

GIS-BASED EPISODE RECONSTRUCTION USING GPS DATA

GIS-BASED EPISODE RECONSTRUCTION USING GPS DATA
FOR ACTIVITY ANALYSIS AND ROUTE CHOICE MODELING

By RON DALUMPINES, B.Sc., DURP, DTP, M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

© Copyright by Ron Dalumpines, May 2014

DOCTOR OF PHILOSOPHY (2014)
(School of Geography and Earth Sciences)

McMaster University
Hamilton, Ontario

TITLE: GIS-based Episode Reconstruction Using GPS Data for Activity Analysis and
Route Choice Modeling

AUTHOR: Ron Dalumpines,

B.Sc. (University of the Philippines, 1999)

Diploma in Urban & Regional Planning (University of the Philippines, 2002)

Diploma in Transportation Planning (University of the Philippines, 2003)

M.Sc. (Int'l Inst. for Geo-Information Science & Earth Observation, 2008)

SUPERVISOR: Professor Darren M. Scott

NUMBER OF PAGES: xviii, 172

Abstract

Most transportation problems arise from individual travel decisions. In response, transportation researchers had been studying individual travel behavior – a growing trend that requires activity data at individual level. Global positioning systems (GPS) and geographical information systems (GIS) have been used to capture and process individual activity data, from determining activity locations to mapping routes to these locations. Potential applications of GPS data seem limitless *but* our tools and methods to make these data usable lags behind. In response to this need, this dissertation presents a GIS-based toolkit to automatically extract activity episodes from GPS data and derive information related to these episodes from additional data (e.g., road network, land use).

The major emphasis of this dissertation is the development of a toolkit for extracting information associated with movements of individuals from GPS data. To be effective, the toolkit has been developed around three design principles: transferability, modularity, and scalability. Two substantive chapters focus on selected components of the toolkit (map-matching, mode detection); another for the entire toolkit. Final substantive chapter demonstrates the toolkit’s potential by comparing route choice models of work and shop trips using inputs generated by the toolkit.

There are several tools and methods that capitalize on GPS data, developed within different problem domains. This dissertation contributes to that repository of tools and methods by presenting a suite of tools that can extract all possible information that can be

derived from GPS data. Unlike existing tools cited in the transportation literature, the toolkit has been designed to be complete (covers preprocessing up to extracting route attributes), and can work with GPS data alone or in combination with additional data. Moreover, this dissertation contributes to our understanding of route choice decisions for work and shop trips by looking into the combined effects of route attributes and individual characteristics.

Acknowledgements

I am grateful to my supervisor, Dr. Darren M. Scott, for the opportunity he has given me to pursue research at McMaster, and for his unwavering support that led to the completion of this dissertation. I am thankful to my co-supervisors, Dr. Pavlos Kanaroglou and Dr. Antonio Páez, for the constructive comments and suggestions that help improved the quality of this dissertation. I am also thankful to Dr. John Eyles, Dr. Niko Yiannakoulis, and Dr. Hanna Maoh for being part of my comprehensive exam committee. Drs. Scott, Kanaroglou, Páez, and Yiannakoulis have been very supportive not only in academic matters, but also in giving personal and professional advice.

I would like to acknowledge Dr. Kay Axhausen and Dr. Nadine Schuessler for sharing their code on data filtering and smoothing that inspired the development of the toolkit's preprocessing module presented in this dissertation. Thanks to David Wynne of ESRI for his point-to-line script, which was incorporated in the early versions of the GIS-based map-matching algorithm. Also, thanks to the feedback of Dr. Marius Thériault in a conference that led to the improvement of the map-matching algorithm.

Thank you to all the Pythonistas, particularly to Guido van Rossum for making Python[®] available to everyone, ArcGIS[®] script contributors, and software developers around the world for sharing their code and insights.

I would like to thank my batchmates: Jia, Xudong, Monir, Tufayel, Jeff, and Nay for their help during my early years at Mac. I am also lucky to meet very accommodating

and helpful seniors: Dr. Dominik Papinski, Dr. Steven Farber, and Dr. Andrew Clark. Dr. Farber has been very helpful in sharing his academic experience, and selfless in giving professional advice. I am also thankful for all the support of my labmates: Ben, Randy, Charles, and Justin, and to all my friends and inquisitive GIS lab students at SGES.

I am thankful to Dr. Michael Goodchild for promptly responding to my email and openly shared his academic writing experience, and to Dr. Arthur Getis for sharing his insights on the pursuits of academic career.

I would like to thank Dr. Pavlos Kanaroglou, Dr. Mark Ferguson, and Deane Maynard for granting me the opportunity to work at McMaster Institute for Transportation and Logistics (MITL), which provided additional financial support for my family. Also, I am thankful for the support and encouragement from Dr. Matthias Sweet, Tom Lavery, and Anita Toth at MITL.

One could not forget the many nice faces at SGES office. Although I could not remember all their names (some work on part-time basis, others retired or moved to another job), I would like to express my utmost thanks for the administrative assistance they provided. In particular, I would like to thank Ann Wallace, Sal Santos, and Kath Philp - I think SGES will not function well without you guys!

I would like also to acknowledge Pat De Luca and Laura Labate for their assistance in the form of GIS data support and MITL reports. Rick Hamilton has been very helpful in printing posters, sometimes at a very short notice. I am also thankful to

Cathy Moulder (previously with Mac Library) for the spatial datasets I used in this dissertation, and to all the staff at Mac Libraries for their excellent library services.

During my early years in Hamilton, I am privileged to meet a lot of people who had helped me in many ways. Jos and Juni Samuel were my first friends in Hamilton who treated me as a family. The families of Percy Fernando and Rudy Yap provided a nice environment for my stay. Through Rudy and Ann Yap, I became part of the HIBC family. I wanted to acknowledge all of HIBC here but the space is not enough. I am thankful to Fevel Toledo and Dr. Weng Monton for their encouragement. From abroad, I am grateful to the Martin family (Ate Ning, Kuya Lito, Clarisse), Nang Niknik, Auntie Eva Reccius, and my friends at ITC and ESRI for all their support.

I am very grateful to my wife, Delecia, and our three wonderful kids (Aiyana, Kearn, and Kaidden) who stood with me until the completion of this dissertation. Thank you very much guys for your understanding when I could not joined you at dinner frequently, when I have to sleep in the morning after working overnight to do scripting and debugging for several days in a row. I am also grateful to my family and relatives, especially to Manang Kiking and her two children, Dan-dan and Caesar, who supported my mother in my absence.

Finally, I would like to dedicate this dissertation to my mother, and in loving memories of my beloved father and eldest brother. And to the God Almighty, I am forever grateful for “He is before all things, and in Him all things hold together” (Colossians 1:17, NASB).

Table of Contents

Abstract.....	iii
Acknowledgements.....	v
List of Figures.....	xii
List of Tables.....	xiv
List of Abbreviations.....	xv
Preface.....	xvii
Chapter 1 Introduction.....	1
1.1 Justification of research topic.....	1
1.1.1 Reconstructing activity episodes from GPS data for activity analysis.....	3
1.1.2 Generating route choice sets from GPS data for route choice modeling.....	8
1.2 Research objectives.....	9
1.3 Dissertation contents.....	12
1.4 References.....	16
Chapter 2 GIS-based Map-matching: Development and Demonstration of a Postprocessing Map-matching Algorithm for Transportation Research.....	22
2.1 Introduction.....	22
2.2 Introducing GIS platform for postprocessing map-matching.....	25
2.2.1 Postprocessing map-matching merits a different approach.....	26
2.2.2 Premise to the use of the GIS platform for map-matching.....	27
2.3 Constraints and limitations of existing algorithms.....	28
2.3.1 Problem with the initial map-matching process.....	28
2.3.2 Calibration of threshold values.....	29
2.3.3 Problems at complex intersections.....	29

2.4 GIS-based map-matching algorithm.....	31
2.5 Results.....	35
2.5.1 Data input and preprocessing.....	35
2.5.2 Accuracy.....	37
2.5.3 Computing speed.....	40
2.5.4 Advantages and limitations.....	42
2.6 Conclusion.....	44
2.7 Acknowledgements.....	46
2.8 References.....	46

Chapter 3 Making Mode Detection Transferable: Extracting Activity and Travel Episodes from GPS Data Using the Multinomial Logit Model and Python.....

Chapter 3 Making Mode Detection Transferable: Extracting Activity and Travel Episodes from GPS Data Using the Multinomial Logit Model and Python.....	51
3.1 Introduction.....	51
3.2 Data and methods.....	56
3.2.1 GPS and time-use diary (TUD) data.....	56
3.2.2 The proposed method: GERT’s Extraction and Mode Detection Module (MDM).....	58
3.2.3 Multinomial logit (MNL) model as classifier.....	64
3.3 Results and discussion.....	66
3.3.1 Classification of activity episodes.....	67
3.3.2 Accuracy.....	72
3.3.3 Performance.....	75
3.4 Conclusion.....	76
3.5 Acknowledgements.....	78
3.6 References.....	78

Chapter 4 GIS-based Episode Reconstruction Toolkit (GERT): A

Transferable, Modular, and Scalable Framework for Automated

Extraction of Activity Episodes from GPS Data.....	84
4.1 Introduction.....	84
4.1.1 Challenges in developing tools and methods for extracting episodes from GPS data.....	86
4.1.1.1 Lack of transferability.....	87
4.1.1.2 Processing demands of huge GPS data.....	88
4.1.1.3 Incomplete set of tools.....	89
4.1.2 Addressing challenges through GIS-based episode reconstruction toolkit (GERT).....	90
4.2 GIS-based episode reconstruction toolkit (GERT).....	91
4.2.1 GERT’s framework: transferability, modularity, and scalability.....	91
4.2.2 GERT’s modules: from GPS data preprocessing to route choice data generation.....	94
4.3 Data and experimental design for validation.....	99
4.3.1 STAR dataset.....	99
4.3.2 Experimental design for validation.....	101
4.4 Results and discussion.....	103
4.4.1 Comparison of TUD and GPS episodes.....	103
4.4.2 Computational performance.....	108
4.5 Conclusion.....	110
4.6 Acknowledgements.....	113
4.7 References.....	114

Chapter 5 Determinants of Route Choice: A Comparison of Shop versus Work Trips Using the Potential Path Area - Gateway (PPAG) Algorithm and Path-Size Logit.....	119
5.1 Introduction.....	119
5.2 Data and methods.....	124
5.2.1 Space-Time Activity Research (STAR) data.....	124
5.2.2 Path generation using the Potential Path Area - Gateway (PPAG) algorithm.....	127
5.2.3 Specification of the Path-Size Logit (PSL) model.....	129
5.2.4 Scale factor estimation and likelihood ratio tests.....	131
5.3 Results and discussion.....	135
5.3.1 Descriptive statistics of work and shop routes generated by the PPAG algorithm.....	135
5.3.2 Route choice behavior for shop trips versus that for work trips.....	138
5.3.3 Statistical test of utility parameters and scale equality.....	141
5.3.4 Difference between shop and work route choice highlighted by interaction variables.....	142
5.4 Conclusion.....	147
5.5 Acknowledgements.....	150
5.6 References.....	150
Chapter 6 Conclusion.....	156
6.1 Contributions to activity analysis and route choice modeling.....	157
6.2 Practical and theoretical implications.....	161
6.3 Directions for future research.....	164
6.4 References.....	168

List of Figures

Figure 2.1	Sample GUI of the GIS-based map-matching algorithm showing the input data and the buffer distance parameter.....	32
Figure 2.2	Portion of the attribute table generated by the GIS-based map-matching algorithm showing the important attributes for each observed route (i.e., travel time (minutes), travel distance (meters), and the number of left and right turns) relevant to route choice modeling.....	34
Figure 2.3	An example of a complex intersection where the GIS-based map-matching algorithm accurately generated the route for the GPS trace....	39
Figure 2.4	The algorithm sticks to the nearest road whenever some network link is missing.....	40
Figure 2.5	Plot of the computation time over route length for a sample of 104 routes.....	41
Figure 2.6	Sensitivity of map-matching algorithm to buffer distance: (a) buffer distance = 50 m and (b) buffer distance = 60 m.....	43
Figure 3.1	GPS trajectory subdivided into points, segments, and episodes.....	52
Figure 3.2	Four-stage workflow of GIS-based episode reconstruction toolkit (GERT), the proposed method of extracting and classifying episodes from GPS data is highlighted in bold.....	60
Figure 3.3	Graphs of GPS episodes in terms of predicted probabilities when varying one of the explanatory variables while setting the rest at their mean values (sample of 16 cases). The effects of varying median speed to predicted probabilities are shown in (a), median change in heading in (b), and total duration in (c).....	73

Figure 4.1	GPS trajectory subdivided into points, segments, and episodes.....	86
Figure 4.2	Four-stage workflow of the GIS-based episode reconstruction toolkit (GERT).....	92
Figure 4.3	Example of GERT’s CSV and shapefile outputs.....	100
Figure 4.4	Episode distribution, TUD versus GPS: (a) by episode type, (b) by episode ID (sequence or order of episode in a person-day; for example, at episode ID 5 the figure shows that GERT generated more 5 th episodes than reported in TUD), and (c) by duration (minutes).....	104
Figure 4.5	Duration distribution per episode type, TUD versus GPS: (a) stop episodes, (b) car episodes, (c) walk episodes, (d) bus episodes, and (e) other episodes.....	106
Figure 4.6	Distribution of travel episodes: (a) daily travel episodes (trips), TUD versus GPS; and (b) GPS-derived distances by travel mode.....	107
Figure 5.1	Route distance distribution for work and shop routes.....	137

List of Tables

Table 3.1	An example of statistical descriptors used as MNL predictors.....	63
Table 3.2	Average values of selected statistical descriptors by episode types.....	64
Table 3.3	Model estimation results for classification of GPS episodes.....	68
Table 3.4	Percent change in odds to assess ability of selected predictors in classifying GPS episodes.....	69
Table 3.5	Classification table of observed (TUD) versus predicted (GPS) episodes.....	74
Table 4.1	Average computational performance of GERT's main modules.....	109
Table 5.1	Individual characteristics used with route attributes to create interaction terms.....	126
Table 5.2	Selected route attribute statistics for work routes and shop routes.....	136
Table 5.3	PSL estimates for separate and combined datasets: work versus shop trips.....	139
Table 5.4	PSL estimates for shop and work, with interaction terms.....	143

List of Abbreviations

ABMS	Agent-Based Modeling and Simulation
ALIM	Activity Locations Identification Module
ALXM	Activity Locations (Stops) Extraction Module
CAD	Canadian Dollar
CF	Commonality Factor
CPU	Central Processing Unit
CSGM	Choice Set Generator Module
CSV	Comma-Separated Values
DMTI	Desktop Mapping Technologies Inc.
GB	Gigabyte
GERT	GIS-based Episode Reconstruction Toolkit
GEV	Generalized Extreme Value
GIS	Geographic Information System
GMM	GIS-based Map-Matching
GPM	GPS Preprocessing Module
GPS	Global Positioning System
GSP	Gateway Shortest Path
GUI	Graphical User Interface
HDOP	Horizontal Dilution of Precision
IIA	Independence from Irrelevant Alternatives
IID	Independently Identically Distributed
LNPS	Natural Logarithm of Path Size
MDM	GPS Episodes Extraction and Mode Detection Module
MNL	Multinomial Logit Model
MTP	Mode Transfer Point
PAL	Potential Activity Locations

PC	Personal Computer
PPA	Potential Path Area
PPAG	Potential Path Area - Gateway
PS	Path Size
PSCR	Point-Segment Classification Routines
PSL	Path-Size Logit
RAM	Random Access Memory
RCA	Route Choice Analysis
RVGM	RCA Variables Generator Module
SHP	Shapefile
ST	Shop Trips
STAR	Space-Time Activity Research
TGEM	TUD-GPS Trip Segments Extraction Module
TSEM	Trip Segments Extraction Module
TSP	Traveling Salesman Problem
TUD	Time-Use Diary
VIF	Variance Inflation Factor
WT	Work Trips

Preface

This dissertation is presented as a compendium of four substantive chapters either accepted, submitted, or in preparation for peer-reviewed publications. For this reason, there is some degree of repetition among the substantive chapters, particularly in the description of common toolkit modules, datasets, and illustrations. While the substantive chapters have been co-authored with the research supervisor, the content of each chapter was the sole responsibility of the dissertation author. This includes summary of the relevant literature, data processing and organization, script programming and debugging, specification and estimation of statistical models, and interpretation of results. The supervisor's contribution included critical appraisal of manuscripts prior to journal submission, editorial advice concerning articles yet to be submitted for publication, and discussion of empirical results and future research. These substantive chapters are as follows:

Chapter 2:

Dalumpines, R., Scott, D.M., 2011. GIS-based map-matching: development and demonstration of a postprocessing map-matching algorithm for transportation research. In S. Geertman, W. Reinhardt & F. Toppen (Eds.), *Advancing Geoinformation Science for a Changing World* (Vol. 1, pp. 101-120): Springer Berlin Heidelberg.

Chapter 3:

Dalumpines, R., Scott, D.M., 2014. Making mode detection transferable: extracting activity and travel episodes from GPS data using multinomial logit model and Python. *Submitted to Transportation Research Part C: Emerging Technologies*.

Chapter 4:

Dalumpines, R., Scott, D.M., 2014. GIS-based episode reconstruction toolkit (GERT): a transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data. *Submitted to Travel Behaviour and Society*.

Chapter 5:

Dalumpines, R., Scott, D.M., 2013. Determinants of route choice behavior: a comparison of shop versus work trips using the potential path area - gateway (PPAG) algorithm and path-size logit. *Submitted to Journal of Choice Modelling*.

Chapter 1

Introduction

1.1 Justification of research topic

Several procedures have been developed to extract information from person-based global positioning system (GPS) data primarily to supplement data from recall-based surveys (e.g., Chung and Shalaby, 2005; Stopher et al., 2005; Schuessler and Axhausen, 2009). However, most of these procedures suffer from specific data requirements and complexity that limit their transferability to other application environments. Further, they have a limited set of modules to extract all necessary information (e.g., automatic extraction of route attributes), and are not specifically designed to handle huge GPS datasets. To deal effectively with these problems, this dissertation presents a framework based on three design principles (transferability, modularity, and scalability), and introduces the geographic information system (GIS)-based episode reconstruction toolkit (GERT) based on this framework, for automated extraction of activity episodes from GPS data.

The problems mentioned earlier, which relate to GERT's three design principles, have not been explicitly addressed in the development of tools and methods in the past. In general, this dissertation argues for the importance of a framework that guides the development of an integrated set of practical tools. This framework was applied in the development of GERT. Without an effective framework and a toolkit to implement this

framework, GPS data cannot be fully utilized for the following reasons: (1) it is difficult to adopt tools developed by other researchers for lack of transferability, (2) there exists limited ability to derive more information from GPS data for lack of an integrated set of modules, and (3) high computational costs and lack of automatic procedures in dealing with huge datasets. These three key issues tend to hinder progress in the development of tools and methods in processing GPS data to support activity analysis in general and route choice modeling in particular. Data limitations are among the factors that have hindered research progress in activity analysis (Kitamura, 1988; Jones et al., 1990; Ortúzar & Olszewski, 2009) and route choice modeling (Prato, 2009).

This dissertation contributes to the repository of tools and methods that extract information from GPS data, and subsequently contributes to the advancement of travel behavior research at the individual level, by introducing a framework for toolkit development and demonstrating the potential of GERT (based on this framework) to automatically generate inputs useful for activity analysis and route choice modeling.

This introductory chapter presents the research context that motivates this dissertation (Sections 1.1.1 and 1.1.2), and the four research objectives that deal with the development and demonstration of GERT's key components (Section 1.2). Finally, this chapter also presents an overview of the contents of this dissertation (Section 1.3), which highlights the links between the four objectives and the four substantive chapters, and provides a brief summary of the findings and study implications.

1.1.1 Reconstructing activity episodes from GPS data for activity analysis

The availability of datasets pertaining to stationary activity and travel episodes plays an important role in activity analysis. Pas (1997, p. 96) envisioned that these datasets “...will very likely stimulate and facilitate continuing research and development of activity-based travel models...”; more so in the age of GPS technology that peoples’ movement can be captured in greater detail and higher frequency, potentially creating a wealth of data useful for activity analysis. Activity analysis is defined as a “framework in which travel is analyzed as daily or multi-day patterns of behavior, related to and derived from differences in life styles and activity participation among the population” (Jones et al., 1990, p. 34). The early roots of activity analysis were attributed to the contributions to time geography (e.g., Hägerstrand, 1970), planning theory (e.g., Chapin, 1974), and psychology (e.g., Fried et al., 1977). As a more holistic view of travel, activity analysis¹ focuses on the complete understanding of travel behavior and the likely effects of transport-related policies on travel by primarily viewing travel as part of an activity-based framework (Jones et al., 1990). As such, activity analysis serves as a tool in the practice of transport planning and policy development; and supports transport policy decision making (Kitamura, 1988; Jones et al., 1990; Pendyala, 2009).

¹ Quite often, the term *activity analysis* is used interchangeably in the travel behavior literature as *activity approaches*, *activity-based analysis*, *activity-based travel demand analysis*, or *activity-based travel demand modeling* (e.g., Kitamura, 1988; Jones et al., 1990; Pendyala, 2009). The latter two terms refer to the specific application of activity analysis to travel demand forecasting. For consistency, activity analysis is used throughout this dissertation as an umbrella term that encompasses all aspects (e.g., theories, applications, methods) of the activity-based framework.

GPS data are naturally suited to the needs of the *common features* of activity analysis, such as the focus on sequences or patterns of behavior and detailed timing and duration of activity and travel episodes (Jones et al., 1990), since person-based GPS devices capture time, distance, and route information better than traditional recall-based surveys. Also, the use of GPS in activity/travel surveys reduces respondent burden and improves trip reporting (Wolf et al., 2003; Stopher & Shen, 2011). Moreover, GPS data collected over long periods of time create new research opportunities for the exploration of the dynamics of travel behavior (Gonzalez et al., 2008; Ortúzar & Olszewski, 2009). In understanding travel patterns, Jones et al. (1990) identified three ways that travel can be measured: the number of stops in a tour or chain, the duration of travel or activities, and the sequence of events. Efficient processing of GPS data basically provides the inputs for these travel measurements. Burnett and Hanson (1979, cited in Damm (1983), p. 7) provided a list of the dimensions of activity behavior with the highest theoretical plausibility:

1. timing (clock time and relative time)
2. duration
3. location (absolute location as given by coordinates and relative location such as distance from the last stop)
4. mode(s) to reach activity (including non-motorized)
5. frequency of participation
6. sequencing

7. flexibility or elasticity (how easily moved in time and space)
8. importance or priority
9. variety (over periods longer than one day)

Note that a large majority of the items in the list can be captured directly or derived from GPS data. However, efficient processing and extraction of these dimensions from large GPS datasets for the benefit of travel behavior research remains a challenge. At the 11th International Conference on Travel Behavior Research, Ortúzar and Olszewski (2009) called for further development of methods in automating the extraction of meaningful travel information from tracking data, building upon several proposed methods (Doherty et al., 2001; Asakura and Hato, 2004; Chung and Shalaby, 2005; Stopher et al., 2005; Schuessler & Axhausen, 2009).

Existing procedures in extracting or reconstructing episodes from GPS data can be categorized into several modules: preprocessing (data filtering and smoothing), extraction of episodes (stages or segments), mode detection (assignment of mode to travel episode), route detection (map-matching), and purpose detection. GPS data consist of a series of points with latitude, longitude, and time – trajectories that represent stationary activity and travel episodes over space and time. These data need cleaning to remove erroneous or invalid points (preprocessing) before they are categorized into stationary activity or travel episodes (segmentation). Travel episodes are assigned to the most likely travel modes (mode detection), and trajectories corresponding to travel episodes are matched to a digital road network (map-matching) to determine travel routes and extract route

attributes. Points corresponding to stationary activity episodes can be analyzed with land use and other additional data to determine the most likely activity types (purpose detection).

