terms of memory and speed. As already discussed in this report, the current version of GPS-TDG detects trips based on both time-gaps and nonmovement. Thus, the software is applicable for use with both switched-power and continuous-power systems. However, GPS data streams from continuous powered systems are significantly larger because data are recorded every second regardless of vehicle movement. Thus efforts are required aimed at enhancing the computational efficiency of the software. Further, the software can be expected to be deployed to process data from several hundreds of vehicles. Hence evaluation and enhancement of the software performance under such large-input conditions is necessary.