To the author's knowledge, original attempts to automate the extraction of episodes in the transportation literature (Chung and Shalaby, 2005; Stopher et al., 2005; Schuessler and Axhausen, 2009) suggest the lack of transferability of existing modules (e.g., non-generic variables used in preprocessing), an incomplete set of modules (e.g., no modules for purpose detection in two studies while not fully automatic for route detection), and only one study was specifically designed for large GPS data (Schuessler and Axhausen, 2009) (for related studies in other fields, see Biljecki (2010) and Bolbol et al. (2012)). Only a few studies (Schuessler and Axhausen, 2009; Bohte and Maat, 2009) specifically addressed the development of tools and methods in extracting travel episodes and trip purposes from large-scale GPS datasets.

Existing methods used unique inputs or variables to filter valid points, and extract and classify episodes. For example, some researchers (e.g., Wolf et al., 2000; Stopher et al., 2005; Chung and Shalaby, 2005) used the number of satellites, heading, and horizontal dilution of precision (HDOP) in a preprocessing module to remove outliers and invalid GPS points; in the absence of the above inputs, Schuessler and Axhausen (2009) instead used the known altitude of Switzerland to remove low quality or erroneous GPS points. Other researchers (e.g., Chung and Shalaby, 2005; Bohte and Maat, 2009; Gong et al., 2012) used proximity measures (e.g., distances to bus, subway, and railway stations)

to determine probable travel modes but threshold distances significantly vary among studies. Lawson et al. (2010) recognized the difficulty of directly comparing different approaches because of different data used in developing these methods, not to mention the different variables required by each approach.

In recent years, there has been an increase in large-scale GPS data used for travel episode (trip) extraction (e.g., Zheng et al., 2008; Schuessler and Axhausen, 2009; Bohte and Maat, 2009; Biljecki, 2010; Millward and Spinney, 2011). Hence manual procedures are no longer practical in dealing with huge GPS data that span millions of records. The availability of huge GPS data and the high potential to collect more make it necessary to come up with efficient procedures that can automatically extract information from these data.

From the perspective of activity analysis, most of the existing methods did not fully capture valuable information from GPS trajectories for these methods were focused more on mode detection, that is, the extraction of travel episodes and classifying these episodes into several types based on travel modes (Stopher et al., 2005; Schuessler and Axhausen, 2009; Gong et al., 2012). Hence no modules were specifically developed to extract information associated with activity locations (*stop* episodes), wherein more information can be extracted with the aid of additional data such as land use and potential activity locations (PAL), and information on observed routes (road attributes) connecting these locations. The current practice of extracting routes (map-matching) is a tedious process of tracing the routes manually from GPS trajectories in a GIS (e.g., Ramming,

2002; Papinski et al., 2009; Winters et al., 2010) or asking respondents directly through web questionnaires (e.g., Kaplan and Prato, 2012). While some researchers had incorporated route detection using map-matching routines (e.g., Chung and Shalaby, 2005; Tsui and Shalaby, 2006), the map-matching lacks integration with route choice set generation algorithms (e.g., Prato and Bekhor, 2006) to automatically generate alternative routes for route choice modeling. In addition, existing modules that automatically generate route attributes for map-matched routes (e.g., Papinski and Scott, 2011) are not integrated with route detection and route choice set generation modules.

To the author's knowledge, no toolkit has been developed that integrates all the above modules, and at the same time include extra modules that automatically generate route choice sets and route attributes from GPS trajectories as further discussed in Section 1.1.2.

1.1.2 Generating route choice sets from GPS data for route choice modeling

The lack of input data for route choice modeling adds to the difficulty in developing better route choice models as indicated by the limited number of observations used across several studies (Ramming, 2002; Zhang & Levinson, 2008; Bekhor & Prato, 2009). Nowadays, advances in geospatial technologies such as GIS and GPS, along with computer-aided surveys, provide more accurate and less costly route choice data (e.g., Frejinger & Bierlaire, 2007) than the tedious and expensive roadside/mail-back surveys used in the past (e.g., Ben-Akiva et al., 1984). However, these advances present a

computing challenge involving the automatic extraction of useful inputs from huge GPS datasets for route choice modeling.

To the author's knowledge, no toolkit has been developed that automates the preprocessing of raw GPS data up to the generation of route choice data inputs. In particular, there exists the lack of sound behavioral basis and computing efficiency among existing route choice set generation algorithms (Prato, 2009). These algorithms are specified based on the experience and knowledge of the analyst, which is not reflective of the actual behavior of drivers. Since these methods take into account the universal set or the complete list of possible route alternatives, computing for the subset of relevant route alternatives takes a huge amount of computing time. The potential path area (PPA) approach (Papinski, 2010), which constrains the universal set of route alternatives based on the activity spaces of drivers, has potential in addressing the two aforementioned issues. Hence this dissertation introduces a modified PPA-based path generation algorithm, which is incorporated in the proposed GIS-based episode reconstruction toolkit.

1.2 Research objectives

The main objective of this dissertation is to advance the current tools and methods for automatically generating information from GPS data to support travel behavior research at the individual level, specifically to provide inputs for activity analysis in general and route choice modeling in particular. This main objective entails the

development of the GIS-based episode reconstruction toolkit (GERT) to automatically extract activity episodes (includes travel episodes) from GPS data and derive information related to these episodes from additional data (e.g., road network, land use). GERT consists of several components (tools or modules) that address key issues identified in Section 1.1.1: lack of transferability of existing tools and methods, an incomplete set of tools to extract information from GPS data, and the inability of existing tools to deal with huge GPS datasets. Because of GERT's broad range of tools, this dissertation focuses on specific objectives that relate to the key components, which deal with the key issues identified earlier. To achieve the main objective, this dissertation deals with the four specific objectives as outlined below.

Objective 1: Develop a simple yet effective approach in matching GPS trajectories to the road network, an approach that can support an integrated set of tools for path generation and extraction of route attributes. This objective lays down one of the key components of the toolkit in addressing the lack of transferability among existing tools to maximize the potential of GPS data for transportation research. The lack of transferability is often associated with complexity (i.e., many assumptions and user-defined parameters) that inhibits the utility of tools dealing with GPS data. As GIS becomes prevalent among transportation researchers, it is best to use GIS as a platform in the development of a postprocessing map-matching algorithm. This algorithm is an integral component of the toolkit designed to manipulate the outputs of another key component as addressed in Objective 2.

Objective 2: Develop a transferable and efficient method of extracting and classifying activity episodes from GPS data, without additional information. This objective addresses the need to develop practical tools that automatically extract activity episodes (i.e., stationary activity and travel episodes) from GPS data in the *absence* of time use or travel diary data. Again, this objective emphasizes that the method should be transferable, that is, applicable to different GPS datasets collected from different locations without the need of an extensive diary survey. Because most of existing methods in classifying episodes (mode detection algorithms) require more assumptions and user-specified values, this objective suggests a minimalist approach. Hence, it calls for the development of an efficient mode detection algorithm as an integral component of the toolkit, which is further expanded to include more functionalities (components) to fully automate, if possible, the extraction of information from GPS data as addressed in Objective 3.

Objective 3: Design and develop an integrated set of tools that automatically extract activity episodes from GPS data and derived information related to these episodes from additional data, a toolkit that includes tools from GPS data preprocessing to route choice data generation. This objective highlights the importance of design principles (i.e., transferability, modularity, and scalability) that can be used to explicitly address the challenges identified in Section 1.1.1. These design principles serve as the framework in the development of independent but integrated components (tools), which, *taken together*, support the data needs of activity analysis in general and route choice modeling in

particular. The framework also serves to guide the adoption and modification of effective algorithms from the literature, and the development of new procedures to fill-in the gaps. Furthermore, this objective lays down the main components of the GIS-based episode reconstruction toolkit, and its application demonstrated as addressed in Objective 4.

Objective 4: Demonstrate the application of the GIS-based toolkit in generating inputs for route choice modeling, a modeling exercise that aims to test whether route choice preference for work trips differs from that of shop trips. This objective emphasizes the capability of the toolkit in automating the often tedious process of preparing data inputs for route choice modeling. Also, this objective highlights the flexible nature of the toolkit – an ability to automatically extract travel episodes (of different trip purposes) from GPS data, given the corresponding time use diary data.

In summary, Objectives 1 and 2 deal with the development and validation of the key components of the toolkit, while Objective 3 highlights all the toolkit's main components, including those components introduced in the first two objectives. Finally, Objective 4 provides a demonstration of the toolkit's ability in supporting the data needs of route choice modeling. The four objectives correspond to the four substantive chapters, which are outlined in Section 1.3.

1.3 Dissertation contents

The remainder of this dissertation is organized into four substantive chapters (either published or submitted for publication), and one concluding chapter on research

contributions and future directions. Corresponding to the four objectives, the four substantive chapters (papers) form a coherent substantial body of work that focuses on the advancement of tools and methods for extracting information from GPS data, which in turn supports the data needs of activity analysis in general and route choice modeling in particular.

Chapter 2 presents a GIS-based map-matching (GMM) algorithm that makes use of geometric, buffer, and network functions in a GIS – to illustrate the suitability of a GIS platform in developing a postprocessing map-matching algorithm for transportation research applications such as route choice analysis. Moreover, this chapter laid down one of the key components of GERT, a component that supports the extraction of segment-related attributes and other route generation procedures. The GMM algorithm was tested using a GPS-assisted time use survey that involved nearly 2,000 households in Halifax, Nova Scotia, Canada. Actual routes taken by household members who travelled to work by car were extracted using the GPS data and the GMM algorithm. The test results suggest that the GMM algorithm can be used as a practical tool in extracting routes based on GPS trajectories.

Chapter 3 presents a transferable and efficient method of extracting and classifying activity episodes from GPS data – an important component of GERT that provides inputs to GERT's subsequent modules *in the absence of time use or travel diary data*. The proposed method, developed using Python[®], introduces the use of the multinomial logit (MNL) model in classifying extracted episodes into different types:

stop, car, walk, bus, and other travel episodes. The proposed method is demonstrated using GPS data from the Space Time Activity Research (STAR) project in Halifax, Canada. The GPS data consisted of 5,127 person-days (about 47 million points) – complementing a time use diary data that provided 7,271 reported episodes that matched GPS episodes within five minutes of episode’s start or end time. The demonstration shows that the proposed method proved to be simple yet effective in extracting and classifying episodes from GPS data, and consequently suggests that the method has potential as a transferable and efficient alternative among mode detection algorithms.

Chapter 4 presents GERT’s framework based on three design principles (transferability, modularity, and scalability), and the entire set of GERT’s components (modules) including the two previously demonstrated in Chapters 2 and 3. This chapter puts emphasis on the linkages among components in the automated extraction of activity episodes from GPS data, and deriving additional information related to extracted episodes from additional data such as road network and land use. To validate that GERT works properly at the aggregate level (i.e., in terms of episode and duration distributions), time use diary (TUD) and GPS episodes were matched based on start/end time difference and total duration. Overall, the validation results indicate that GERT provides a transferable, modular, and scalable set of practical tools in automatically reconstructing episodes from GPS data, and potentially supports the data needs of activity analysis in general and route choice modeling in particular.

Chapter 5 presents a comparison of the route choice models for work and shop trips in order to test whether route choice decision processes differ by trip purpose. In the process, this chapter introduces a practical path generation method, called Potential Path Area - Gateway (PPAG) algorithm, which automatically generates route choice sets from GPS trajectories. Moreover, this chapter demonstrates the capability of GERT in generating route attributes automatically, given a road network dataset and GPS trajectories. The GERT-generated route choice sets were used as inputs to Path-Size Logit modeling (Ben-Akiva & Bierlaire, 1999), which is in turn used as the basis for the scaling estimation method and likelihood ratio test (Swait & Louviere, 1993) to check whether the utility and scale parameters are different for separate route choice models of work and shop trips. The results show that route choice preferences vary by trip purpose, and suggest that route choice behavior for work trips tend to be *restrictive* while *nonrestrictive* for shop trips. In addition, descriptive analysis of route choice sets and intuitive model results suggest the utility of GERT in generating inputs for route choice modeling, clearly reducing the burden often associated with reproducing actual routes taken by survey respondents.

Finally, Chapter 6 wraps up the four substantive chapters by summarizing the contributions of this dissertation to activity analysis and route choice modeling, and identifying the results from each substantive chapter that make substantial contributions to the literature. This chapter also discusses the practical implications of GERT and identifies possible directions for future research.

1.4 References

- Asakura, Y., & Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12(3–4), 273-291.
- Bekhor, S., & Prato, C. G. (2009). Methodological transferability in route choice modeling. *Transportation Research Part B: Methodological*, 43(4), 422-437.
- Ben-Akiva, M. E., Bergman, M. J., Daly, A. J., & Ramaswamy, R. (1984). Modeling interurban route choice behavior. In J. Volmuller & R. Hamerslag (Eds.), *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory: 11-13 July, 1984, Delft, The Netherlands* (pp. 299-330). Utrecht, The Netherlands: VNU Science Press.
- Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In R. Hall (Ed.), *Handbook of Transportation Science* (Vol. 23, pp. 5-33): Springer US.
- Biljecki, F. (2010). Automatic segmentation and classification of movement trajectories for transportation modes (Master's thesis). Delft University of Technology, Delft, The Netherlands.
- Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285-297.
- Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36(6), 526-537.

- Chapin, F. S. (1974). *Human activity patterns in the city: things people do in time and in space*. New York: Wiley.
- Chung, E.H., & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28(5), 381 - 401.
- Damm, D. (1983). Theory and empirical results: a comparison of recent activity-based research. In S. Carpenter & P. Jones (Eds.), *Recent advances in travel demand analysis* (pp. 3-33). Aldershot: Gower Publishing.
- Doherty, S. T., Noel, N., Gosselin, M. L., Sirois, C., & Ueno, M. (2001). Moving beyond observed outcomes: integrating global positioning systems and interactive computer-based travel behavior surveys (pp. 449-466): *Transportation Research Board*.
- Frejinger, E., & Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological*, 41(3), 363-378.
- Fried, M., Havens, J., & Thall, M. (1977). *Travel behaviour - a synthesized theory*. Final report to the National Cooperative Highway Research Program, Washington, DC.
- Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36(2), 131-139.
- Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782.
- Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science*, 24(1), 6-21.

- Jones, P., Koppelman, F. S., & Orfueil, J. P. (1990). Activity analysis: state-of-the-art and future directions. In P. Jones (Ed.), *Developments in dynamic and activity-based approaches to travel analysis* (pp. 34-55). Avebury: Gower Publishing.
- Kaplan, S., & Prato, C. G. (2012). Closing the gap between behavior and models in route choice: the role of spatiotemporal constraints and latent traits in choice set formation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 15(1), 9-24.
- Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation*, 15(1-2), 9-34.
- Lawson, C. T., Chen, C., & Gong, H. (2010). *Advanced applications of person-based GPS in an urban environment*. New York: New York University at Albany. Retrieved from http://www.utrc2.org/sites/default/files/pubs/advanced-applications-gps1-final_2.pdf
- Millward, H., & Spinney, J. (2011). Time use, travel behavior, and the rural-urban continuum: Results from the Halifax STAR Project. *Journal of Transport Geography*, 19(1), 51-58.
- Ortúzar, J. d. D., & Olszewski, P. (2009). Advances in data acquisition. In R. Kitamura, T. Yoshii & T. Yamamoto (Eds.), *The expanding sphere of travel behavior research: selected papers from the 11th Conference of the International Association for Travel Behavior Research* (pp. 447-455). Bingley, UK: Emerald Group Publishing.
- Papinski, D. (2010). *Investigating route choice decisions using GPS and prompted-recall diary data* (Doctoral dissertation). McMaster University, Hamilton, Ontario, Canada.

- Papinski, D., & Scott, D. M. (2011). A GIS-based toolkit for route choice analysis. *Journal of Transport Geography*, 19(3), 434-442.
- Papinski, D., Scott, D. M., & Doherty, S. T. (2009). Exploring the route choice decision-making process: a comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4), 347-358.
- Pas, E. I. (1997). Recent advances in activity-based travel demand modeling. In L. J. Engelke (Ed.), *Activity-based travel forecasting conference: summary, recommendations and compendium of papers* (pp. 79-102). Arlington, TX: Texas Transportation Institute.
- Pendyala, R. M. (2009). Challenges and opportunities in advancing activity-based approaches for travel demand analysis. In R. Kitamura, T. Yoshii & T. Yamamoto (Eds.), *The expanding sphere of travel behavior research: selected papers from the 11th Conference of the International Association for Travel Behavior Research* (pp. 303-335). Bingley, UK: Emerald Group Publishing.
- Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modeling*, 2(1), 65-100.
- Prato, C. G., & Bekhor, S. (2006). Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board*, 1985, 19-28.
- Ramming, M. S. (2002). *Network knowledge and route choice* (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, USA.

- Schuessler, N., & Axhausen, K. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105, 28-36.
- Stopher, P., Jiang, Q., & FitzGerald, C. (2005). Processing GPS data from travel surveys. Paper presented at the 2nd International Colloquium on Behavioral Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications, Toronto, Ontario, Canada.
- Stopher, P., & Shen, L. (2011). In-depth comparison of global positioning system and diary records. *Transportation Research Record: Journal of the Transportation Research Board*, 2246, 32-37.
- Swait, J., & Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, 30(3), 305-314.
- Tsui, S., & Shalaby, A. (2006). Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 38-45.
- Winters, M., Teschke, K., Grant, M., Setton, E., & Brauer, M. (2010). How far out of the way will we travel? *Transportation Research Record: Journal of the Transportation Research Board*, 2190, 1-10.
- Wolf, J. (2000). Using GPS data loggers to replace travel diaries in the collection of travel data (Doctoral dissertation). Georgia Institute of Technology, Atlanta.
- Wolf, J., Oliveira, M., & Thompson, M. (2003). Impact of underreporting on mileage and travel time estimates: results from global positioning system-enhanced household

travel survey. *Transportation Research Record: Journal of the Transportation Research Board*, 1854, 189-198.

Zhang, L., & Levinson, D. (2008). Determinants of route choice and value of traveler information: a field experiment. *Transportation Research Record: Journal of the Transportation Research Board*, 2086, 81-92.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the web. Paper presented at the 17th World Wide Web Conference, Beijing, China.

Chapter 2

GIS-based Map-matching: Development and Demonstration of a Postprocessing Map-matching Algorithm for Transportation Research

2.1 Introduction

The increasing popularity of Global Positioning Systems (GPS) inspires some renewed interests in travel behavior research. Person-based GPS devices are increasingly used in travel or time use surveys (Doherty, 2001; Murakami and Wagner, 1999; Ogle et al., 2002; Wolf et al., 1999; Casas and Arce, 1999; Yalamanchili et al., 1999; Draijer et al., 2000; Pearson, 2001; Wagner, 1997). Matching the GPS coordinates to the digital road network has become an accepted approach in determining the actual routes taken by travelers, thus improving travel behavior analysis by providing a more accurate account of observed routes (Chung and Shalaby, 2005; Marchal et al., 2005; Schuessler and Axhausen, 2009). This approach is commonly known as map-matching in the field of car navigation and transportation research. Map-matching is a method of tracing the path or route taken by a traveler (represented by a sequence of GPS points) relative to a digital road network map. The underlying issues in map-matching has been extensively explored within the broader field of geographic information science focusing on different applications: geographic integration (Devoegele, 2002; Harvey, 1994, 2005; Harvey and Vaughlin, 1996a, 1996b; Walter and Fritsch, 1999), similarity measures for feature/geographic data matching (Bel Hadj Ali, 1997; Lemarié and Raynal, 1996;

Vauglin and Bel Hadj Ali, 1998), spatiotemporal databases and moving objects data (Brakatsoulas et al., 2005; Cao and Wolfson, 2005).

Map-matching can be classified generally into real-time and postprocessing map-matching (Quddus et al., 2007). Real-time map-matching captures the location of a traveler in the road network with a real-time feed of GPS locations (often augmented by data from dead reckoning devices). Postprocessing map-matching takes GPS data recorded from a travel or time-use survey and matches it to the road network to trace the routes taken by travelers. This postprocessing procedure allows the integration of network attributes with the socio-economic information of travelers, providing data that can be used for analysis and model estimation.

Although most map-matching approaches use geometric and topological analysis — two common built-in functions in most GIS packages — nevertheless very limited studies have attempted to develop a postprocessing map-matching algorithm in a GIS platform because most published articles on map-matching focus on real-time navigation applications (Quddus et al., 2007), and the perceived slow performance of map-matching in a GIS (Schuessler and Axhausen, 2009). In the case where it is used for real-time map-matching, GIS use is limited only to visualize the map-matching results (Taylor et al., 2006). Few studies have been published on postprocessing map-matching algorithms specifically for transportation research. At least 35 articles on map-matching are claimed to be published for the period 1989-2006 (Quddus et al., 2007) although an extensive literature exists in geographic information science pertaining to similar concept but used

for different applications (e.g. Brakatsoulas et al., 2005; Cao and Wolfson, 2005). Yet in the context of transportation research, to the authors' knowledge, only two articles are written on postprocessing map-matching developed and implemented in a GIS platform (Chung and Shalaby, 2005; Zhou, 2005). The large majority of the articles focus on real-time map-matching algorithms generally developed for navigation purposes. This suggests the lack of research for postprocessing map-matching algorithms and the need for more of these tools for transportation research particularly in travel behavior analysis. This paper helps to fill this gap by introducing a postprocessing map-matching algorithm developed in a GIS platform to support travel behavior research in exploiting the increasing popularity of GPS in travel or time-use studies. Hence, this paper argues that a GIS is an ideal platform for the development of a postprocessing map-matching algorithm for transportation research.

The research presented in this paper is unique in two respects: technique and data input. Compared to previous map-matching algorithms, this is the first purely GIS-based map-matching algorithm for postprocessing person-based GPS data. Only two other published studies used GIS as a platform in developing postprocessing map-matching algorithms; however, they employed real-time map-matching procedures such as a reconfiguration of Greenfeld's (Greenfeld, 2002) weighted topological algorithm (Chung and Shalaby, 2005), and multiple-hypothesis testing matching with rank aggregation (Zhou, 2005). The algorithm presented in this paper utilizes mainly built-in functions in a GIS such as buffer analysis and route analysis tools. In terms of data input, the algorithm

uses the largest GPS-assisted time use survey undertaken to date (Bricka, 2008). This research is a novel attempt to extract observed routes for work trips using GPS data and time diaries (episode data file). This is different from the two related studies on postprocessing GPS data. The first used large GPS records without any additional information (Schuessler and Axhausen, 2009). The second had travel survey data but need to re-enact the trip data using a person-based GPS (Chung and Shalaby, 2005). Also, the proposed algorithm fully utilizes the network dataset from a private data provider (DMTI) that includes attribute information such as turn restrictions, one-way street information, road classification, road speed, etc. Such effective use of a network dataset in a GIS platform for postprocessing map-matching has not been done before.

The GIS-based map-matching algorithm generates the actual routes taken by respondents based on the GPS data and time diary. The actual routes (or observed routes) serve as the dependent variable in route choice modeling. Aside from the observed routes, the algorithm also generates a travel time, route distance, and number of left and right turns for each observed route — used as independent variables in route choice models.

2.2 Introducing GIS platform for postprocessing map-matching

A map-matching problem is characterized by two objectives: 1) identify the link traversed by the traveler, and 2) find his/her actual location within that link (Quddus et al., 2007; White et al., 2000). Postprocessing map-matching algorithms focus only on the first objective while real-time map-matching algorithms need to address the two

objectives. Postprocessing and real-time map-matching algorithms also differ in data inputs. Road network map and GPS data are often enough for postprocessing map-matching. Real-time map-matching requires other data (e.g., from dead reckoning devices, elevation models, etc.) usually to augment the inaccuracies of GPS in urban environments. The kind and nature of these data inputs and the purpose of the algorithm largely influence the development of the map-matching procedures. For example, postprocessing procedures can create a polyline feature from the entire series of GPS points, which are already available, and match this line to the road network (i.e., global map-matching procedure (Lou et al., 2009)). This is not possible in real-time map-matching because the map-matching needs to process the GPS coordinates as they are being updated online (i.e., incremental map-matching procedure).

2.2.1 Postprocessing map-matching merits a different approach

Adopting procedures originally developed for real-time map-matching to postprocessing map-matching restricts the search for more appropriate procedures specifically for postprocessing map-matching. For example, the shortest path algorithm has not been used for real-time map-matching but can be appropriately used for postprocessing map-matching applications. Zhou (2005) cites that the shortest path algorithm can be utilized for postprocessing map-matching algorithm but did not proceed on exploring the idea. Hence, the core element of the GIS-based postprocessing map-matching as proposed in this paper, which is the use of the shortest path algorithm, is not a new idea. But this idea, to the authors' knowledge, has not been fully explored

particularly in a time when advances in GIS platforms offer more flexibility and advanced functionality.

Furthermore, there should be a clear distinction as far as postprocessing map-matching is concerned. For this reason, the development of postprocessing map-matching algorithms should take a different approach from those of real-time map-matching. The dominance of the real-time map-matching procedures in the literature leads to the ongoing adoption of these procedures to postprocessing applications. Since real-time map-matching procedures have not established an affinity with GIS platforms, postprocessing map-matching procedures currently focus on developing procedures in non-GIS platforms.

2.2.2 Premise to the use of the GIS platform for map-matching

The platforms used for the development of the map-matching algorithms reveal the range of techniques that can be employed. GIS provides excellent data models and tools in dealing with spatial data. For example, ArcGIS[®] provides an advanced network data model that allows for the modeling of complex road layouts by taking into account road design parameters such as turn restrictions, road hierarchy, and impedances. This strength of GIS makes it an ideal platform in developing a postprocessing map-matching algorithm that fully integrates network topology and attributes to match streams of GPS points to the road network.

However, GIS packages are often proprietary, comprehensive, and platform-dependent. For these reasons, most map-matching procedures are developed and

implemented in non-GIS platforms. For example, Java is free, platform-independent and used in some postprocessing procedures (Marchal et al., 2005; Schuessler, and Axhausen, 2009). Even so, non-GIS platforms have limited capability in handling spatial data models such as road networks and thus rely on a planar network that consists of nodes and arcs (links). This paper provides evidence that a postprocessing map-matching algorithm that utilizes a GIS network data model is effective in integrating topological information and resolving the map-matching problems of complex road intersections.

2.3 Constraints and limitations of existing algorithms

The existing literature identifies the constraints and limitations of the current map-matching algorithms for transportation applications (Quddus et al., 2007; White et al., 2000). Although the literature review by Quddus et al. (2007) focuses on real-time map-matching algorithms, some of the major constraints and limitations they identified also apply to postprocessing map-matching algorithms. These constraints and limitations refer to problems associated with the identification of initial links, calibration of threshold values used in decision processes, and the difficulty in correctly matching locations in complex road layouts (e.g., cloverleaf interchanges, flyovers).

2.3.1 Problem with the initial map-matching process

One of the problems of existing map-matching algorithms is the identification of the initial link. The existing map-matching techniques use an error ellipse or circle to snap the GPS point/s to the junction (intersection) node to identify the initial link. The

problem occurs if the junction node falls outside this error region. Moreover, the entire length of the link is assumed to be traversed once an initial link is identified. This is problematic for trip ends covering only a portion of the road link because the travel distance will be overestimated if computed based on all the links traversed. This problem is avoided when using the GIS-based map-matching algorithm. This algorithm snaps the initial GPS point to the nearest link instead of the junction node, and the route length is calculated from the portion of the link covered by the GPS trajectory.

2.3.2 Calibration of threshold values

All of the existing map-matching algorithms use some parameters. For example, a postprocessing map-matching algorithm developed by Marchal et al. (2005) depends on two parameters, N (number of candidate paths) and α (u-turn parameter). The number of parameters increases as the map-matching algorithm becomes complicated. Often it is difficult to recommend default values and this becomes an issue when applying the map-matching technique to a different operational environment (Quddus et al., 2007). A map-matching algorithm that uses a minimum number of parameters that can be calibrated easily will be helpful to transportation researchers.

2.3.3 Problems at complex intersections

Development of map-matching algorithms should effectively address the problems that arise at intersections. Route changes occur at intersections making it difficult for map-matching processes to identify the next link (White et al., 2000). This

difficulty is more pronounced in complex intersections (e.g., cloverleaf interchanges, flyovers) and further exacerbated by GPS errors. For this reason, the existing literature repeatedly suggests the integration of network topology in map-matching procedures (Marchal et al., 2005; Quddus et al., 2007; White et al., 2000; Quddus et al., 2003). Most of the existing algorithms did not go beyond the simple connectivity rules, often incorporated in some scoring procedures (or set of rules) to determine the next link after the intersection. Hence, Quddus et al. (2007) recommends the use of road design parameters (e.g., turn restrictions, road classification, etc.) and Marchal et al. (2005) hinted at the use of turn rules to improve map-matching performance particularly at intersections.

To the authors' knowledge, no map-matching algorithm has taken full advantage of these road attributes. Quddus et al. (2007) attributed this to unavailability of data but it can be argued that the standard network data model (planar network model) limits the inclusion of road attributes. Spatial road network data are available from governments for free and private providers sell more comprehensive data at a reasonable cost. Private data vendors provide route logistics or road network data in GIS formats (e.g., ArcGIS[®], MapInfo[®]), which the existing map-matching methods have not taken full advantage.

Therefore, this paper argues that a GIS platform should be used in developing a postprocessing map-matching algorithm to utilize the network topology and road attributes that can be handled easily in a GIS environment. The next section describes the GIS-based postprocessing map-matching algorithm followed by the testing results using

the Halifax STAR Project dataset (focusing on routes taken by 104 individuals during their drive to work in the morning).

2.4 GIS-based map-matching algorithm

The core of the GIS-based map-matching algorithm is the use of a route analysis tool in ArcGIS® (Network Analyst extension). This tool uses a shortest path algorithm and basic inputs (stops, barriers) to generate the shortest path or route. Stops and barriers are the basic parameters that need to be set by the user. Stops refer to the origin and destination locations (i.e. trip origin and destination) used by the shortest path algorithm to generate the best or shortest route. Barriers play a significant role in controlling the shortest path algorithm to generate only the route based on the streams of GPS points that represent a trip. The algorithm creates a buffer region around GPS trajectories to produce a set of barriers that control the route analysis tool to correctly generate the observed routes. These routes are automatically stored in a file geodatabase feature class format that contains relevant attributes for the route choice analysis (e.g., travel distance, travel time, number of left and right turns). These attributes are automatically added by the algorithm to every route generated. The route generated depends on the impedance or cost defined by the user. Travel time is the default impedance used by the algorithm but the user can change it to travel distance if desired.

The postprocessing GIS-based algorithm is developed and implemented in ArcGIS® v9.3.1 using Python scripting language. Python scripting is free and well

supported in ArcGIS® and works well with ArcObjects™ - the building blocks of ArcGIS® software. The algorithm can be run as a standalone program or added as a tool in ArcGIS®. The standalone implementation saves some processing overhead and hence it has computational speed advantage. The latter approach provides a user-friendly GUI, allowing users to specify the input data and parameters. To run the algorithm via GUI (Figure 1), the user specifies the following parameters: the workspace location of the file geodatabase containing the GPS data, a sample GPS data file from the file geodatabase (for the script to read the attribute fields of the GPS data file), line field and sort fields used to convert the GPS points into a polyline feature, the network dataset, and the buffer distance in meters (50 m is the default value).

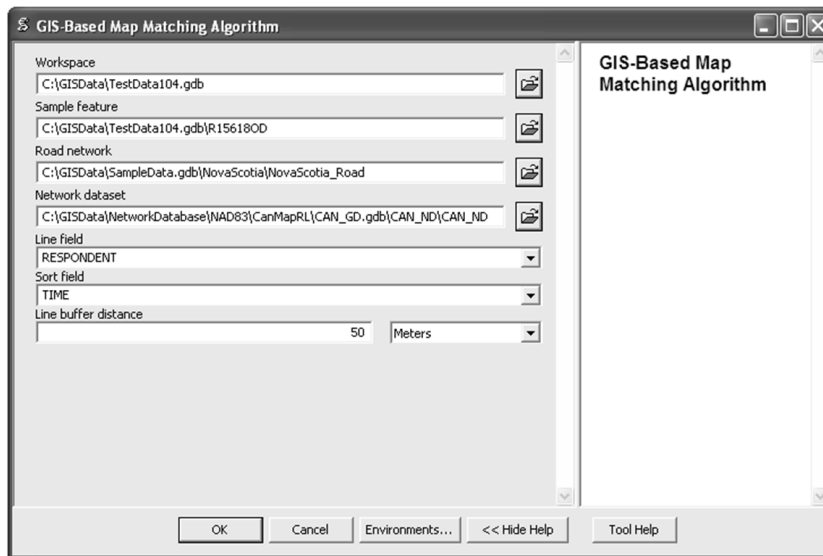


Figure 2.1 Sample GUI of the GIS-based map-matching algorithm showing the input data and the buffer distance parameter

Python scripting is used to automate the detailed steps for the GIS-based map-matching, the steps are described as follows:

1. Convert the stream of GPS points (representing the trip made by a traveler) into a polyline feature. The first and the last GPS points in the sequence are designated as stops (origin and destination points) in the network analyst module in ArcGIS®. The intermediate points between the stops are used to generate the polyline feature; this feature is the basis for the buffer in step 2.
2. Create a buffer around the polyline feature based on user-defined distance. The buffer distance should be, more or less, 5x to 6x the horizontal accuracy of the GPS data. This is based on the results of the sensitivity analysis, which is explained further in the next section on results (section 5.4). The experiment for GPS data with a horizontal accuracy of 10 m revealed that a buffer of 50 m produces accurate results. This was set as the default distance in the algorithm but can be changed by the user.
3. Assign the stops and barriers for the route solver (from ArcGIS® Network Analyst module). Start and end points define the stops. (The route solver can also work with multiple stops in between the start and end points, similar to the Traveling Salesman Problem (TSP)). Barriers are defined by the intersection of the boundary of the buffer region created in the previous step and surrounding links. Barriers ensure the accuracy and efficiency of the shortest path algorithm. This step

assumes that there are no errors particularly gaps in the GPS data. Outliers and other errors are handled by the GPS data preprocessing module.

4. The observed route is generated (or not generated) depending on the buffer distance specified by the user. This route is the shortest or best route generated by the shortest path algorithm (using the origin and destination points) inside the buffer region. Topological rules and road attributes (e.g. one-way restrictions, road hierarchy, etc.) are used by the built-in shortest path algorithm in ArcGIS® in generating the shortest path between origin and destination points.
5. The network attribute table is updated for the number of left and right turns in traversing the observed route, aside from the travel distance and travel time that are automatically generated by the route solver (Figure 2).

ROUTE #	RESPONDENT ID	TRAVEL TIME (min)	TRAVEL DISTANCE (m)	LEFT TURNS	RIGHT TURNS
1	15618	4	3,478	3	2
2	15626	5	6,044	4	3
3	15639	2	1,523	2	1
4	15702	8	10,855	2	0
5	15770	18	21,333	5	3
6	15787	17	19,867	15	9
7	15849	5	4,575	4	3
8	15864	1	1,073	1	2
9	15938	19	23,185	5	3
10	15956	7	7,496	4	3

Figure 2.2 Portion of the attribute table generated by the GIS-based map-matching algorithm showing the important attributes for each observed route (i.e., travel time (minutes), travel distance (meters), and the number of left and right turns) relevant to route choice modeling

2.5 Results

2.5.1 Data input and preprocessing

The Halifax Space-Time Activity Research (STAR) Project was claimed to be the world's first largest GPS-assisted prompted-recall time diary survey (Bricka, 2008). The survey was conducted for a 2-day period covering approximately 2,000 households in Halifax, Nova Scotia, Canada from 2007 to 2008. Person-based GPS devices were used. The GPS data have a spatial resolution of within 10 meters (but generally <3m) and a temporal resolution of 3 recordings every 2 seconds. About 47 million GPS points were collected. The GPS data were obtained in SPSS format from the Halifax STAR Project then converted into GIS format as point features. Start time and end time corresponding to trip ends for work trips were extracted from a time diary episode data file into a matrix of respondent IDs, start time, and end time. The matrix was used to extract the portion of the daily trips corresponding to work trips by car using a Python script in ArcGIS®. Work trips by car were extracted because most individual daily trips consist of this kind of trip. Moreover, work trips are extensively studied in the field of transportation, particularly in route choice modeling. The selection of the sample is motivated by the potential application of the GIS-based algorithm to generate the input data for route choice modeling. Thus, the selection focused on interzonal, home-based work trips that are at least a kilometer in length, and performed by unique individuals. Out of 3,023 simple work trips from the STAR time diary - episode data file, about 574 home-based work

trips are selected. Some of the reported work trips in episode data file have missing GPS trajectories. All the GPS trajectories representing home-based work trips are preprocessed. Data preprocessing involved removal of outliers and gaps. GPS points with horizontal dilution of precision (HDOP) value greater than 2 are removed. Also, “position jumps” are removed if the calculated speed between two consecutive GPS points exceeds 50 m/s (Schuessler and Axhausen, 2009). Gaps are filled in using proximity analysis and data management tools in ArcGIS[®]. After data preprocessing, 104 work trips are finally selected.

A first experiment is performed on the sample of 104 work trips that accounts for about one percent of the data (46K points) or about 18 percent of total home-based work trips. Each work trip from the sample begins at home and ends at work place. The small sample was chosen because it is easy to manage and helps facilitate the manual validation of the routes generated by the algorithm – enough to illustrate the performance of the GIS-based map-matching algorithm. Future experiments will attempt to use the entire GPS dataset, starting with the application of the algorithm to extract routes for the 3,023 simple work trips. The validation process involved visual checking of home and work place locations and GPS trajectories with the aid of contextual information from time diary data, and satellite image. The preprocessed GPS data were stored in ArcGIS[®] file geodatabase ready for map-matching, representing about 104 individual work trips.

The GIS-based map-matching algorithm requires two inputs: (1) the preprocessed GPS data file stored in file geodatabase format, and (2) the network dataset. The

preprocessed GPS data for 104 individual work trips comprise about 440 points per trip. The network dataset was from DMTI Spatial CanMap® Route Logistics Version 2008.3 that provides a detailed road and highway network for Canada. A subset of this network was extracted for Nova Scotia because some of the trips go beyond the Halifax region. The road network for Nova Scotia consists of 116,647 links and 98,132 junctions.

2.5.2 Accuracy

The algorithm correctly generated the routes for 88 percent of the work trips (91 routes). The few inaccuracies are mainly attributed to the wrong turn restrictions in the network dataset from DMTI. Manual correction of the wrong turn restrictions produced accurate results. However, the algorithm performed well at complex intersections (Figure 3). The validation is performed by visually retracing the routes taken by respondents using time diary records for each of the 104 respondents. The advanced network data model and the shortest path algorithm in the GIS platform effectively utilized the network topological information enabling the GIS-based algorithm to produce accurate results. Other map-matching algorithms that use the planar network model have difficulty in matching GPS trajectories at complex intersections (Quddus et al., 2007). This is because of the limited capability of the planar network in modeling complex road layouts.

Unlike the planar network data model commonly used in existing map-matching algorithms (Marchal et al., 2005; White et al., 2000; Quddus et al., 2003), the proposed map-matching uses an advanced network data model of ArcGIS®. The planar network data model is a simple representation of road network in terms of nodes and arcs. It is

computationally efficient but very limited in making use of topological rules and road attributes to model the actual road network. For this reason, White et al. (2000), Quddus et al. (2003), and Marchal et al. (2005) call for the integration of network topology to improve accuracy, and address the limited capability of map-matching when it comes to complex road layouts, particularly at road intersections. The network dataset in ArcGIS® is an advanced network data model that fully utilizes connectivity rules and road attributes (e.g., costs, restrictions, road classification) that allow for the modeling of complex scenarios. The route analysis tool makes use of the ArcGIS® advanced network data model and has a potential in addressing the two prevailing issues in the literature: effective use of topological information and dealing with the problems that arise in road intersections. The route analysis uses the shortest path algorithm to solve for the best route based on the cost or impedance parameter. This produces accurate results in a matter of seconds. This is made possible with the use of the advanced network connectivity and attribute data model.

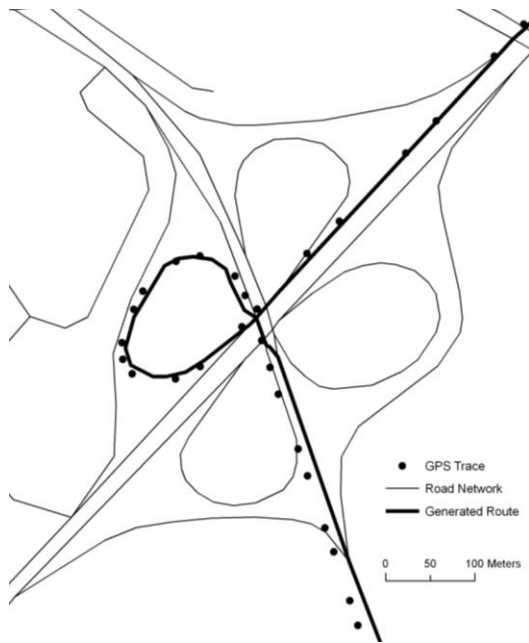


Figure 2.3 An example of a complex intersection where the GIS-based map-matching algorithm accurately generated the route for the GPS trace

The GIS-based map-matching algorithm avoids the problem with the initial map-matching process. The snapping function in the GIS environment works effectively in snapping the initial GPS point to the nearest road link. However, the resolution or the quality of the road network often affects the accuracy of the algorithm at the start and end locations. Access roads that connect parking areas or home locations to main roads are missing in most digital road network. At the start or end of the trip, the GIS-based map-matching uses the nearest road to match the GPS trajectories when access roads are missing (Figure 4). Based on the experiment results, the missing access roads to home locations or parking areas account for about 10 percent difference between the actual trip distances and generated routes. An in-depth analysis of problems associated with missing

links has not been fully addressed here but would be interesting to investigate in the future.

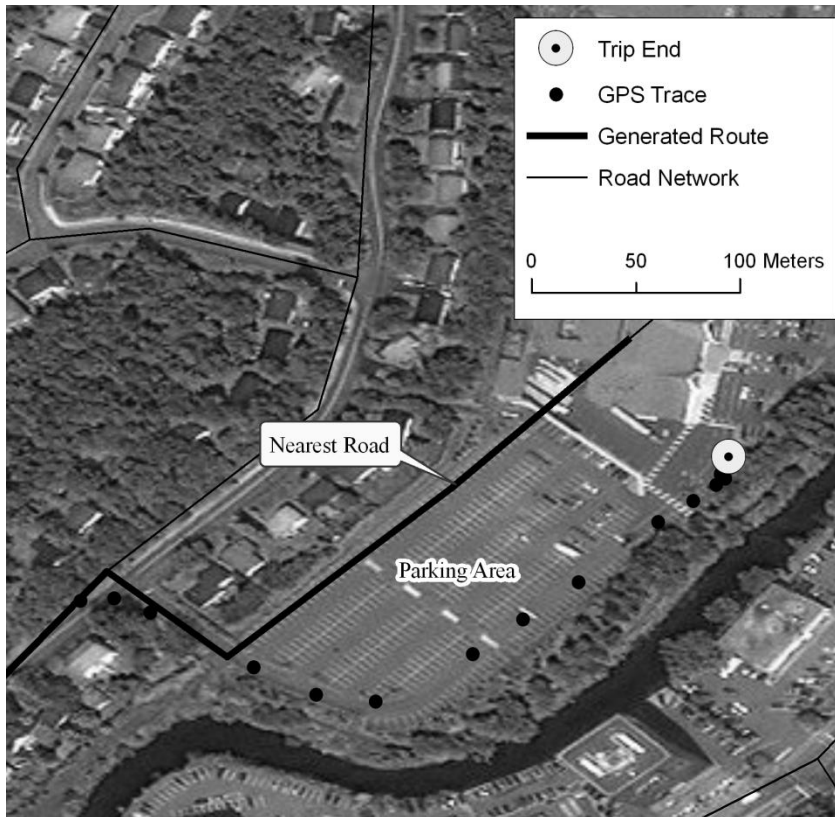


Figure 2.4 The algorithm sticks to the nearest road whenever some network link is missing

2.5.3 Computing speed

The testing of the algorithm is implemented in a PC with an Intel Core Duo processor clocked at 2.66 GHz with 3 gigabytes of physical memory. The experiment revealed an average computing speed of one minute per trip. This is approximately 6 s per point relative to the sample. The computing speed per trip seems to remain constant

regardless of an increase in the number of trips or route length (Figure 5). However, computing speed gradually increases with the increase in buffer distance or the increase in complexity of the GPS trajectory that prolongs the creation of the buffer region. Chung and Shalaby (2005) reported a computing speed of 2-6 minutes per trip in a PC with an Intel Pentium III 1 GHz. No direct comparison can be made with other previous studies because of the lack of objective and comparable performance indicators. The computing speed can be improved by minimizing the overhead processing through efficient coding.

Although considered important, computing performance is not the top priority for postprocessing map-matching for transportation research. However, the GIS-based map-matching algorithm demonstrated an acceptable computational speed in generating accurate routes for 88 percent of the work trips tested.

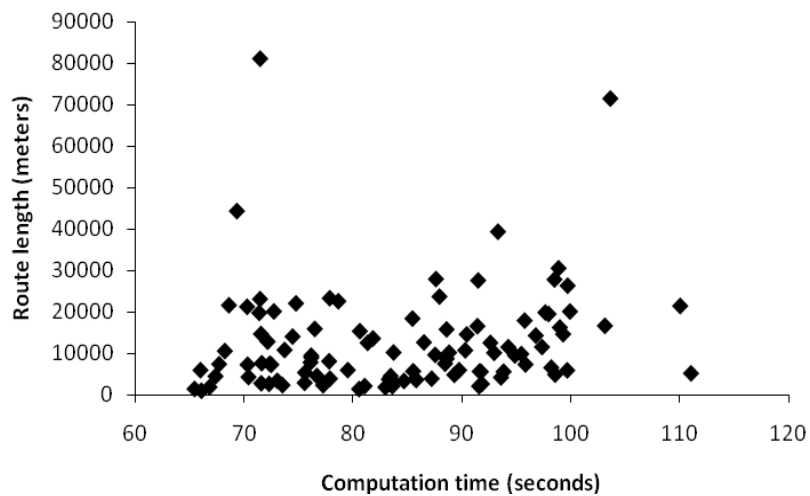


Figure 2.5 Plot of the computation time over route length for a sample of 104 routes

2.5.4 Advantages and limitations

The accuracy of the GIS-based map-matching algorithm is sensitive to the buffer distance. A sensitivity analysis is conducted on some randomly selected trips. These trips are selected because they are more complex than the rest, often with loops and sharp curves. Buffer distances of 10 m, 15 m, 20 m, ..., 100 m are tested. The results show that no routes are generated for buffer distances below 50 m. Inaccurate routes are generated for buffer distances of 60 m and above. Figure 6 shows the effect of buffer distances to the map-matching accuracy. Therefore, the buffer distance for the GPS trajectory should be, more or less, 5x to 6x the horizontal accuracy of GPS data. This range of buffer distance values accounts for the width of the roads, the sharpness of curves, and GPS positioning errors. Values greater than this threshold will cover irrelevant links resulting to incorrect routes generated while, values lower than the threshold will be too restrictive and no shortest path or route will be generated. The buffer distance that will produce accurate map-matching results depends on the complexity of the road network and the horizontal accuracy of GPS device. But this distance can be easily set by the user unlike some threshold values in other map-matching algorithms that need in-depth empirical study to determine the appropriate values for several parameters (Marchal et al., 2005; Quddus et al., 2003). Future research should perform a more thorough investigation concerning the buffer distance parameter and computing performance (e.g. using different GPS and road network datasets).

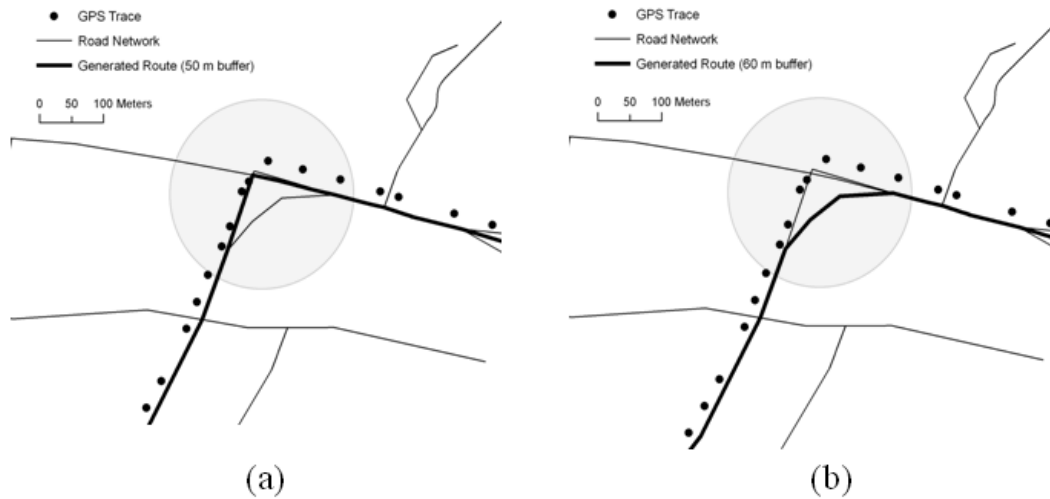


Figure 2.6 Sensitivity of map-matching algorithm to buffer distance: (a) buffer distance = 50 m and (b) buffer distance = 60 m

The shortest path algorithm in ArcGIS[®] is the core component of the map-matching algorithm proposed in this research. The shortest path algorithm alone is computationally efficient in generating routes and could also take into account multiple stops or destinations. With some modification, the algorithm can be extended or expanded to automatically extract multi-modal trips, similar to the trip reconstruction tool (Chung and Shalaby, 2005) or a GPS postprocessing tool (Schuessler and Axhausen, 2009).

The GIS-based algorithm is timely for the increasing availability of road network data from government sources and private data vendors. Rich network datasets of good quality are readily available, mostly from private data vendors for a reasonable price. This reduces the time to provide road network data required for the map-matching algorithm. In the absence of road network data, new data can be created in the GIS environment.

In summary, the advantages of the proposed algorithm are the simple user interface (GUI), parameters can be changed to suit the demands of a particular dataset (i.e. using the appropriate buffer distance), can be expanded to perform more functions (e.g. extract multi-modal trips), generates accurate routes within a reasonable amount of time, and its portability. The algorithm was developed using Python script and can be easily added in ArcGIS® as a tool. The portability of the script makes it available to many users. Also, the script can be easily edited and improved by accessing the script file in any text reader application.

2.6 Conclusion

This paper argues that a GIS is the ideal platform for the development of postprocessing map-matching algorithm for transportation research like route choice modeling because it is easier to develop and implement, is scalable, and generates accurate results at an acceptable computing cost. To support this argument, this paper presented a postprocessing algorithm developed and implemented in a GIS platform. The development of the algorithm is easy and fast by making use of the functionalities already available in commonly used GIS platforms. As shown in this paper, the GIS-based map-matching algorithm is able to deal effectively with complex road intersections and generate accurate routes at reasonable computing cost. The script can be improved and can be easily employed by researchers with GIS in their research environment. Basically the algorithm makes use of buffer and network analysis that can effectively be done in

GIS. The increasing availability of commercially available network datasets and GPS-assisted time use or activity surveys provide a timely basis for this kind of algorithm. This algorithm can be easily tailored to the needs of researchers in analyzing route choice behavior.

However, several issues need to be resolved for further improvement of the algorithm. Computing speed can be improved by the use of efficient coding and moving computing intensive processes to a faster programming language like C++. Seamless integration with the GPS data preprocessing is needed and this is another research direction that the authors will undertake. This integration may also include the development of a new module that will enable users to easily link GPS data with a time diary episode data file or travel survey data to enable extraction of reported trip ends, travel time and other information.

The GIS-based map-matching algorithm can be expanded to automatically detect and extract trips made by other modes such as public transportation, walking and cycling. The development of this trip reconstruction tool is perfectly suited for the Halifax STAR dataset and the authors are currently working towards this direction by utilizing GIS as a development platform. Moreover, the GIS-based map-matching algorithm presented in this paper is part of an on-going effort to develop a GIS-based toolkit for route choice modeling.

2.7 Acknowledgements

Financial support for this project was provided by a grant awarded to Darren M. Scott from the Natural Sciences and Engineering Research Council of Canada (261850-2009). The authors greatly acknowledge suggestions from the three anonymous reviewers in improving the quality of this paper. The authors also acknowledge the Halifax STAR Project for the GPS and time diary datasets used for the development and testing of the map-matching algorithm, David Wynne for the point-to-line script he made available for free, and the GEOmatics for Informed DEcisions (GEOIDE) for the travel grant awarded to the first author to present this work at the 14th AGILE Conference on Geographic Information Science.

2.8 References

- Bel Hadj Ali, A. (1997) Appariement geometrique des objets géographiques et étude des indicateurs de qualité. Saint-Mandé (Paris), Laboratoire COGIT.
- Brakatsoulas, S., Pfoser, D., Salas, R. and Wenk, C. (2005) On Map-Matching Vehicle Tracking Data. VLDB 2005, pp. 853-864.
- Bricka, S. (2008) Non-Response Challenges in GPS-Based Surveys, paper at the 8th International Conference on Travel Survey Methods, May 2006, Annecy, France.
- Cao, H. and Wolfson, O. (2005) Nonmaterialized Motion Information in Transport Networks. ICDT 2005, pp. 173-188.

- Casas, J., and Arce, C. H. (1999) Trip Reporting in Household Travel Diaries: A Comparison to GPS-Collected Data. Presented at *78th Annual Meeting of the Transportation Research Board*, Washington, D.C.
- Chung, E., and Shalaby, A. (2005) A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, Vol. 28, No. 5, pp. 381-401.
- Devogele, T. (2002) A new merging process for data integration based on the discrete Frechet distance. In *Advances in Spatial Data Handling*, D. Richardson and P. van Oosterom. New York, Springer Verlag, pp. 167-181.
- Doherty, S. (2001) Meeting the Data Needs of Activity Scheduling Process Modeling and Analysis. Presented at *80th Annual Meeting of the Transportation Research Board*, Washington, D.C.
- Draijer, G., Kalfs, N. and Perdok, J. (2000) Global Positioning System as Data Collection Method for Travel Research. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1719, TRB, National Research Council, Washington, D.C., pp. 147-153.
- Greenfeld, J. S. (2002) Matching GPS observations to locations on a digital map. Papers presented at the 81th Annual Meeting of the Transportation Research Board. CD-ROM. January 2002, Washington, DC.
- Harvey, F. (1994) Defining unmoveable nodes/segments as part of vector overlay: The alignment overlay. In *Advances in GIS Research*, T. C. Waugh and R. C. Healey. London, Taylor and Francis, 1, pp. 159-176.
- Harvey, F. (2005) Aligning or Matching: Cartographic Perspectives on Geographic Integration. AutoCarto 2005, Las Vegas, NV, ACSM.

- Harvey, F. and Vauglin, F. (1996a) Geometric match processing: Applying Multiple Tolerances. The Seventh International Symposium on Spatial Data Handling (SDH'96), Delft, Holland, International Geographical Union (IGU).
- Harvey, F. and Vauglin, F. (1996b) Geometric match processing: Applying Multiple Tolerances. In *Advances in GIS Research*, Proceedings of the Seventh International Symposium on Spatial Data Handling, M. J. Krakk and M. Molenaar. London, Taylor & Francis, 1, pp. 155-171.
- Lemarié, C. and Raynal, L. (1996) Geographic data matching: First investigations for a generic tool. GIS/LIS '96, Denver, Co, ASPRS/AAG/URISA/AM-FM.
- Lou, Y., Xie, X., Zhang, C., Wang, W., Zheng, Y. and Huang, Y. (2009) Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS)*, pp. 544-545.
- Marchal, F., Hackney, J. and Axhausen, K.W. (2005) Efficient map matching of large global positioning system data sets. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1935, Transportation Research Board of the National Academies, Washington, D.C., pp. 93-100.
- Murakami, E., and Wagner, D.P. (1999) Can Using Global Positioning System (GPS) Improve Trip Reporting? *Transportation Research Part C*, Vol. 7, pp. 149-165.
- Ogle, J., Guensler, R., Bachman, W., Koutsak, M. and Wolf, J. (2002) Accuracy of Global Positioning System for Determining Driver Performance Parameters. In *Transportation Research Record: Journal of the Transportation Research Board: No. 1818*, Transportation Research Board of the National Academies, Washington, D.C., pp. 12-24.

Pearson, D. (2001) Global Positioning System (GPS) and Travel Surveys: Results from the 1997 Austin Household Survey. Presented at 8th Conference on the Application of Transportation Planning Methods, April 2001, Corpus Christi, Texas.

Quddus, M. A., Ochieng, W.Y. and Noland, R.B. (2007) Current map-matching algorithms for transport application: State-of-the-art and future research directions. *Transportation Research Part C*, Vol. 15, pp. 312-328.

Quddus, M. A., Ochieng, W.Y., Zhao, L. and Noland, R.B. (2003) A general map matching algorithm for transport telematics applications. *GPS Solutions*, Vol. 7, pp. 157-167.

Schuessler, N. and Axhausen, K.W. (2009) Processing raw data from Global Positioning Systems without Additional Information. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2105, Transportation Research Board of the National Academies, Washington, D.C., pp. 28-36.

Taylor, G., Brunsdon, C., Li, J., Olden, A., Steup, D. and Winter, M. (2006) GPS accuracy estimation using map matching techniques: Applied to vehicle positioning and odometer calibration. *Computers, Environment and Urban Systems*, 30, pp. 757-772.

Vauglin, F. and Bel Hadj Ali, A. (1998) Geometric matching of polygonal surfaces in GISs. ASPRS Annual Meeting, Tampa, FL, ASPRS.

Wagner, D. P. (1997) Lexington Area Travel Data Collection Test: GPS for Personal Travel Surveys. Final Report. Office of Highway Policy Information and Office of Technology Applications, Battelle Transport Division, FHWA, Sept. 1997, Columbus, Ohio.

- Walter, V. and Fritsch, D. (1999) Matching spatial data sets: a statistical approach. *International Journal of Geographic Information Science*, 13(5), pp. 445-473.
- White, C. E., Berstein, D. and Kornhauser, A.L. (2000) Some map matching algorithms for personal navigation assistants. *Transportation Research Part C*, 8, pp. 91-108.
- Wolf, J., Hallmark, S., Oliveira, M., Guensler, R. and Sarasua, W. (1999) Accuracy Issues with Route Choice Data Collection by Using Global Positioning System. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1660, TRB, National Research Council, Washington, D.C., pp. 66-74.
- Yalamanchili, L., Pendyala, R.M., Prabakaran, N. and Chakravarty, P. (1999) Analysis of Global Positioning System-Based Data Collection Methods for Capturing Multistop Trip-Chaining Behavior. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1660, TRB, National Research Council, Washington, D.C., pp. 58-65.
- Zhou, J. (2005) A three-step general map matching method in the GIS environment: Travel/transportation study perspective. UCGIS Summer Assembly 2005. Wyoming. <http://www.ucgis.org/summer2005/studentpapers.htm>. Last date accessed 07.2010.

Chapter 3

Making Mode Detection Transferable:

Extracting Activity and Travel Episodes from GPS Data

Using the Multinomial Logit Model and Python

3.1 Introduction

The prevalence of global positioning systems (GPS) devices and their increasing use in transportation research (e.g., Murakami & Wagner, 1999; Draijer et al., 2000; Bohte & Maat, 2009) and other fields (e.g., Maddison & Ni Mhurchu, 2009) calls for practical approaches in leveraging GPS for research purposes. GPS data provide detailed accounts of stationary activity and travel episodes, traditionally captured by time-use surveys, but with more accuracy, better frequency and lesser burden to respondents (Wolf, 2000; Stopher et al., 2005; Bricka, 2008; Stopher & Shen, 2011). Extracting activity episodes from GPS data involves two main processes: (1) the extraction of segments (i.e., sequence of points classified as stop, trip, or mode transfer points) and (2) classification of these segments into stop episodes (stationary trajectories) or various travel episodes such as walk, car, and so on (movement trajectories); see Figure 3.1. In this paper, the authors propose a transferable and efficient method of automatically extracting and classifying activity episodes (referred hereafter as episodes) from GPS data without any additional information. The proposed method uses a multinomial logit (MNL), which provides a transferable and efficient approach in classifying extracted

episodes into different types. The preliminary results are promising in the light of the transferability issue that impedes the widespread use of GPS devices for transportation research. Transferability, as used here, refers to the ability of a classification method to be applied in different environments with minimal effort due to its minimal dependencies and objectivity in determining threshold values in predicting episode types.

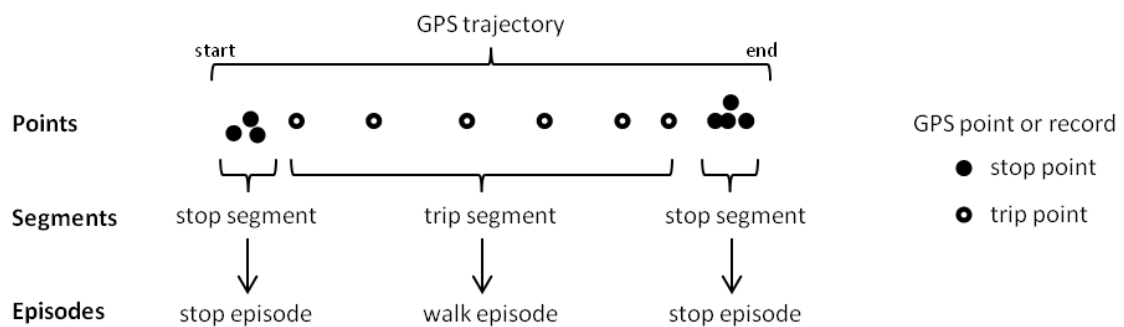


Figure 3.1 GPS trajectory subdivided into points, segments, and episodes

In passive mode, a person-based GPS data logger can collect huge amounts of raw data spanning several days. The raw data provide useful information on activity episodes - typically consisting of information on the location, duration, travel mode, and route. Several methods have been developed to automatically extract episodes from GPS data, particularly in imputing travel modes based on these episodes. These methods of classifying segments from GPS data into different activity episodes are called mode detection or movement trajectories classification (e.g., Schuessler & Axhausen, 2009; Biljecki, 2010). Often dictated by available data and unique research purposes, these methods are difficult to apply in different geographic areas. Transferability of the existing

methods remains a challenge. This issue was identified by Lawson et al. (2010) based on their experiment that aimed to replicate currently used methods on mode detection. Their experiment provides an objective examination of currently used mode detection algorithms to date. They recognized the difficulty of directly comparing different approaches because of different data used in developing these methods, not to mention the different variables required by each approach. Transferability of various methods of extracting episodes from GPS data is difficult, and often researchers narrowly focus on the performance of their proposed methods. This problem can be attributed to many, if not, unique input variables required and the lack of clear-cut input values that will yield the best results. Moreover, the authors believe that the main reason for the lack of transferability is the method used in classifying episodes extracted from GPS data. These classification methods have specific data requirements that are hard to replicate in different application environments, not to mention the subjectivity of selecting threshold values used in classification.

Following Lawson et al. (2010), there are three basic methodologies in classifying episodes: rule-based, neural networks, and fuzzy logic (for a comprehensive list, see Biljecki, 2010). Rule-based methods are linked with data-intensive processes and require many trials to determine the best input values that often vary in different application environments (e.g., Chung & Shalaby, 2005; Bohte & Maat, 2009; Gong et al., 2012). In addition, rules vary across different researchers and data used. For these reasons, the rule-based approach is the least transferable among the three methodologies. Neural networks

(e.g., Gonzalez et al., 2008) need a good quality training dataset and it is difficult to determine the best variables automatically except to compare percent of test samples accurately predicted. Unlike rule-based methods, neural networks can be adapted to different training datasets, which makes them more transferable. However, the problem with neural networks is the fine-tuning of input weights to achieve the best predictive accuracy. Fine-tuning is often an ad hoc process lacking in clear guidelines, aside from little guidance given on the selection of neural network architecture (Hunt & Lyons, 1994). As training datasets vary across different application environments, it is difficult to determine how much training is sufficient. Fuzzy logic (e.g., Tsui & Shalaby, 2006; Schuessler & Axhausen, 2009) suffers the same limitation of rule-based methods, as it is difficult to find the right combination of rules that will produce the best results. These three basic methodologies also lack the flexibility of accommodating additional input variables as dictated by different application environments. For example, the classification of GPS segments into bus episodes often assumes that segment endpoints are within a certain distance of a bus stop (e.g., Tsui & Shalaby, 2006). This distance threshold used as a predictor is ineffective in environments where most bus stops are located where car episodes also begin or end such as shopping malls, government buildings, street intersections, and so on. Eventually this predictor will be dropped and one has to search for another input variable in its place. Within the frameworks of existing methodologies, searching for additional input variables takes many trials and errors. Logit models, as used in travel demand models, have been investigated for their

transferability in space and time (e.g., Atherton & Ben-Akiva, 1976; Wilmot, 1995; Cotrus et al., 2005). As logit models appear to be stable across different data sources, we can adopt this technique as a more transferable and efficient method of classifying episodes extracted from GPS data.

This paper argues that the transferability of the method of extracting episodes automatically from GPS data can be significantly improved by using generic variables and MNL to classify GPS segments into episodes. By using MNL, input weights are automatically estimated - significantly reducing the burden of determining thresholds for input variables as in rule-based methods. One also avoids the complexity of setting up analysis layers associated with neural networks by having a parsimonious MNL model. In addition, MNL specification and estimation is easier to implement than setting up rules when using fuzzy logic. Utility specification in MNL provides flexibility for its use in various classification problems (e.g., easy to add variables in utility specification), and has been pointed out as its strength (e.g., Koutsopoulos et al., 1994; Sorci et al., 2010). Following Karlaftis and Vlahogianni (2011), all mentioned methods have advantages and limitations but often simple ones give as good results as complex ones. MNL looks promising as a transferable and efficient method in classifying GPS segments into activity and travel episodes, with an overall accuracy of 90% as demonstrated in this paper. Reported overall accuracies of existing methods range from 70 to 95%, based on the comparative experiments done by Lawson et al. (2010) and the extensive list reviewed by Biljecki (2010).

The following sections discuss the proposed method along with the sample data used to demonstrate the automatic extraction and classification of episodes into five types: *stop*, *walk*, *car*, *bus*, and *other* (travel) episodes. Data and methods are presented briefly introducing existing and refined procedures adopted from existing algorithms for GPS data preprocessing. The remaining sections focus on the specification and validation of a MNL model for classifying GPS segments into different activity episodes. The concluding section summarizes the potential of MNL along with the segmentation techniques in extracting activity episodes from GPS data.

3.2 Data and methods

This section describes the GPS and time-use diary data used in the development and testing of the proposed method of extracting activity and travel episodes. Also, the GIS-based Episode Reconstruction Toolkit (GERT; Dalumpines & Scott, 2014) components for mode detection is presented, summarizing the key steps involved in generating statistical descriptors for extracted GPS episodes. Finally, this section presents the MNL model specification based on the statistical descriptors.

3.2.1 GPS and time-use diary (TUD) data

The proposed method, consisting of GERT's extraction and mode detection component (Figure 3.2; further discussed in Section 3.2.2), is demonstrated using a GPS data set from the Space-Time Activity Research (STAR) project - a comprehensive survey of time-use and travel activity conducted in Halifax, Nova Scotia, Canada from

April 2007 to May 2008 (for more information on this data set, see Millward and Spinney, 2011). Apart from a time-use diary, respondents carried a GPS-equipped mobile device (Hewlett Packard iPAQ hw6955), which recorded a location every second and with a horizontal accuracy ≤ 10 m. The GPS data logger collected positional data that includes unit ID, date, time, x-coordinate, x-direction (north/south), y-coordinate, y-direction (east/west), speed, altitude, horizontal dilution of precision (HDOP), and the number of satellites; 1,967 respondents collected over 47 million points for two survey days (equivalent to 5,127 person-days). STAR time-use diary (TUD) data provided a sample of 7,271 reported episodes from 1,277 respondents that matched extracted GPS episodes within five minutes of the TUD episode's start or end time. Of particular interest, each TUD episode has information on the activity location such as home, workplace, car, bus, and so on.

A helper module written in Python[®] as part of GERT automatically extracts the matching episodes from diary and GPS data. In extracting matching episodes, this module automatically classifies TUD episodes as *stop* episodes if their associated locations were outside of travel modes (e.g., activity episodes performed at home were labeled as *stop* episodes). This was done to correctly match TUD episodes with that of GPS episodes. Out of 7,271 episodes, 49% (3,569) were reported by survey respondents as *stop* episodes, 43% (3,131) as *car* episodes, 5% (390) were recorded as *walk* episodes, while the minority consists of 2% (110) for *bus* episodes and 1% (71) for *other* (travel) episodes. Since other travel episodes have a very small share in the sample, these minor

travel episodes were grouped together under *other* episodes. These minor travel episodes included the following travel modes: bicycle (42), boat ferry (19), motorcycle (9), and refused to be reported (1). *Bus* episodes were retained to reflect public transportation modes in the classification scheme. In general, about 90% (6,544) of the sample from the STAR data was used for the MNL model estimation, and the remaining 10% (727) for model validation.

3.2.2 The proposed method: GERT's Extraction and Mode Detection Module (MDM)

The proposed method (of extracting and classifying episodes from GPS data) is one of the main components or modules in GERT, referred to as the GPS Episodes Extraction and Mode Detection Module (MDM) or the Stage 2 in GERT's workflow (Figure 3.2). The development of GERT was motivated by the lack of practical tools that can automatically extract information from GPS data for activity analysis in general and route choice modeling in particular. To make it practical, GERT used a framework built around three design principles: modularity, transferability, and scalability (Dalumpines & Scott, 2014). GERT and its modules were designed to work on minimal input requirements: latitude, longitude, and time; but can easily scale-up to accommodate additional inputs apart from GPS data or additional modules without losing its integrity. As part of GERT, the proposed method or MDM was designed to be transferable to different application environments because it only relies on location coordinates and time stamps (generic inputs), and takes advantage of the efficiency and flexibility provided by MNL in classifying episodes. GERT's workflow (Figure 3.2) is discussed in detail in

another paper by the authors dealing with the entire GERT components (Dalumpines & Scott, 2014). The rest of this sub-section provides an overview of the inputs and processes in GERT's MDM, and summarizes the main steps involved.

Each GPS point $p_i = \{\text{latitude, longitude, time}\}$, has a coordinate (latitude, longitude) and time stamp. A sequence of GPS points $p_i \in \{p_1, p_2, \dots, p_n\}$, called GPS trajectory (Figure 3.1), represents an individual's movement for a 24-hour period. The coordinates and time stamps were used to extract distance, heading, duration, speed, and acceleration information. This extracted information was used to derive statistical descriptors (observed characteristics), later used in establishing decision rules for segmentation of preprocessed or valid GPS trajectories (i.e., slicing each trajectory into stationary or movement segments). Then derived statistical descriptors were used as explanatory variables or predictors in the classification of episodes into several episode types using MNL.

GERT's GPS Preprocessing Module (GPM in Figure 3.2) generates valid GPS trajectories as inputs to MDM. GPM removes redundant records (i.e., records with the same coordinates) and outliers with speed greater than or equal to 50 m/s (Wolf et al., 1999; Schuessler & Axhausen, 2009). As a data-preprocessing component, GPM adopted data cleaning procedures introduced by Schuessler and Axhausen (2009), with additional rules that allowed automatic removal of errors associated with urban canyon effects.

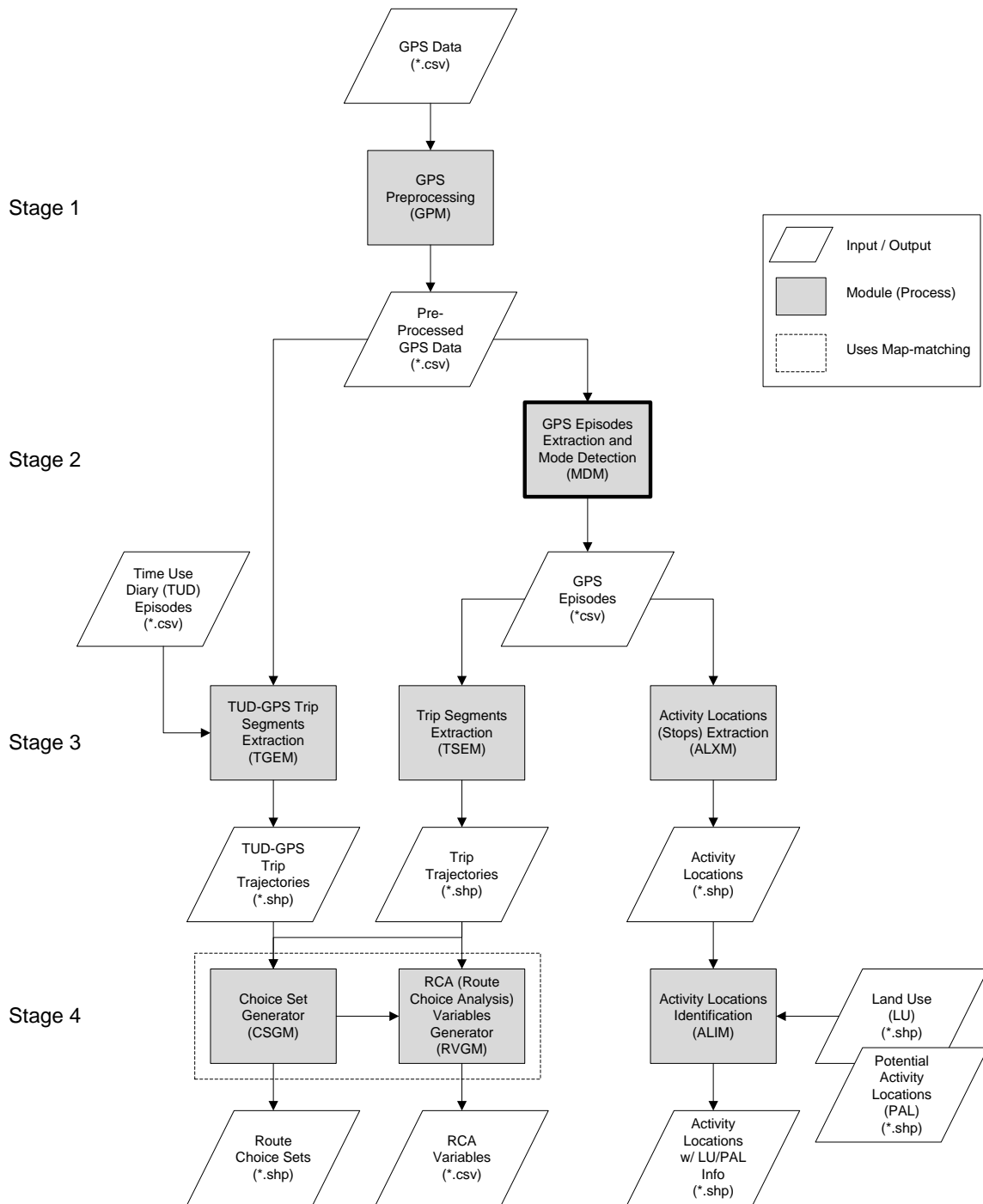


Figure 3.2 Four-stage workflow of GIS-based episode reconstruction toolkit (GERT), the proposed method of extracting and classifying episodes from GPS data is highlighted in bold

GPM is further discussed in detail in another paper by the authors (Dalumpines & Scott, 2014). Given an input of valid GPS trajectories, the episode extraction and classification component (MDM) consists of the three steps as summarized below:

1. Using the concept of stops as location anchors, the first step classified each record from valid GPS trajectories into three types of points: stop, trip, or mode transfer point (Chung & Shalaby, 2005; Zheng et al., 2008). Generally, records classified as stop points have speed ≤ 0.15 m/s (based on typical GPS device velocity accuracy with horizontal accuracy ≤ 10 m), and duration or dwell time ≥ 120 s as used in previous studies (e.g., Wolf et al., 1999; Stopher et al., 2005). Mode transfer points (MTP) are similar to stops but with duration of less than 120 s. Trip points are those records that did not fall under stop or MTP categories. Then, similar records were merged together to form two types of segments: stop segments comprised of stop points, and trip segments comprised of trip points or MTPs. Stop segments were initially classified into *stop* episodes.
2. The second step partitioned travel episodes into walk or non-walk, where non-walk episodes have an average speed greater than 0.91 m/s (LaPlante & Kaeser, 2007; Gong et al., 2012), distance ≥ 55 m (Schuessler & Axhausen, 2009), average heading less than 30 degrees, and consist of at least three records (Bohte & Maat, 2009). Generally, travel episodes, comprised of MTPs, were classified as walk episodes if their duration exceeded 60 s;

otherwise, they were merged with non-walk episodes. The threshold values were determined based on a series of experiments (e.g., average heading < 30 degrees) and previous studies as cited above. At this step, \mathbf{K} statistical descriptors were computed for each episode (i.e., *stop*, *walk*, or *non-walk*), where $\mathbf{K} = \boldsymbol{\mu}'\mathbf{x}$; $\boldsymbol{\mu} = [\text{min, max, mean, median, standard deviation, sum}]$; $\mathbf{x} = [\text{distance, heading, duration, speed, change in heading, acceleration}]$ - a total of 36 descriptors excluding the count of GPS points in each episode; see Table 3.1 for an example of these descriptors. Note that \mathbf{x} variables were used as these variables can be easily derived from any GPS data (i.e., *generic*) and commonly used in previous studies, particularly speed and acceleration (e.g., Chung & Shalaby, 2005; Stopher et al., 2005; Zheng et al., 2008; Bohte & Maat, 2009; Dodge et al., 2009; Biljecki, 2010; Lawson et al., 2010). As an overview, average values of selected statistical descriptors are shown in Table 3.2.

3. Finally, a MNL model was specified for all activity episode types of interest, which include the travel episodes. In this research, the other travel or non-walk episodes were classified into travel by *car*, *bus*, or *other* (travel) modes. The MNL model specification used the statistical descriptors as explanatory variables for the classification of all episodes into J episode types, where $J = \{\text{stop, walk, bus, car, other}\}$. Dodge et al. (2009) also used statistical descriptors effectively as inputs to supervised vector machine in classifying

trajectories. Each episode was assigned to the type with maximum probability (Zheng et al. 2008; Schuessler & Axhausen, 2009).

Table 3.1 An example of statistical descriptors used as MNL predictors

episode	start_time	end_time	<i>dist_min</i>	<i>dist_max</i>	<i>dist_median</i>	<i>dist_avg</i>	<i>dist_stdev</i>	<i>dist_sum</i>
1	4:00:50	8:24:12	8.7	15.6	12.4	12.3	3.1	49.1
2	8:24:12	8:28:54	2.6	67.7	29.0	26.4	13.2	2349.2
3	8:28:54	8:32:01	8.5	8.5	8.5	8.5	0.0	8.5
4	8:32:01	8:36:49	1.6	62.4	23.2	22.5	13.2	2412.0
5	8:36:49	8:41:21	10.6	10.6	10.6	10.6	0.0	10.6
6	8:41:21	8:43:07	1.2	11.3	4.3	5.9	3.6	81.9
7	8:43:07	8:45:56	22.2	22.2	22.2	22.2	0.0	22.2
8	8:45:56	8:52:04	0.6	59.9	26.2	28.5	13.6	4154.1
9	8:52:04	11:06:55	14.1	20.1	17.1	17.1	3.0	34.2
10	11:06:55	11:11:48	2.4	68.9	27.4	29.3	14.9	3430.8
11	11:11:48	11:19:56	0.7	11.5	6.1	6.1	5.4	12.2
12	11:19:56	11:24:44	2.5	118.6	25.9	29.1	18.1	2361.1
13	11:24:44	11:36:52	7.5	7.5	7.5	7.5	0.0	7.5
14	11:36:52	11:40:34	4.2	62.2	24.6	26.2	14.2	1835.9
15	11:40:34	12:34:02	8.8	8.8	8.8	8.8	0.0	8.8
16	12:34:02	12:43:24	1.0	192.5	24.1	26.7	22.1	3233.5
17	12:43:24	16:46:21	6.7	11.1	8.9	8.9	2.2	17.8
18	16:46:21	16:53:44	1.0	87.4	32.1	35.0	18.5	5946.6
19	16:53:44	17:21:28	6.4	8.3	7.4	7.4	1.0	14.7
20	17:21:28	17:23:40	2.3	62.0	29.1	31.8	17.7	1652.8
21	17:23:40	17:26:41	5.0	5.0	5.0	5.0	0.0	5.0
22	17:26:41	17:30:21	1.6	160.5	27.2	31.8	24.1	1687.7
23	17:30:21	18:03:05	2.0	8.2	5.1	5.1	3.1	10.2
24	18:03:05	18:09:50	3.9	84.1	33.5	36.5	19.7	5955.2
25	18:09:50	23:58:45	7.7	12.0	9.9	9.9	2.1	19.8

Distance in meters; *dist_min* = minimum distance, *dist_max* = maximum distance, *dist_median* = median distance, *dist_avg* = average distance, *dist_stdev* = standard deviation distance, and *dist_sum* = total distance. Similar set-up for *heading*, *duration*, *speed*, *change in heading*, and *acceleration* but not shown for lack of space.

Table 3.2 Average values of selected statistical descriptors by episode types

Selected predictors ^a	Episodes ($n = 6,544$)				
	Stop ($n = 3,208$)	Car ($n = 2,823$)	Walk ($n = 355$)	Bus ($n = 96$)	Other ($n = 62$)
Median speed (m/s) ^b	0.7	12.0	1.7	7.4	4.9
Maximum speed (m/s)	1.3	40.0	8.5	39.0	22.4
Median change in heading (degrees) ^b	99.6	4.6	12.4	5.7	4.9
Maximum change in heading (degrees)	121.0	139.7	123.3	167.8	136.4
Average acceleration (m/s ²)	0.054	0.319	0.073	0.242	0.163
Median acceleration (m/s ²)	0.040	-0.095	-0.005	-0.025	-0.041
Total distance (m)	27.1	11,606.1	515.2	11,143.8	7,112.3
Total duration (min) ^b	218.9	16.0	8.4	29.0	27.8

^a Average values may be higher in some cases due to the presence of noise in the sample GPS trajectories (e.g., the median speed for walk episode is quite high compared to the 1.5 m/s average walking speed reported by Knoblauch et al. (1996)).

^b Used in final MNL model.

3.2.3 Multinomial logit (MNL) model as classifier

Using random utility models, each episode n has an episode type utility function $U_{nj} = V_{nj} + \varepsilon_{nj}$ that can be decomposed into a deterministic component V_{nj} and an error (random) component ε_{nj} . The deterministic part (representative utility) is assumed to contain variables derived from extracted GPS episodes, while the random part corresponds to the unaccounted factors that affect the utility not included in V_{nj} . The representative utility was designated as a linear function of the episode \mathbf{K} statistical descriptors, which can be assumed as characteristics or indicators of J episode types. Thus, the representative utility for episode n classified as episode type $j \in J$ can be written as:

$$V_{nj} = \alpha_j + \sum_{k=1}^K \beta_{kj} x_{kn}$$

where

α_j = alternative-specific constant for episode type j ;

β_{kj} = coefficient of statistical descriptor k for episode type j ; $k=1, \dots, K$; and

x_{kn} = statistical descriptor k for episode n .

Following McFadden (1974), the logit model was derived by assuming that the error components are independently, identically distributed (IID) extreme value. With this assumption, the probability that episode n will be classified as episode type i becomes

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j \in J} e^{V_{nj}}}.$$

Because of the IID assumption, the MNL model restricts the odds of choosing one episode type over another to be independent of other episode types, known as independence from irrelevant alternatives (IIA) property. This restriction implies that the introduction of a new episode type in the set will affect all other episode types proportionately. This assumption can be avoided by allowing for correlated errors through other model specifications like nested logit and mixed logit. The nested logit model groups episode types that share unobserved attributes at different nest levels, which allows error terms within the nest to be correlated. As a highly flexible model, mixed logit allows for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. More details on these two alternative model structures can be found in the work of Train (2009).

In this study, it was found that the MNL model was more appropriate for the data (see Table 3.1 for sample data format). Nested logit and mixed logit require multiple observations for each episode, where each observation consists of indicators for each episode type. Since there is only one observation per episode, nested logit and mixed logit were not implemented. Diagnostics tests were used to check if the IIA property was violated. The Hausman test (Hausman & McFadden, 1984) and the Small-Hsiao test (Ben-Akiva & Lerman, 1985) of the IIA assumption were conducted. Diagnostic test results showed no strong evidence that IIA was violated. To explicitly deal with multicollinearity, predictor variables were also evaluated based on their variance inflation factors (VIFs) and those with lower VIFs were finally selected (O'Brien, 2007). MNL estimation and diagnostics tests were performed using Stata[®].

3.3 Results and discussion

Implementation of the proposed method (GERT's MDM) extracted episodes from GPS data, and generated statistical descriptors. Around 36 descriptors were used to specify the MNL model, which was used in the classification of extracted episodes into five types (*stop*, *walk*, *car*, *bus*, and *other* episodes). In this section, model results are presented that highlight the typical characteristics of episodes captured by the selected descriptors. Moreover, the classification results are evaluated using a validation sample described in the previous section. Lastly, this section describes the average performance of the proposed method in comparison with existing methods cited in Section 3.1.

3.3.1 Classification of activity episodes

After exploring a variety of different model specifications, the final model that consisted of the following statistical descriptors or variables: median speed (m/s), median change in heading (degrees), and total duration (min), had the best model fit (Table 3.3). The final model has an adjusted rho-squared of 0.81, significant at the 0.01 level, and was estimated using a random sample of 6,544 episodes from STAR data (90% of 7,271 episodes; remaining 10% were used for model validation). Moreover, the final MNL model has a VIF below three.

In general, the signs of the parameters as shown in Table 3.3 were consistent with the expected characteristics of travel episodes (i.e., *car*, *walk*, *bus*, and *other*; *stop* is base outcome) in terms of three variables: median speed, median change in heading, and total duration. The positive signs for median speed imply that the utility of an episode type and the probability that it will be chosen (or classified as that type) increases as the median speed of that episode increases. The coefficients of median speed across travel episode types indicate a hierarchy – faster travel episodes have higher utilities compared to slower travel episodes. The *other* episode is a collection of episodes that involve other modes of travel, and the model results suggest that this travel episode is faster than *walk* episodes but slower than *car* or *bus* episodes. On the other hand, the negative signs of median change in heading and total duration imply that the utility of an episode type and the probability that it will be classified as that type decreases as median change in heading and total duration of that episode increases. In other words, median speed is a utility

while median change in heading and total duration are disutilities in the classification of episodes extracted from GPS data. The alternative-specific constants are considered to represent the average effect of all factors that influence the classification but are not included in the utility specification.

Table 3.3 Model estimation results for classification of GPS episodes

Independent variable	Coefficient (t-statistic) ^a			
	Car	Walk	Bus	Other
Alternative-specific constant	-0.6012 (-2.31)	2.7734 (9.74)	-1.3019 (-2.83)	2.3317 (3.96)
Median speed (m/s)	1.2477 (12.84)	0.2120 (2.07)	0.9445 (9.33)	0.5991 (5.50)
Median change in heading (degrees) ^b	-0.1185 (-9.25)	-0.1436 (-11.82)	-0.1921 (-5.02)	-0.5383 (-7.82)
Total duration (min)	-0.0394 (-7.60)	-0.0200 (-6.19)	-0.0060 (-2.24)	-0.0118 (-3.21)

^a Number of observations = 6,544; null log-likelihood = -6389.13; final log-likelihood = -1192.66; adjusted rho-squared = 0.811. All coefficients were significant at 95%. Stop episode is base outcome with coefficients restricted at zero.

^b The median of the differences in bearings (in degrees) between two consecutive points, for all latitude/longitude points in the sample.

Table 3.4 shows that for a unit increase in median speed, the odds (or relative risk) of the episode being classified as a *car* episode relative to *stop* episode increases by 248%, holding median change in heading and total duration constant. Using the same episode type comparison, a unit increase in median change in heading and total duration decreases the odds by 11% and 4%, respectively.

Table 3.4 Percent change in odds to assess ability of selected predictors in classifying GPS episodes

Odds comparing Episode <i>A</i> vs Episode <i>B</i>		Percent Change in Odds due to a Unit Change in		
		Median speed (m/s)	Median change in heading (degrees)	Total duration (min)
Walk	Car	-64.5	—	2.0
Walk	Bus	-51.9	—	-1.4
Walk	Other	-32.1	48.4	—
Walk	Stop	23.6	-13.4	-2.0
Car	Walk	181.7	—	-1.9
Car	Bus	35.4	—	-3.3
Car	Other	91.3	52.2	-2.7
Car	Stop	248.2	-11.2	-3.9
Bus	Walk	108	—	1.4
Bus	Car	-26.2	—	3.4
Bus	Other	41.3	41.4	—
Bus	Stop	157.1	-17.5	-0.6
Other	Walk	47.3	-32.6	—
Other	Car	-47.7	-34.3	2.8
Other	Bus	-29.2	-29.3	—
Other	Stop	82.0	-41.6	-1.2
Stop	Walk	-19.1	15.4	2.0
Stop	Car	-71.3	12.6	4.0
Stop	Bus	-61.1	21.2	0.6
Stop	Other	-45.1	71.3	1.2

— not significant at 95% level. Boldface indicates highest value.

Based on the highest values of percent change in odds (values highlighted in bold in Table 3.4), the model tends to be most sensitive to: (1) change in median speed when comparing *car* over *stop* episodes, (2) median change in heading when comparing *stop* over *other* episodes, and (3) total duration when comparing *stop* over *car* episodes. Also, the percent change in odds that are not significant (Table 3.4) indicate that the median change in heading and total duration perform poorly in differentiating *walk*, *bus*, or *other* episodes from the rest of the episode types. Among the three explanatory variables in the

MNL model, median speed has greater influence in classifying episodes because of higher magnitudes of percent change in odds for this variable. However, total duration has the least influence among the three probably because episodes exhibit different duration regardless of episode types. Note, for example, the closeness of average episode duration for *bus* and *other* travel episodes in Table 3.2. On the other hand, there are strong contrasts among episode types in terms of average values of their median speeds (Table 3.2).

In most cases, there exists a strong hierarchy in terms of median speed with *car* episodes having higher median speeds followed by *bus*, *other* (travel), *walk*, and *stop* episodes, in that order. There also exists a hierarchy among episode types in terms of median change in heading, with high variations for GPS trajectories within *stop* episodes followed by *walk*, *bus*, and *other* episodes; however, the values for motorized travel episodes (i.e., *car*, *bus*, and *other*) are quite similar but largely different when compared with *walk* and *stop* episodes. As expected, GPS trajectories will show high heading variations in indoor locations (*stop* episodes) or when walking versus when driving a car. In general, these observations reflect the expected patterns of GPS trajectories under different episode types and were captured successfully by the MNL model. These patterns are visually reflected in Figure 3.3, which shows how predicted probabilities of episode types vary with changes in one of the explanatory variables while controlling for the rest. For example, there was a huge difference in terms of predicted probabilities when comparing *car* over *stop* episodes when varying median speed values while controlling

for the other two explanatory variables (set at their mean values). This is illustrated in Figure 3.3a, at median speed below 9 m/s, episodes are likely to be classified as *stop* episodes relative to other episode types; beyond this point *car* episodes tend to dominate (with higher predicted probabilities) over the rest of episode types. It is interesting to note that Chung and Shalaby (2005) classified episodes with maximum speed over 8.9 m/s to be car episodes; though median speed as a predictor is more robust to extreme values than maximum speed. With the exception of *stop* and *car* episodes, *bus* episodes tend to dominate at 10 m/s over other travel episodes.

When varying values for median change in heading (Figure 3.3b) while setting median speed and total duration at their mean values, *other* episodes tend to dominate below median change of heading of five degrees, from that point until 19 degrees *bus* episodes take over, and over 19 degrees *stop* episodes dominate. Among travel episodes, *car* episodes tend to dominate at median change in heading of over 20 degrees followed by *bus*, *walk*, and *other* episodes in that order, but the differences in their predicted probabilities narrow down as values approach 60 degrees.

When holding median speed and median change in heading at their mean values while varying total duration of episodes (Figure 3.3c), *car* episodes dominate over the rest of episode types if total episode duration is below 15 minutes; after this point *stop* episodes take over. A closer look at Figure 3.3c reveals that *walk* episodes slightly dominate over other travel episodes with the exception of *car* episodes.

In summary, the parsimonious MNL model using only three predictors (median speed, median change in heading, and total duration) provided an efficient method of classifying GPS episodes. As expected, model estimates confirm the typical characteristics of different episode types in terms of the three predictors as follows: (1) episodes with faster median speed are more likely associated with faster travel modes, (2) slower travel episodes or *stop* episodes tend to be associated with higher variation in heading, and (3) longer duration episodes are more likely linked to *stop* episodes. Moreover, GERT's preprocessing and mode detection modules written in Python[®] demonstrated potential in making automatic classification of episodes transferable to GPS data collected from different locations. These modules easily provided around 36 statistical descriptors that can be used for MNL model specification. Aside from being transferable and efficient, validation results suggest that the MNL model has potential as an episode classifier as shown in the next subsection.

3.3.2 Accuracy

In Table 3.5, about 727 episodes (10% of the 7,271 episodes) from the STAR data were used for validation. This validation set was selected under the restriction that total daily duration of the episodes must have nearly the same duration as the reported episodes from time-use diaries (i.e., within five minutes of episode's start or end time). The restriction ensures that reported episodes from time-use diaries are reasonably comparable with those extracted from GPS trajectories, a more conservative restriction than the 30-minute difference suggested by Stopher et al. (2011).

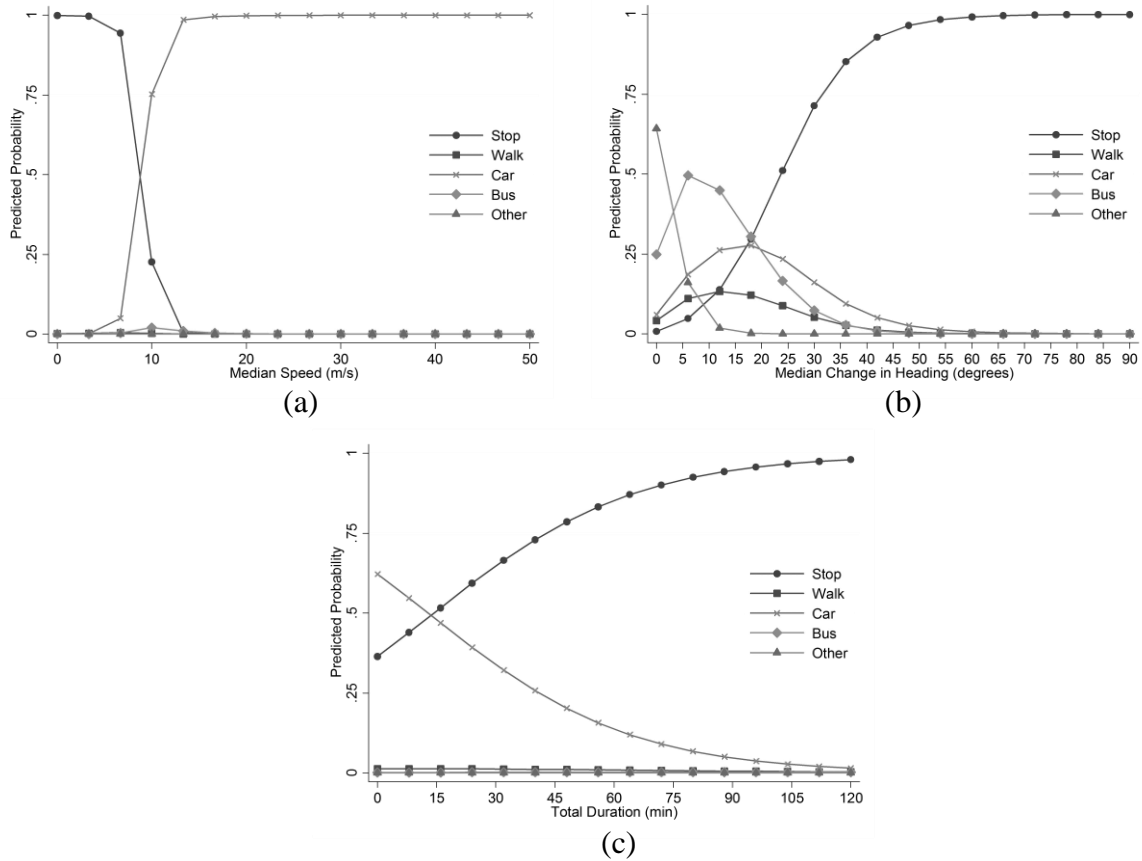


Figure 3.3 Graphs of GPS episodes in terms of predicted probabilities when varying one of the explanatory variables while setting the rest at their mean values (sample of 16 cases). The effects of varying median speed to predicted probabilities are shown in (a), median change in heading in (b), and total duration in (c)

Table 3.5 shows the observed and predicted values from the MNL model estimated using three statistical descriptors or independent variables: median speed (m/s), median change in heading (degrees), and total duration (min). The row percentages indicate the percentage of observed episodes in a given type that was predicted by the MNL model as *stop*, *car*, *walk*, *bus*, or *other* episode. It shows that the model is highly successful in classifying *stop* episodes (98% were predicted correctly), *car* episodes

(98%), and *walk* episodes (94%), than *bus* and *other* episodes (no episodes were predicted correctly).

Table 3.5 Classification table of observed (TUD) versus predicted (GPS) episodes

<i>Observed Episode</i>	<i>Predicted Episode</i>					Row Total
	Stop	Car	Walk	Bus	Other	
Stop	353 (97.8)	1 (0.3)	6 (1.7)	—	1 (0.3)	361
Car	1 (0.3)	303 (98.4)	4 (1.3)	—	—	308
Walk	—	2 (5.7)	33 (94.3)	—	—	35
Bus	—	12 (85.7)	2 (14.3)	—	—	14
Other	—	6 (66.7)	2 (22.2)	1 (11.1)	—	9
Column Total	354 (48.7)	324 (44.6)	47 (6.5)	1 (0.1)	1 (0.1)	727

Values in parentheses indicate row %, highlighted in bold indicate correctly classified; — = zero.

Based on Table 3.5, the calculated kappa statistic is 0.91 ($p < 0.001$), which indicates almost perfect agreement between observed episodes and those predicted by the MNL model (Landis & Koch, 1977). However, the model was ineffective in predicting *bus* and *other* episodes due to the limited samples used in the estimation for these episodes (only 96 and 62 episodes, respectively). The model tends to misclassify *bus* episodes as *car* episodes (86%), and *other* episodes as *stop* episodes (67%). The rest of *bus* (14%) and *other* episodes (22%) were misclassified as *walk* episodes. Moreover, a small number of *walk* episodes (6%) were misclassified as *car* episodes. The misclassifications seem to indicate the huge effect of median speed in favor of *car* episodes over *bus* and *other* episodes; that effect far outweighs the effects of the other two predictors combined (Table 3.4). This suggests that similarity in speeds among episode types, perhaps tempered by noisy GPS readings, makes it difficult to distinguish

one episode type from another (Biljecki, 2010). This also suggests that the model may be sensitive to the noise in median speed values, often favoring faster travel modes over slower ones. We tested the use of a dummy variable for bus route (1 if travel episode traverses a bus route, 0 otherwise) and found it to be very effective in differentiating *bus* episodes from the rest of the episode types. Hence, future developments should consider other predictors commonly derived from transportation networks and other data sources.

Based on the validation results (Table 3.5), the mode detection using a MNL model has an overall accuracy of 90% (adjusted count R^2 ; Long, 1997, p. 108), which shows that the model reduces the errors in prediction by 90%. Reported overall accuracies of existing methods range from 70 to 95% (Biljecki, 2010; Lawson et al., 2010). In this context, the use of MNL in classifying activity episodes looks promising, given a parsimonious model consisting only of three independent variables: median speed (m/s), median change in heading (degrees), and total duration (min).

3.3.3 Performance

The procedures presented were developed and implemented in Python[®], a free scripting language (www.python.org). The scripting language facilitates fast development and allows easy integration with the map-matching algorithm (Dalumpines & Scott, 2011), which was also written in Python[®]. These procedures and algorithms are integral part of GIS-based episode reconstruction toolkit (GERT), which was developed to automatically extract activity episodes from GPS data and derive associated information for each extracted episode.

Running the entire procedure for 5,127 person-days (equivalent to 47 million GPS points) took 24 hours. This can be considered not very fast in terms of algorithm performance but can be considered much faster than manual recording of trip information (normally requiring days to complete). Considering both processing time and overall accuracy, the proposed method performed quite well. However, more testing needs to be done using different datasets, preferably from different geographic areas, to see if similar results are found.

3.4 Conclusion

This article presented a method that automatically extracts activity episodes from GPS data (GERT's MDM), introducing the use of MNL as a classifier to offer an alternative method that is more transferable than existing ones. The use of MNL in classifying GPS episodes can be a reasonable alternative because of its efficiency in differentiating activity episodes using generic variables, objective techniques in determining significant predictors, and limited dependencies in input variables (reliant only on GPS-derived predictors yet flexible to accommodate additional variables). About 36 statistical descriptors were automatically derived from GPS data; median speed, median change in heading, and total duration are found useful in differentiating activity episodes. Model estimates confirmed the typical characteristics of different episodes as distinguished by three selected predictors: median speed, median change in heading, and total duration.

This paper shows that MNL can be easily adopted in the design of algorithms used to extract activity episodes from GPS data, an approach that resulted in significant time savings in algorithm development and provided reasonable predictive accuracy. Efficient methods will require a minimal amount of time for data preparation and in performing the extraction of episodes from GPS data. Lawson et al. (2010) reported that it took them more than two months to implement the rule-based GIS method and two days for neural networks, the proposed method took only about a day in extracting five episode types from about 47 million GPS points. The proposed method achieved an overall accuracy of 90%; however, limited observations for *bus* and *other* episodes resulted in poor classification performance for these two types of episodes. Nevertheless, the use of MNL in classifying activity episodes extracted from GPS data looks promising in terms of methodological transferability and efficiency that the proposed classifier provides.

The results suggest for the refinement of methods in preprocessing of GPS data to minimize noise caused by the inherent limitations of satellite signals and other factors. As GPS technology matures coupled by the increasing popularity of location-based services, better preprocessing will significantly contribute to improvement in predictive accuracy of GPS data mining techniques. The proposed method of extracting and classifying activity episodes from GPS data shows potential but also left some room for improvement. Future work will focus on improving further the performance of the proposed method. Several considerations for improvement in predicted accuracy of the proposed method include: (1) postprocessing of classification that takes account of

transition probability; i.e., considering the episode type probabilities of adjacent segments or episodes, similar to Zheng et al. (2008); (2) using large samples for all episode types, particularly for those travel episodes that are underrepresented, to be modeled by conducting GPS surveys for all episodes of interest; (3) adding relevant explanatory variables or descriptors such as a dummy variable for bus routes; and (4) improving the preprocessing algorithms to minimize errors in speed, heading, and duration values, for example, by using smoothing techniques. Future work should also consider an implementation of the proposed method using GPS data taken at different locations and different temporal resolution.

3.5 Acknowledgements

Financial support for this project was provided by a grant awarded to Darren M. Scott from the Natural Sciences and Engineering Research Council of Canada (261850-2009). The authors acknowledge the Halifax STAR Project for the GPS and time diary data sets used for the development of preprocessing modules.

3.6 References

- Atherton, T. J., & Ben-Akiva, M. E. (1976). Transferability and updating of disaggregate travel demand models. *Transportation Research Record: Journal of the Transportation Research Board*, 610, 12-18.
- Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete choice analysis: theory and application to travel demand*. MIT Press, Cambridge, Mass.

Biljecki, F., 2010. Automatic segmentation and classification of movement trajectories for transportation modes. Unpublished MSc Thesis, Delft University of Technology, Delft, The Netherlands.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285-297.

Bricka, S. (2008). Non-response challenges in GPS-based surveys. Paper presented at the 8th International Conference on Survey Methods in Transport.
<http://www.nustats.com/nustats_dot_com/templates/yes_again_new-menu/docs/great_reads/Nonresponse_GPS_BasedSurveys.pdf>.

Chung, E.-H., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology* 28 (5), 381–401.

Cotrus, A.V., Prashker, J.N., Shiftan, Y., 2005. Spatial and temporal transferability of trip generation demand models in Israel. *Journal of Transportation and Statistics* 8 (1), 37-56.

Dalumpines, R., Scott, D.M., 2011. GIS-based map-matching: development and demonstration of a postprocessing map-matching algorithm for transportation research. In: S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing Geoinformation Science for a Changing World* (Vol. 1, pp. 101-120): Springer Berlin Heidelberg.

Dalumpines, R., Scott, D.M. (2014). GIS-based episode reconstruction toolkit (GERT): a transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data. Manuscript in preparation.

- Dodge, S., Weibel, R., Forootan, E., 2009. Revealing the physics of movement: comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems* 33(6), 419-434.
- Draijer, G., Kalfs, N., Perdok, J., 2000. Global positioning system as data collection method for travel research. *Transportation Research Record: Journal of the Transportation Research Board* 1719, 147–153.
- Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36(2), 131-139.
- Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P.L., Georggi, N.L., Perez, R., 2008. Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones, 15th World Congress on Intelligent Transportation Systems, New York, New York.
- Hausman, J.A., McFadden, D.L., 1984. Specification tests for the multinomial logit model. *Econometrica* 52, 1219-1240.
- Hunt, J.G., Lyons, G.D., 1994. Modeling dual carriageway lane changing using neural networks. *Transportation Research Part C: Emerging Technologies* 2(4), 231–245.
- Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies* 19(3), 387-399.
- Knoblauch, R., Pietrucha, M., Nitzburg, M., 1996. Field studies of pedestrian walking speed and start-up time. *Transportation Research Record: Journal of the Transportation Research Board* 1538, 27-38.

- Koutsopoulos, H.N., Kaptis, V.I., Downey, A.B., 1994. Improved methods for classification of pavement distress images. *Transportation Research Part C: Emerging Technologies* 2(1), 19-33.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- LaPlante, J., Kaeser, T.P., 2007. A history of pedestrian signal walking speed assumptions, 3rd Urban Street Symposium: Uptown, Downtown, or Small Town: Designing Urban Streets That Work.
- Lawson, C.T., Chen, C., Gong, H., 2010. Advanced Applications of Person-based GPS in an Urban Environment, Technical Report, University at Albany, New York. <http://www.utrc2.org/sites/default/files/pubs/advanced-applications-gps1-final_2.pdf>.
- Long, J.S., 1997. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage Publications.
- Maddison, R., Ni Mhurchu, C., 2009. Global positioning system: a new opportunity in physical activity measurement. *International Journal of Behavioral Nutrition and Physical Activity*, 6(1): 73.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (Ed.), *Frontiers in Econometrics*. Academic Press, New York, pp. 105-142.
- Millward, H., Spinney, J., 2011. Time use, travel behavior, and the rural-urban continuum: results from the Halifax STAR project. *Journal of Transport Geography* 19(1), 51-58.

Murakami, E., Wagner, D.P., 1999. Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C*, Vol. 7, pp. 149–165.

O'Brien, R., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* 41(5), 673-690.

Schuessler, N., Axhausen, K.W., 2009. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105, 28-36.

Sorci, M., Antonini, G., Cruz, J., Robin, T., Bierlaire, M., Thiran, J.-P., 2010. Modelling human perception of static facial expressions. *Image and Vision Computing*, 28(5), 790-806.

Stopher, P., Shen, L., 2011. In-depth comparison of global positioning system and diary records. *Transportation Research Record: Journal of the Transportation Research Board*, 2246, 32-37.

Stopher, P.R., Jiang, Q., FitzGerald, C., 2005. Processing GPS data from travel surveys. Presented at 2nd International Colloquium on Behavioral Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications, Toronto, Ontario, Canada, June 2005.

Train, K., 2009. *Discrete choice methods with simulation* (2nd ed.). Cambridge, UK ; New York: Cambridge University Press.

Tsui, A., Shalaby, A., 2006. Enhanced system for link and mode identification for personal travel surveys based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 38–45.

Wilmot, C. (1995). Evidence on transferability of trip-generation models. *Journal of Transportation Engineering*, 121(5), 405-410.

Wolf, J. (2000). Using GPS data loggers to replace travel diaries in the collection of travel data. Unpublished Thesis, Georgia Institute of Technology, Atlanta.

Zheng, Y., Liu, L., Wang, L., Xie, X., 2008. Learning transportation mode from raw GPS data for geographic applications on the web. 17th World Wide Web Conference, Beijing, 21-25 April 2008.

Chapter 4

GIS-based Episode Reconstruction Toolkit (GERT): A Transferable, Modular, and Scalable Framework for Automated Extraction of Activity Episodes from GPS Data

4.1 Introduction

Person-based global positioning system (GPS) devices capture the start and end times of activity episodes, their duration, and travel routes in greater spatial and temporal resolution than traditional recall-based surveys. Because not all episode information can be directly captured by these devices, GPS is increasingly used to supplement traditional survey methods primarily to increase the accuracy of data collection and reduce burden among respondents (Wolf, 2000; Bricka, 2008; Millward and Spinney, 2011). Several procedures have been developed to extract information from person-based GPS data in order to supplement data from recall-based surveys (e.g., Stopher et al., 2005; Chung and Shalaby, 2005; Tsui and Shalaby, 2006; Zheng et al., 2008; Schuessler and Axhausen, 2009; Bohte and Maat, 2009). However, most of these procedures suffer from specific data requirements and complexity that limit their transferability to other application environments. Further, they have a limited set of modules to extract all necessary information such as route attributes, and were not specifically designed to handle huge GPS datasets (Schuessler and Axhausen, 2009; Biljecki, 2010; Lawson et al., 2010). To deal effectively with these problems, this paper presents a framework based on three

design principles (transferability, modularity, and scalability), and a GIS-based toolkit (based on this framework) for automated extraction of activity episodes from GPS data. Without an effective framework and a toolkit to implement this framework, we cannot take full advantage of GPS data for the following reasons: (1) difficult to adopt tools developed by other researchers for lack of transferability, (2) limited ability to derive more information from GPS data for lack of an integrated set of modules, and (3) high computational costs and lack of automatic procedures in dealing with huge datasets. Before elaborating on these issues (lack of transferability, huge GPS data, incomplete set of tools), it is best to clarify some key terms used in this paper.

In the context of activity analysis, a person's 24-hour (daily) activities can be subdivided into episodes, which are differentiated based on location (inside a building or travel mode); hence an activity episode can be a stationary episode (*stop* episode) or a travel episode (e.g., *car* episode). Travel episodes are synonymous to *trips*. In this paper, a segment refers to a sequence of GPS points similarly classified as *stop* or *trip* points (Figure 4.1), while an episode refers to the diary-equivalent classification of a segment based on location (e.g., *stop* episode, *walk* episode, *car* episode, and so on) with attributes such as start time, end time, duration, and distance. In different contexts, some researchers used the term "stages" (Schuessler and Axhausen, 2009), "objects" (Dodge et al., 2009), and "segments" (Zheng et al. 2008; Gong et al., 2012) to refer to episodes.

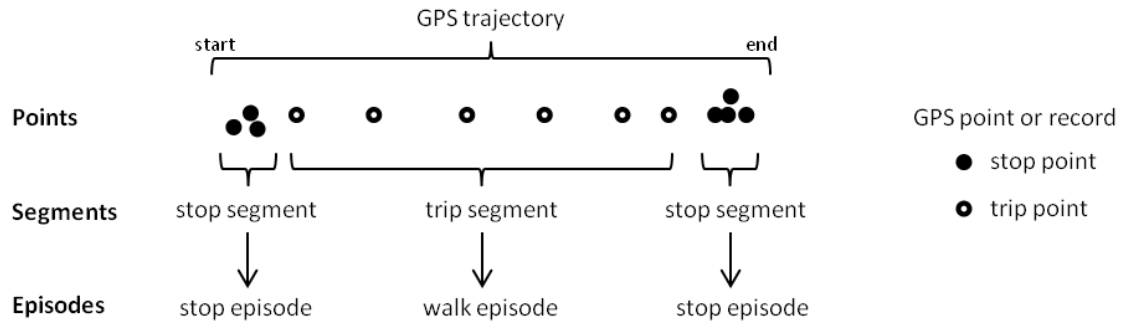


Figure 4.1 GPS trajectory subdivided into points, segments, and episodes

4.1.1 Challenges in developing tools and methods for extracting episodes from GPS data

Existing procedures in extracting or reconstructing episodes from GPS data can be categorized into several modules: preprocessing (data filtering and smoothing), extraction of episodes (stages or segments), mode detection (assignment of mode to travel episode), route detection (map-matching), and purpose detection. To the authors' knowledge, original attempts to automate the extraction of episodes in the transportation literature (Chung and Shalaby, 2005; Stopher et al., 2005; Schuessler and Axhausen, 2009) suggest the lack of transferability of existing modules (e.g., non-generic variables used in preprocessing), an incomplete set of modules (e.g., no modules for purpose detection in two studies while not fully automatic for route detection), and only one study was specifically designed for large GPS data (Schuessler and Axhausen, 2009) (for related studies in other fields, see Biljecki (2010) and Bolbol et al. (2012)). Only a few studies (Schuessler and Axhausen, 2009; Bohte and Maat, 2009) specifically addressed the

development of tools and methods in extracting travel episodes and trip purposes from large-scale GPS datasets.

4.1.1.1 Lack of transferability

Existing methods used unique inputs or variables to filter valid points for extracting and classifying episodes. Because procedures are often designed to make the most out of the data available, the decision rules vary depending on the characteristics of the input data. For example, some researchers (e.g., Wolf et al., 2000; Stopher et al., 2005; Chung and Shalaby, 2005) used the number of satellites, heading, and horizontal dilution of precision (HDOP) in a preprocessing module to remove outliers and invalid GPS points; in the absence of the above inputs, Schuessler and Axhausen (2009) instead used the known altitude of Switzerland to remove low quality or erroneous GPS points. Other researchers (e.g., Chung and Shalaby, 2005; Bohte and Maat, 2009; Gong et al., 2012) used proximity measures (e.g., distances to bus, subway, and railway stations) to determine probable travel modes; however, threshold distances vary significantly among studies. For rule-based procedures, there is absence of clear guidelines in finding optimum threshold values as implemented in various modules because often the cut-off values are based on subjective judgment and the quality of GPS data.

Researchers have focused more on the effectiveness of their tools and methods, and little attention had been given to the transferability of the tools and methods. Eventually each researcher has developed a unique set of tools using unique inputs for specific purposes. Transferability of the existing methods remains a challenge. This issue

had been identified by Lawson et al. (2010) based on their experiment that aims to replicate currently used methods on mode detection. Their experiment provides an objective examination of currently used mode detection algorithms to date. They recognized the difficulty of directly comparing different approaches because of different data used in developing these methods, not to mention the different variables required by each approach. At this juncture, the existing methods are mostly not generic, making it difficult to apply the same methods in different environments with minimal effort.

4.1.1.2 Processing demands of huge GPS data

Current personal-based GPS devices are becoming widespread because of better accuracy, more portability (pocket-size), and lower costs. In recent years, we observed an increase in large-scale GPS data used for travel episode (trip) extraction: more than 20,000 km of GPS trajectories (Zheng et al., 2008); 64.5 million GPS points (Schuessler and Axhausen, 2009); 17.6 million GPS points (Bohte and Maat, 2009; Biljecki, 2010); 47.3 million GPS points (Millward and Spinney, 2011); and perhaps more. Manual procedures are no longer practical in dealing with huge GPS data that span millions of records (Schuessler and Axhausen, 2009). The availability of huge GPS data and the high potential to collect more make it necessary to come up with efficient procedures that can automatically extract information from these data.

4.1.1.3 *Incomplete set of tools*

From the perspective of activity analysis, most of the existing methods did not fully capture valuable information from GPS trajectories for these methods were focused more on mode detection, that is, the extraction of travel episodes and classifying these episodes into several types based on travel modes (Stopher et al., 2005; Schuessler and Axhausen, 2009; Gong et al., 2012). Hence no modules were specifically developed to extract information associated with activity locations (*stop* episodes), wherein more information can be extracted with the aid of additional data such as land use and potential activity locations (PAL), and information on observed routes (road attributes) connecting these locations. Although Stopher et al. (2005) and Bohte and Maat (2009) also tried to capture trip purposes (the former asked respondents to provide addresses of activities while the latter used points of interest and GPS-derived trip endpoints), both lack modules to *automatically* capture more information from detailed land use; information on land use and points of interest can be used to automatically classify *stop* episodes and assign trip purposes to travel episodes.

Traditional survey methods fell short of providing data of good quality for route choice modeling. GPS-assisted surveys can fill this gap; however automatic postprocessing is required. The current practice of extracting routes (map-matching) is a tedious process of tracing the routes manually from GPS trajectories in a GIS (e.g., Ramming, 2002; Papinski et al., 2009; Winters et al., 2010) or directly asking respondents through web questionnaires (e.g., Kaplan and Prato, 2012). While some researchers had

incorporated route detection using map-matching routines (e.g., Chung and Shalaby, 2005; Tsui and Shalaby, 2006), the map-matching lacks integration with route choice set generation algorithms (e.g., Prato and Bekhor, 2006) to automatically generate alternative routes for route choice modeling. In addition, existing modules that automatically generate route attributes for map-matched routes (e.g., Papinski and Scott, 2011) are not integrated with route detection and route choice set generation modules. To the authors' knowledge, no toolkit has been developed that integrates all the above modules, and at the same time include extra modules that automatically generate route choice sets and route attributes from GPS trajectories.

4.1.2 Addressing challenges through GIS-based episode reconstruction toolkit (GERT)

A framework needs to be developed that explicitly addresses the above issues (lack of transferability, an incomplete set of tools, and computational demands of huge GPS data). Therefore, this paper presents a GIS-based episode reconstruction toolkit (GERT), which was built on the framework of three design principles (modularity, transferability, and scalability), to automatically extract activity episodes from GPS data with or without additional information. The development of the toolkit was motivated by the need to provide data to support route choice modeling in particular and activity analysis in general (Ortúzar and Olszewski, 2009). The data generated by the toolkit can be used to supplement time-use diary (TUD) data. In addition, the generated data can be used for other applications such as understanding activity patterns over time and space (dynamics), and finding determinants of route choice behavior. The toolkit reduces the

burden placed on respondents to record their routes. Through a validation experiment, we found that GERT's core modules work properly in reconstructing episodes from GPS data.

The rest of this paper is organized as follows: the next section presents GERT's components, starting with a functional overview of all components and followed by a detailed discussion of its main modules. To find out if GERT works well in reconstructing episodes from GPS data, a validation experiment and its results are discussed next. The concluding section highlights how GERT's framework addressed the main challenges faced by existing procedures, as well as GERT's potential in supporting analysis and model estimation tasks. Future research directions are also discussed.

4.2 GIS-based episode reconstruction toolkit (GERT)

We developed GERT using a framework that specifically deals with lack of transferability, an incomplete set of tools, and demands for automatic procedures associated with huge GPS data. This section further explains GERT's framework and presents the main components of the toolkit.

4.2.1 GERT's framework: transferability, modularity, and scalability

Inspired by best practices in software development, GERT's framework (Figure 4.2) was developed around three design principles: transferability, modularity, and scalability.

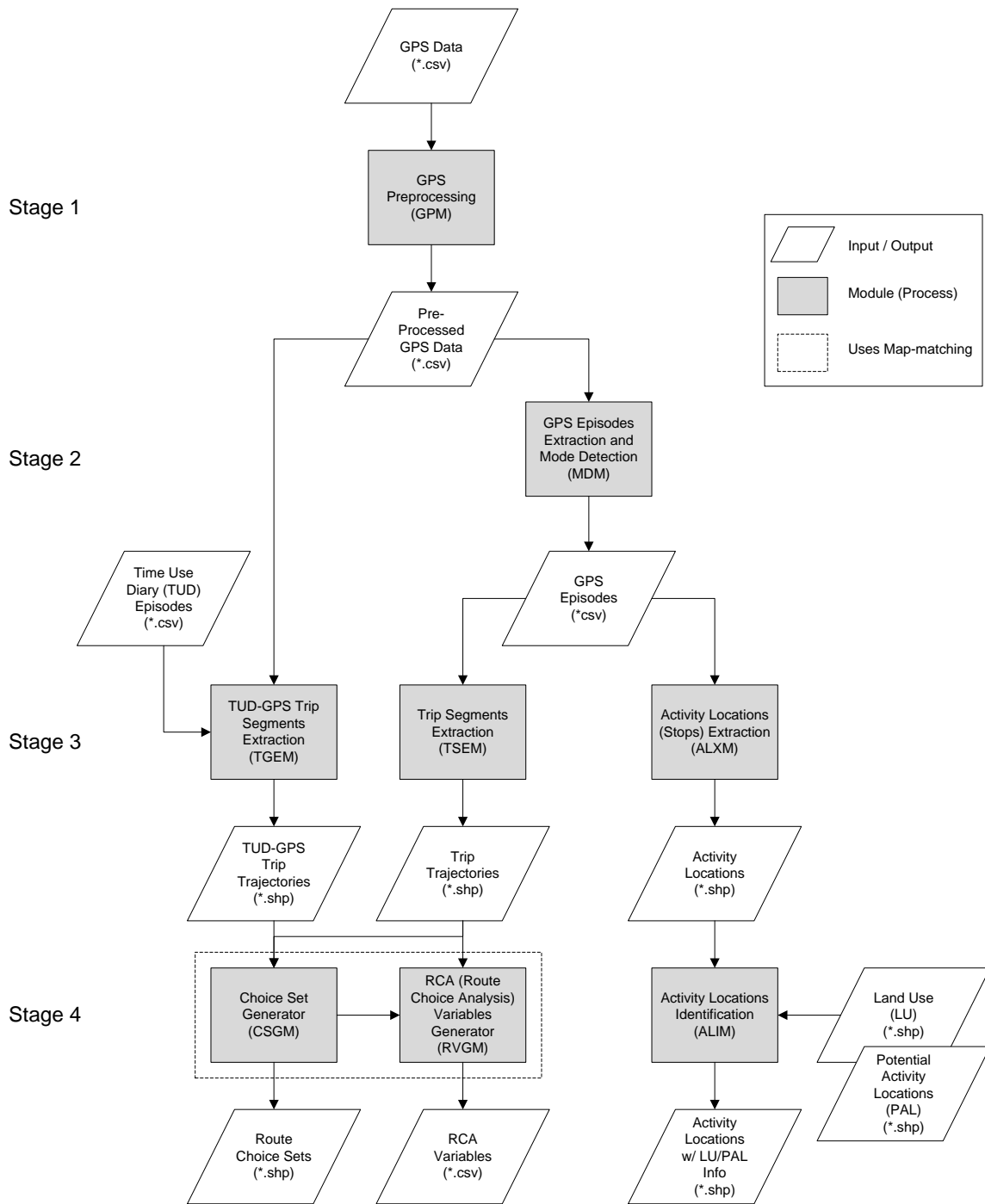


Figure 4.2 Four-stage workflow of the GIS-based episode reconstruction toolkit (GERT)

In order to be transferable, GERT was designed to work on minimal inputs – generic variables such as location coordinates and time stamps. This feature allows implementation in environments without additional information aside from GPS data. Other related features are efficiency and flexibility brought by the mode detection module (MDM), which involves multinomial logit estimation of episode types using standard statistical software outside of GERT, and the resulting model can be brought in for episode classification. Also, GERT generates outputs in comma-separated values (CSV) or shapefiles (SHP). Both are cross-platform data formats for spreadsheet and GIS applications, respectively – making it easy to manipulate GERT’s outputs in other applications.

GERT was also developed into separate but interconnected modules, which makes it easy to update existing modules and add new ones without affecting its integrity. For example, the MDM uses a multinomial logit (MNL) model to automatically classify extracted episodes from GPS data into different types. New modules can be added providing for a different logic for MDM implementation (e.g., fuzzy logic, neural networks).

In terms of scalability, GERT was designed to process huge GPS data and can accommodate additional data such as time-use diary, land use, and potential activity locations to enrich GERT’s extracted episodes for exploratory data analysis, route choice modeling, and so on.

4.2.2 GERT's modules: from GPS data preprocessing to route choice data generation

Software design partitions a program into several levels of detail wherein the entire system is organized into sub-systems (more abstract or higher level, referred to here as stages), the sub-systems or stages are further divided into classes (intermediate level), and the classes are divided into routines and data (lower level, where decision rules are implemented); for more information, see McConnell, 2004, pp. 82-87. As a toolkit for GPS data mining, GERT can be seen in a higher design level as a sequence of workflows in four stages (Figure 4.2), each stage producing data for subsequent stages or other processes outside of GERT. At the intermediate level, each stage is further divided into one or more main modules, which in turn are composed of many separate but interconnected sub-modules at the lower design level.

Stage 1 consists of the GPS Preprocessing Module (GPM), a module that removes invalid points from raw GPS data in comma-separated values (CSV) format using data cleaning procedures adopted from previous studies (e.g., Schuessler and Axhausen, 2009; Marchal et al., 2011). Generally, invalid points include redundant points (points with the same coordinates) and outliers (with speed ≥ 50 m/s). GPM uses several algorithms to filter valid trajectories and these algorithms were written in several sub-modules with each sub-module focusing on certain aspects of GPS preprocessing. Preprocessing was organized into two core processes: clustering and segmentation. Clustering divides GPS points into 24-hour trajectories, which represent sequences of points for each person-day. It then further divides each trajectory into sequential clusters of adjacent points based on

speed, distance, heading, and change-in-heading thresholds. After clustering, segmentation uses *point-segment classification routines (PSCR)* to tag each point in the valid trajectory as a stop (stationary) point or a trip (moving) point. PSCR are iterative and sequential sub-processes that classify each point based on its characteristics (i.e., distance, duration, speed, heading, and change in heading), the characteristics of surrounding points (i.e., points before and after), the characteristics of the segment where it belongs (a segment is a sequence of similarly tagged points), and the characteristics of surrounding segments. The next three stages use GPM's output for further processing, an output written in CSV format. GPM's CSV output contains original input fields such as latitude, longitude, time, and so on; and added new fields such as distance (m), duration (s), speed (m/s), heading (degrees), change in heading (degrees), and status (stop or trip point) – useful information for other processes outside of GERT.

Stage 2 uses the GPS Episodes Extraction and Mode Detection Module, in short the Mode Detection Module (MDM), which partitions valid GPS trajectories produced by GPM into segments, then classifies these segments into stationary activity episodes (stop episodes) or travel episodes such as walk, car, bus, and so on. At the core of MDM is a module that classifies GPS segments into different types of episodes using an estimated MNL model based on observed episodes (TUD-reported episodes with GPS trajectories). To support MNL model estimation, several utility modules were also developed to extract statistical descriptors from observed episodes. These statistical descriptors provide information on the signatures of different episode types of interest based on descriptive

statistics of distance, duration, speed, heading, change in heading, and acceleration of GPS segments. The utility modules extract segments from GPS data based on corresponding time-use diary (TUD) episodes, and generate statistical descriptors based on these segments. A MNL model can be estimated based on these sample segments using statistical software and the resulting model can be fed to the classifier module. Using the estimated MNL model parameters, MDM generates a CSV output of classified episodes (called GPS episodes) from GPS trajectories. Attached to each classified episode is the following information: episode ID, episode number, episode type (i.e., stop, car, walk, and so on), mode probability, date, start time, end time, duration, distance, and lat-lon coordinates (if stop episode). Aside from being an input to succeeding stages, GPS episodes can be used for analytical processes outside of GERT that use number of episodes (stop versus travel), their duration, and distances (if travel episodes). For example, actual number of trips used in transportation studies can be derived from the GPS episodes. More details on the development and testing of MDM, with particular emphasis on the transferability of MNL as a classifier, are presented in another paper by the authors (Dalumpines and Scott, 2014b).

Stage 3 consists of three main modules: TUD-GPS Segments Extraction Module (TGEM), Trip Segments Extraction Module (TSEM), and Activity Locations Extraction Module (ALXM). TGEM and TSEM extract trip segments (sequences of GPS points in a travel episode) while ALXM extracts activity locations (stops or endpoints of trip segments) from valid GPS trajectories. TGEM extracts trip segments directly from GPS

trajectories using start and end times of TUD episodes; *the module skips MDM in the process*. TSEM, on the other hand, relies on the classified episodes generated by the MDM. This makes GERT flexible; it can handle GPS data with or without TUD episodes. Stage 3 generates outputs as point or multi-point shapefiles (SHP), a standard format for spatial data across different GIS platforms. Aside from being used as input in the final stage, Stage 3 outputs can be used to visualize activity episodes in a GIS or can be used as inputs to spatial analysis and modeling outside of GERT.

Finally, Stage 4 has been inspired by the need to develop tools to support route choice analysis (RCA). This final stage consists of three main modules: Choice Set Generator Module (CSGM), RCA Variables Generator (RVGM), and Activity Locations Identification Module (ALIM). CSGM and RVGM generate data for route choice modeling; both modules use the map-matching algorithm (Dalumpines and Scott, 2011) that snaps walk and non-walk trip segments to a digital road/pedestrian network to derive the actual travel (observed) routes. CSGM uses a modified potential path area - gateway algorithm (PPAG) to generate alternative routes for trip segments produced by TGEM or TSEM. The PPAG algorithm defines a potential path area based on route travel time or distance and uses random gateways (links) within the area to generate alternative routes. Alternative routes are included in route choice sets if they pass a set of criteria (overlap factor, loop factor, distance factor) adopted from the branch-and-bound algorithm (Prato and Bekhor, 2006). RVGM is an improved version of Papinski and Scott's (2011) RCA toolkit that generates over 50 variables based on road network attributes such as distance,

time, turns, speed, and so on, and those that can be derived from these attributes such as the route directness index (RDI). The improved version fully automates the extraction of RCA variables, correctly measures route length (original version used nearest nodes as route endpoints), adds new variables such as number of intersections and overlap statistics (with other alternatives in a choice set), and makes it compatible with the latest version of ArcGIS®.

In the absence of TUD episodes, CSGM can also use TSEM's output. In turn, CSGM's output can be used as input to RVGM to generate variables for each alternative route. For route choice modeling, CSGM is used first to generate route choice sets in SHP format, where each choice set (containing alternative and actual routes) is stored in a separate folder; then RVGM uses these choice sets to generate variables for each route in a CSV format. RVGM is coupled with a module that calculates route overlap statistics such as percent of route length that overlaps with other routes in a choice set. Overlap statistics variables can be used as inputs to Path-Size Logit (PSL), a modified form of MNL that uses a correction factor in the deterministic part of utility function (Bekhor and Prato, 2009). The outputs of CSGM and RVGM can also support other research applications aside from RCA. For example, CSGM and RVGM can be used to extract and describe observed routes and their shortest paths to determine factors that influence deviation from shortest paths – an investigation of 'route efficiency' (Papinski and Scott, 2013). More details on the design and application of CSGM are presented in another paper by the authors (Dalumpines and Scott, 2014a).

The other main module in Stage 4 is ALIM, which appends additional information, if available, to extracted activity locations and generates output in SHP format. This additional information can be generated from spatial data such as land use and potential activity locations (PAL) using overlay analysis functions in GIS. PAL refers to points of interest that indicate locations of government offices, shopping destinations, banks, and so on. ALIM enriches ALXM's output, making it useful for activity analysis. In the future, another module can be added to GERT to automatically classify activity locations generated in Stage 2 based on land use and PAL information.

Figure 4.3 shows an example of the outputs generated by GERT's modules.

4.3 Data and experimental design for validation

In this section, we describe the TUD and GPS data used in GERT's development and validation. Also, the experimental design is presented to assess the performance of GERT's key modules (GPM and MDM) in reconstructing episodes from GPS data. The purpose of validation is to evaluate the effectiveness of GERT's ability to extract episodes from GPS data, without additional information.

4.3.1 STAR dataset

The development and validation of GERT's algorithms used GPS data from the Space-Time Activity Research (STAR) project. TUD episodes from the STAR project were used as the basis to validate GERT's ability to extract episodes from GPS data.





GERT module	CSV output	Shapefile output
GPM	<pre> gpstid ... time distance heading dursec speed deltahead status 15308757 ... 4:00:50 8.7 165.6 15691 0.90 111.5 stop 15316278 ... 8:23:49 15.6 333.6 88 0.18 137.3 trip 15316281 ... 8:23:57 15.0 163.7 8 1.88 137.3 trip 15316287 ... 8:24:12 9.7 38.3 15 0.65 125.3 trip </pre>	Can be generated, if needed.
MDM	<pre> episode mode likelihood start_time end_time dursec distmeters latitude longitude 1 stop 1.00 4:00:50 8:22:21 15691 8.7 44.64700 -63.47423 2 walk 0.99 8:22:21 8:24:12 111 40.4 na na 3 car 1.00 8:24:12 8:28:54 282 2349.2 na na 4 stop 0.81 8:28:54 8:32:01 187 8.5 44.65789 -63.49093 </pre>	Not applicable.
TGEM / TSEM	<pre> episode mode likelihood start_time end_time dursec distmeters 2 walk 0.99 8:22:21 8:24:12 111 40.4 3 car 1.00 8:24:12 8:28:54 282 2349.2 5 car 1.00 8:32:01 8:36:49 288 2412.0 6 walk 0.98 8:36:49 8:43:07 378 92.5 8 car 1.00 8:45:56 8:52:04 368 4154.1 9 walk 0.99 8:52:04 8:55:09 185 20.1 11 car 1.00 11:06:55 11:11:48 293 3430.8 </pre>	
ALXM	<pre> episode mode likelihood start_time end_time dursec latitude longitude 1 stop 1.00 4:00:50 8:22:21 15691 44.64700 -63.47423 4 stop 0.81 8:28:54 8:32:01 187 44.65789 -63.49093 7 stop 0.94 8:43:07 8:45:56 169 44.66760 -63.48920 10 stop 1.00 8:55:09 11:06:55 7906 44.64699 -63.47430 14 stop 0.96 11:24:44 11:36:52 728 44.65427 -63.48014 </pre>	
CSGM	Depends on road network data.	
RVGM	<pre> routeid ... crowdist turns totdist tottime avgspd crossing pathsize ... ~alternative_6350736 ... 1065.2 6 1684.0 1.9 54.67 14 0.607 ... ~alternative_shortestpath ... 1065.2 2 1341.1 1.4 58.18 10 0.393 ... ~observed_route ... 1065.2 2 1532.6 1.6 57.33 14 0.524 ... </pre>	Can be linked to CSGM's output.
ALIM	<pre> episode lu_match lu_code lu_classification pal_match pal_id ... 1 YES 110 Tax - Residential NO 0 ... 4 YES 110 EMO Sites YES 206083581 ... 7 YES 235 Church/EMO Site YES 255772956 ... 10 YES 110 Tax - Residential NO 0 ... 14 YES 220 Public School/HRM Parks NO 0 ... </pre>	

Figure 4.3 Example of GERT's CSV and shapefile outputs

The STAR project is a comprehensive survey of time use and travel activity conducted in Halifax, Nova Scotia, Canada from April 2007 to May 2008 (Millward and Spinney, 2011). Apart from a time-use diary, respondents carried a GPS-equipped mobile device (Hewlett Packard iPAQ hw6955), which records a location every second and with a horizontal accuracy ≤ 10 m. The GPS data logger collected positional data that included unit ID, date, time, x-coordinate, x-direction (north/south), y-coordinate, y-direction (east/west), speed, altitude, horizontal dilution of precision (HDOP), and the number of satellites. Around 2,000 respondents collected 47.3 million points for two survey days (equivalent to 5,127 person-days), and a total of 108,529 TUD episodes. Of particular interest, each TUD episode has information on the activity location such as home, workplace, car, bus, and so on.

4.3.2 Experimental design for validation

The original GPS data from the STAR project were stored as a single file in SPSS[®] format. This single file was converted into a comma delimited (*.csv) format for use with GERT. Next, GERT was used to extract episodes based on an estimated MNL model (Dalumpines and Scott, 2014b); 25,707 episodes were extracted then further processed using GERT's complete set of tools. About 31% (8,052) of these extracted episodes were found to match those of TUDs, equivalent to 653 person-days (567 respondents). Since most of TUD person-days lasted 20 hours (04:00-11:59), their corresponding GPS person-day trajectories were included in the validation set *if* these person-day trajectories lasted for at least 18 hours (10% lower than TUD person-day's

duration); otherwise, these person-day trajectories were considered incomplete and discarded.

Since not all of the GPS data had equivalent time-use diaries, the GPS data have been processed to match TUDs using GERT's helper modules (part of GERT's package used for converting inputs into a format used by the toolkit or in preparing data for validation). In this way, episodes derived by GERT from GPS data can be compared with TUD episodes. A helper module was used to convert TUD episodes into a format comparable with that of GPS episodes, converting stationary activity locations such as home into stop episodes and mobile activity locations such as car into travel episodes; similar adjacent episodes with the same location were merged (e.g., several activity episodes done at home were merged as a stop episode). These procedures reduced original TUD episodes to 57,775. Another helper module flagged TUD and GPS episodes as a match when the difference in their start and end times did not exceed 30 minutes as suggested by Stopher and Shen (2011). The same authors recommended some strategies for *detailed* comparison and will be considered in the future as part of GERT's performance tuning.

As extensive episode-to-episode comparison is a detailed investigation in itself, this part is also planned for future extension of this paper. Consequently, this paper followed an aggregate comparison methodology as used by Schuessler and Axhausen (2009) to assess, at the general level, GERT's ability to extract episodes (i.e., involving GPM and MDM, which are the key components of the toolkit).

4.4 Results and discussion

About 26,000 episodes were automatically reconstructed using GERT from 47.3 million GPS points collected by the STAR project. To validate that GERT works properly at the aggregate level (i.e., in terms of episode and duration distributions), time-use diary (TUD) and GPS episodes were matched based on two thresholds: difference between start and end times must not exceed 30 minutes, and minimum daily duration of 18 hours. The resulting validation set was used in comparing TUD and GPS episodes at the aggregate level as discussed in this section. After that, GERT's computational performance is assessed to provide a rough assessment of its scalability.

4.4.1 Comparison of TUD and GPS episodes

The validation set consisted of 391 person-days, equivalent to 10,315 episodes. Of these episodes, 53% (5,494) were reported in TUD, and 47% (4,821) were extracted from GPS data using GERT. The number of GPS episodes is 12% (1,343) lower than that of TUD. This difference may be attributed to cases where respondents left their GPS devices at home for some parts of the day. For these cases, GERT correctly captures the stationary activity episode at home but failed to match several travel episodes reported by respondents in TUD.

In Figure 4.4, the distribution of episodes generated by GERT is compared to that of TUD. In terms of episode types (Figure 4.4a), GPS have more *walk* episodes while

TUD seems to dominate in the rest of episode types, which indicates that GERT works properly in detecting short travel episodes often not reported in TUD.

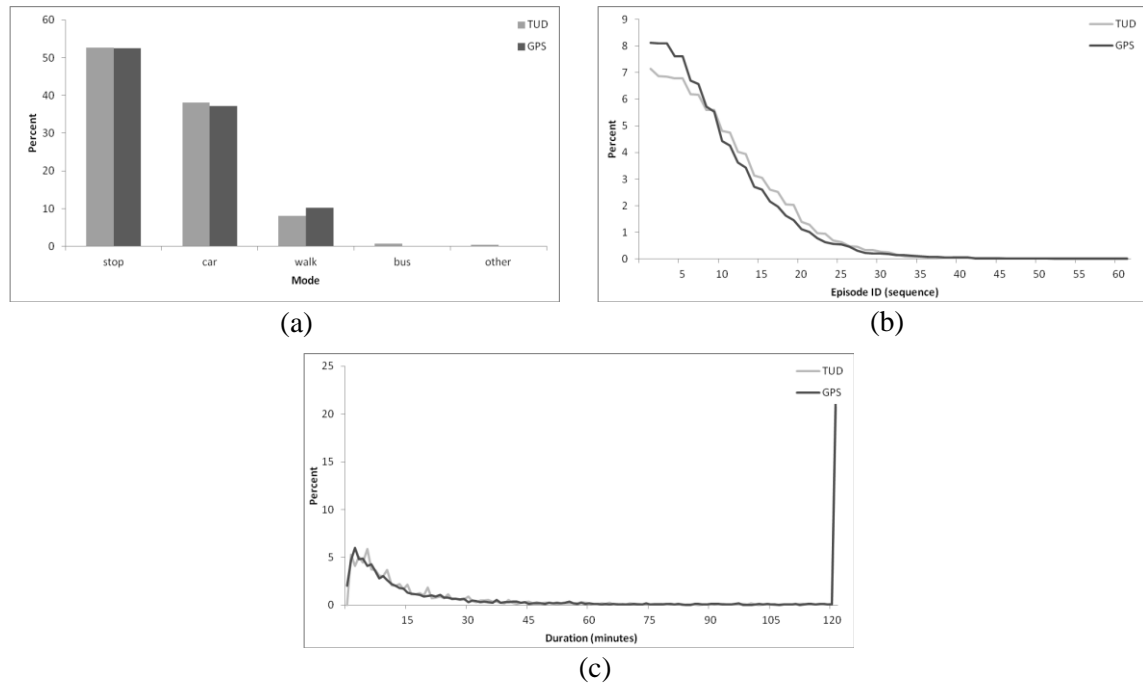


Figure 4.4 Episode distribution, TUD versus GPS: (a) by episode type, (b) by episode ID (sequence or order of episode in a person-day; for example, at episode ID 5 the figure shows that GERT generated more 5th episodes than reported in TUD), and (c) by duration (minutes)

GERT’s ability to detect short travel episodes is also reflected in Figure 4.4b, which shows the distribution of episodes based on their sequence or order in a person-day. This figure indicates that GPS dominates in the lower-range sequence (approximately in the first 8th episodes, where the 1st episode typically represents maintenance activity at home); consequently, TUD dominates in the middle-range sequence (10th-31st episodes) probably because some episodes were misclassified by

GERT's MDM as *stop* episodes. Figure 4.4c displays similar distributions for TUD and GPS episodes (all episodes longer than 2 h have been summarized in the last category).

The distribution of duration per episode type is shown in Figure 4.5. Again all episodes longer than 2 h have been summarized in the last category of each distribution. In general, the distributions reveal similar patterns in TUD and GPS episodes, with the exception of *bus* and *other* (travel) episodes (41 *bus* episodes reported but only one was detected in GPS; for 26 *other* travel episodes, only two were detected). The MNL model used by MDM in classifying episodes (or imputing travel modes) was not able to differentiate *bus* and *other* (travel) episodes from the rest. This was because of the small samples for these two types of episodes used in MNL model estimated by STAR data. Future work will consider experiments to enhance GERT's MDM performance by using a balanced sample of episode types, and by including additional variables in model specification – another test for GERT's transferability and scalability features.

Figure 4.5 also highlights some differences between TUD and GPS episodes. GERT's activity extraction modules (GPM/MDM) generated more *stop* episodes (stationary activities) longer than 2 h (Figure 4.5a). This is perhaps due to GPM's *point-segment classification routines (PSCR)* that merged very short-duration travel episodes with *stop* episodes if located between *stop* episodes. This situation further requires analysis in the future to determine effective thresholds for these routines. Figure 4.5b shows that GERT generated more short-duration *car* episodes than those reported in TUD – attributed to short-duration episodes with high speed.

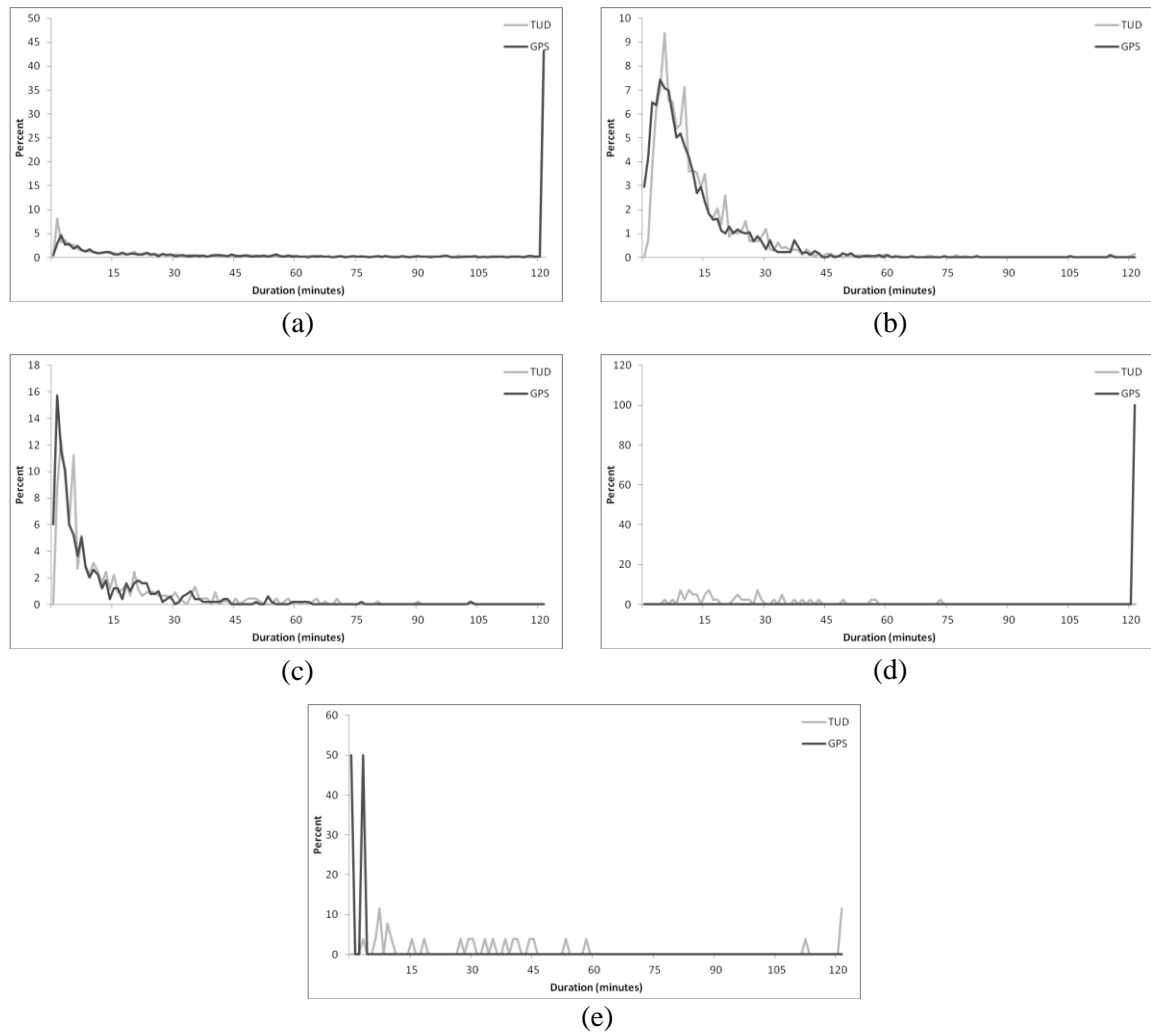


Figure 4.5 Duration distribution per episode type, TUD versus GPS: (a) stop episodes, (b) car episodes, (c) walk episodes, (d) bus episodes, and (e) other episodes

As expected, GERT was able to generate more short-duration *walk* episodes that were not reported in TUD. This ability is reinforced by a finding that about 18% of reasons trips were recorded only by GPS device was “short trip and respondent did not realize that it should be reported” (Stopher and Shen, 2011, p. 36). This situation

highlights the advantage of GPS over recall-based surveys with respect to temporal and spatial accuracy.

Figure 4.6 shows the distributions of daily travel episodes (trips) and GPS-derived distances by travel mode. Trips that occurred more than 18 times per day or longer than 20 km are summarized in the last category for the respective distributions.

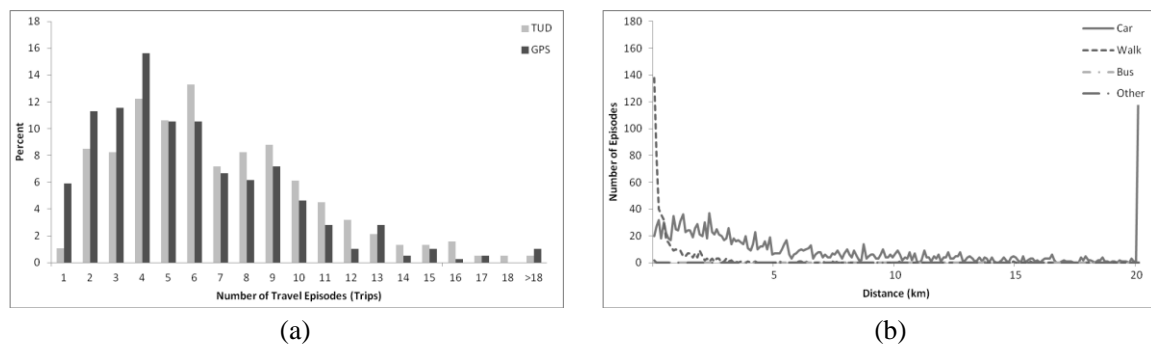


Figure 4.6 Distribution of travel episodes: (a) daily travel episodes (trips), TUD versus GPS; and (b) GPS-derived distances by travel mode

TUD and GPS have similar distributions in terms of the number of trips per day (Figure 4.6a); both are skewed to the right with GPS episodes having a slightly longer tail. GPS episodes dominate in the categories of less than 5 trips per day, which reveals that GERT tends to misclassify some trips as stop episodes hence reducing the number of trips per day. In addition, GERT tends to be effective in detecting short trips that resulted to more cases of > 18 trips per day. Figure 4.6b shows the distribution of GPS-derived distances by travel mode. No comparison was made between TUD and GPS episodes in terms of distances because distances were not available for TUD episodes. However, the distribution of GPS-derived distances reveals typical patterns of the different travel

modes; *walk* episodes are short while *car* and other episodes that use motorized modes cover longer distances (*car* episodes were most dominant).

In summary, the episode and duration distributions reveal similar patterns between TUD and GPS episodes, a similarity that confirms that GERT's components (GPM and MDM) work properly, at the aggregate level, in reconstructing episodes from GPS data. This also means that GERT can be used to reconstruct episodes from a different GPS data, with some calibration of the MNL model that will be used by MDM. Some issues were identified for improvements in the future such as GERT's MDM misclassification of travel episodes as *stop* episodes, and weak detection of underrepresented episode types (i.e., *bus* and *other* travel episodes).

4.4.2 Computational performance

Table 4.1 shows the average performances of GERT's main modules using a desktop PC (Win 7 Pro 64-bit, i7 CPU @ 3.4GHz, 16GB RAM). Preprocessing of 47.3 million GPS points took an average of 6,000 s or about 7,887 points per second. This is slower compared to the reported performance of 9,100 points per second, preprocessing plus mode detection for 64.5 million points using Debian Linux Version 2.618-5-amd64, four dual-core CPUs @ 2GHz, 4GB RAM (Schuessler and Axhausen, 2009). GPM's slower performance may be attributed to its point-segment classification routines to filter out invalid points, particularly caused by multi-path and signal blocking errors in urban environments. This is a tradeoff to the transferability of GPM, and of GERT in general; the use of generic variables offers transferability but requires complicated routines to

filter out invalid points effectively. Overall, GERT's modules tend to have linear running times, processing time increases with the number of inputs (GPS points, stop episodes, travel episodes). RVGM's performance may somewhat vary depending on the density of the road network; more links in the network increases the number of alternative routes, hence increases processing time.

Table 4.1 Average computational performance of GERT's main modules

Module	Function	Average Performance ^a
GPM	Preprocessing (remove invalid points)	1 s per person-day (7,887 points per second) ^b
MDM	Extract episodes and detect modes	0.5 s per episode (2 episodes per second)
TGEM/TSEM	Extract travel episodes and convert to multi-point shapefiles	4 s per travel episode (trip)
ALXM	Extract stop episodes and convert to point shapefiles	486 stop episodes (activity locations) per second
CSGM	Generate route choice set for each travel episode	50 s per travel episode (trip) ^c
RVGM	Generate route attributes	8 s per travel episode (trip)
ALIM	Append information from land use and potential activity locations	2 s per stop episode (activity location)

^a Desktop PC (Win 7 Pro 64-bit, i7 CPU @ 3.4GHz, 16GB RAM).

^b 5,127 person-days, about 47.3 million points.

^c An average of 6 routes (including observed route) per route choice set.

From the standpoint of the three challenges faced by existing methods (huge data, incomplete set of tools, and lack of transferability), GERT's overall performance suggests potential because of its scalability – GERT can scale up to large GPS data (aside from its ability to accommodate additional information); modularity – GERT has a complete set of tools to support analyses and model estimations; and transferability – GERT's reliance on

generic variables (latitude, longitude, time) makes it applicable to many environments with added flexibility to make use of additional information unique to each environment. For instance, one can fully appreciate the potential of GERT in generating data for route choice modeling using thousands of observations; a difficult task, impractical if done manually, but made a lot easier using GERT's modules. A demonstration of this task is described in another paper by the authors (Dalumpines and Scott, 2014a), an article that determines the underlying route choice decisions for shopping and work trips using a sample of 1,462 observed routes.

4.5 Conclusion

Existing methods of extracting episodes from person-based GPS data faced three main challenges: lack of transferability, an incomplete set of tools, and computational demands of huge GPS data. This paper presented a GIS-based episode reconstruction toolkit (GERT) to address these challenges using a framework built around three design principles: transferability, modularity, and scalability. Transferability guided the use of generic variables (latitude, longitude, time) and practical procedures (MNL as classifier for mode detection) that makes GERT transferable to many environments. Modularity allowed the development of an interrelated set of modules that provided a complete set of functionalities (CSGM, RVGM, and ALIM), extensions not currently available in existing toolkits. Scalability refers to GERT's ability to process huge volume of GPS data (e.g., about 127 s in preprocessing a million points) and can accommodate additional

information to enrich GPS-derived data (e.g., land use and points of interest for activity identification of *stop* episodes).

As a complete toolkit, GERT has many modules that cannot be entirely covered for validation in this paper. Hence a validation experiment was only conducted on GERT's ability to extract episodes from GPS data (i.e., GPM and MDM components), as extracted episodes are required for most of GERT's other modules. Since TUD episodes from STAR project underwent good data quality review (Millward and Spinney, 2011), we assumed that TUD episodes correctly represent activities performed by survey respondents. A comparison of the episode and duration distributions reveal similar patterns between TUD and GPS episodes, a similarity that confirms that GERT's components (GPM and MDM) work properly, at the aggregate level, in reconstructing episodes from GPS data. GERT's components, GPM and MDM, made possible the automatic extraction of *stop* episodes, which covers information on stationary activities such as location (latitude, longitude), start time, end time, and duration. With additional data such as land use and potential activity locations (PAL), ALXM and ALIM can automatically attach more information on activity locations. To supplement TUD reporting, GERT's extracted *stop* episodes can be used to determine missing or unreported locations of stationary activities and test the framework proposed by Horner et al. (2012) in the reconstruction of activity destinations. Extracted *stop* episodes can also be used for exploratory spatial data analysis to gain insights from the spatio-temporal distribution of activity locations, and destination choice modeling. In addition to *stop*

episodes, TGEM/TSEM can automatically generate travel segments (GPS trajectories of travel episodes); CSGM generates route choice sets from travel segments and RVGM attaches route attributes. GERT makes it easy to generate inputs for route choice modeling, which for many years has relied on disjointed, and often, manual procedures (e.g., Winters et al., 2010; Kaplan and Prato, 2012). Also, GERT's RVGM can generate observed routes and their corresponding shortest paths that can be used to analyze route choice efficiency (e.g., Papinski and Scott, 2013). Overall, GERT's modules provide transportation researchers with rich datasets (i.e., *stop* and travel episodes, activity locations, travel segments, route choice sets, route attributes) for improving our understanding of activity/travel patterns in general and route choice decisions in particular.

Further research is encouraged to improve the accuracy and applicability of the toolkit. TUDs often suffer from underreporting of short travel episodes (Stopher and Shen, 2011), mostly *walk* episodes. Episodes extracted by GERT from GPS data can be used to identify these unreported episodes. Moreover, extra care is needed when validating the accuracy of the toolkit using TUD because of the inaccurate reporting of short travel episodes. This can be addressed in the future by conducting controlled experiments that ensure accurate reporting of episodes and varied deployment of GPS devices to capture enough samples for different travel modes, giving emphasis on underrepresented modes (e.g., bike, bus, and so on). Consequently, an extensive episode-to-episode validation can be performed along the lines of the framework suggested by

Stopher and Shen (2011). To enhance GERT's MDM, future work should consider the implementation of a routine for a higher threshold value for dwell time together with a sensitivity analysis (e.g., Schuessler and Axhausen, 2009), use a feedback loop from the map-matching algorithm to provide information on the bus/train network, and employ an episode transition probability matrix (e.g., Zheng et al., 2008). Future work may also focus on the new module to extend ALIM's functionality (e.g., Huang et al., 2010), an activity profiler that automatically classifies *stop* episodes into different activity types (e.g., home maintenance, work, leisure, and so on) based on the information harvested by ALIM from land use and points of interest. A thorough performance testing of all the modules would require a lot of time; this will be considered in future plans for fine tuning of GERT's performance. GERT was developed using data captured by person-based GPS devices with high temporal resolution (at least one reading per second) and horizontal accuracy of 10 m or better. Although we assumed that GERT will work with GPS data of low temporal resolution since speed and other thresholds may still hold, it would be interesting to test this assumption in the future and compare GERT's performance with other studies (e.g., Bolbol et al., 2012).

4.6 Acknowledgements

Financial support for this project was provided by a grant awarded to Darren M. Scott from the Natural Sciences and Engineering Research Council of Canada (261850-2009). The authors acknowledge the Halifax STAR project for providing the data for

testing and developing the GIS-based episode reconstruction toolkit; also, Nadine Schuessler and Kay Axhausen for sharing their code on data filtering and smoothing that inspired development of GERT's GPM.

4.7 References

Bekhor, S., Prato, C.G., 2009. Methodological transferability in route choice modeling. *Transportation Research Part B: Methodological* 43 (4), 422-437.

Biljecki, F., 2010. Automatic segmentation and classification of movement trajectories for transportation modes. Master of Science in Geomatics MSc Thesis, Delft University of Technology, Delft, The Netherlands.

Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems* 36 (6), 526-537.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies* 17 (3), 285-297.

Bricka, S., 2008. Non-response challenges in GPS-based surveys. Paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, France.

Available from:

<http://www.isctsc.let.fr/papiers/resourcepaper%20%20final%20version/A2%20RP%20bricka.doc>. [21 May 2014].

Chung, E.H., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology* 28 (5), 381 - 401.

Dalumpines, R., Scott, D.M., 2011. GIS-based map-matching: development and demonstration of a postprocessing map-matching algorithm for transportation research, in: Geertman, S., Reinhardt, W., Toppen, F. (Eds.), *Advancing Geoinformation Science for a Changing World Vol. 1*, Springer, Berlin, pp. 101-120.

Dalumpines, R., Scott, D.M., 2014a. Determinants of route choice behavior: a comparison of shop versus work trips using the potential path area - gateway (PPAG) algorithm and path-size logit. Manuscript in preparation.

Dalumpines, R., Scott, D.M., 2014b. Making mode detection transferable: extracting activity and travel episodes from GPS data using the multinomial logit model and Python. Manuscript in preparation.

Dodge, S., Weibel, R., Forootan, E., 2009. Revealing the physics of movement: comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems* 33 (6), 419-434.

Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36 (2), 131-139.

Horner, M.W., Zook, B., Downs, J.A., 2012. Where were you? Development of a time-geographic approach for activity destination re-construction. *Computers, Environment and Urban Systems* 36 (6), 488-499.

Huang, L., Li, Q., Yue, Y., 2010. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. Paper presented at the Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, San Jose, California.

Kaplan, S., Prato, C.G., 2012. Closing the gap between behavior and models in route choice: the role of spatiotemporal constraints and latent traits in choice set formation. *Transportation Research Part F: Traffic Psychology and Behaviour* 15 (1), 9-24.

Lawson, C.T., Chen, C., Gong, H., 2010. Advanced applications of person-based GPS in an urban environment (final report). New York: University at Albany. Available from: <http://www.utrc2.org/sites/default/files/pubs/advanced-applications-gps1-final_2.pdf>. [21 May 2014].

Marchal, P., Madre, J.L., Yuan, S., 2011. Postprocessing procedures for person-based global positioning system data collected in the French National Travel Survey 2007-2008. *Transportation Research Record: Journal of the Transportation Research Board* 2246, 47-54.

McConnell, S., 2004. *Code complete*, 2nd ed. Microsoft Press, Redmond, Washington.

Millward, H., Spinney, J., 2011. Time use, travel behavior, and the rural-urban continuum: results from the Halifax STAR project. *Journal of Transport Geography* 19 (1), 51-58.

Ortúzar, J.D., Olszewski, P., 2009. Advances in data acquisition, in: Kitamura, R., Yoshii, T., Yamamoto, T. (Eds.), *The expanding sphere of travel behavior research: selected papers from the 11th Conference of the International Association for Travel Behavior Research*, Emerald Group Publishing, Bingley, UK, pp. 447-455.

Papinski, D., Scott, D.M., 2011. A GIS-based toolkit for route choice analysis. *Journal of Transport Geography* 19 (3), 434-442.

Papinski, D., Scott, D.M., 2013. Route choice efficiency: an investigation of home-to-work trips using GPS data. *Environment and Planning A* 45 (2), 263-275.

- Papinski, D., Scott, D.M., Doherty, S.T., 2009. Exploring the route choice decision-making process: a comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour* 12 (4), 347-358.
- Prato, C. G., Bekhor, S., 2006. Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board* 1985, 19-28.
- Ramming, S., 2002. Network knowledge and route choice. Ph. D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.
- Schuessler, N., Axhausen, K., 2009. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board* 2105, 28-36.
- Stopher, P., Jiang, Q., FitzGerald, C., 2005. Processing GPS data from travel surveys. Paper presented at the 2nd International Colloquium on Behavioral Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications, Toronto, Ontario, Canada.
- Stopher, P., Shen, L., 2011. In-depth comparison of global positioning system and diary records. *Transportation Research Record: Journal of the Transportation Research Board* 2246, 32-37.
- Tsui, S., Shalaby, A., 2006. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board* 1972, 38-45.

Winters, M., Teschke, K., Grant, M., Setton, E., Brauer, M., 2010. How far out of the way will we travel? *Transportation Research Record: Journal of the Transportation Research Board* 2190, 1-10.

Wolf, J., 2000. Using GPS data loggers to replace travel diaries in the collection of travel data. Ph.D. Thesis, Georgia Institute of Technology, Atlanta.

Zheng, Y., Liu, L., Wang, L., Xie, X., 2008. Learning transportation mode from raw GPS data for geographic applications on the web. Paper presented at the 17th World Wide Web Conference, Beijing, China.

Chapter 5

Determinants of Route Choice Behavior: A Comparison of Shop versus Work Trips Using the Potential Path Area – Gateway (PPAG)

Algorithm and Path-Size Logit

5.1 Introduction

Route choice behavior is a complex spatial behavior influenced by route attributes and individual characteristics (e.g., Carpenter, 1979; Ben-Akiva et al., 1984; Zhang and Levinson, 2008). The complexity of this behavior arises from the two main challenges in route choice modeling: the very large number of alternative routes, and the prevalent overlapping of these routes. To deal with these challenges, considerable efforts have been made toward the development of path generation methods and route choice models. However, most of these efforts have focused on route choice behavior related to driving from home to work (work trips) and less attention has been given to other types of trips (e.g., shop trips). Shop trips account for a substantial portion of total trips and its growing share suggest that research should not only focus on work trips, but also on shop trips (Nelson and Niles, 2000; Hanson, 2004). To put this issue into proper perspective, this paper seeks to test whether route choice behavior varies for work and shop trips. Failure to consider such systematic variation may result in bias and over/under prediction of the relative effects of policy relevant variables on route alternatives and route choice (Ben-Akiva and Morikawa, 1990; Bradley and Daly, 1994). For this reason, it has been

suggested that such systematic variation should be incorporated in route choice model specification (Ben-Akiva et al., 1984; Zhang and Levinson, 2008).

Significant progress has been achieved in recent years in addressing the challenges of route choice modeling. Many studies have developed techniques for generating alternative routes, classified as deterministic shortest path-based methods, stochastic shortest path-based techniques, constrained enumeration algorithms, and probabilistic approaches (see Prato (2009) for a review of these algorithms). Most of these algorithms lack a sound behavioral basis (specified based on the experience and knowledge of the analyst) and computing efficiency (often accounts for the universal set of route alternatives not a subset of feasible routes). The use of behavioral rules generates a more realistic set of routes than shortest path-based methods (Prato and Bekhor, 2006; Papinski, 2010). The potential path area (PPA) approach is an implementation of the PPA concept to constrain the set of possible routes. Papinski (2010) found that PPA-generated choice sets improve model performance compared to k shortest paths.

To deal with the issue of route overlaps, route choice modeling has relied on a number of discrete choice models (Cascetta et al., 1996; Ben-Akiva and Bierlaire, 1999; Prashker and Bekhor, 2004; Frejinger and Bierlaire, 2007). The Multinomial Logit (MNL) route choice models have been used in investigating the determinants of route choice behavior (Zhang and Levinson, 2008), and was successfully applied in a number of network models (Wen et al., 2006). However, it was discovered that the inherent limitation of the MNL route choice model resulted in over prediction of the route choice

probabilities for overlapping links (e.g., expressways), which led to unrealistic traffic volume estimates along these links (Ben-Akiva et al., 2012). This limitation of MNL is its inability to account for similarities among route alternatives. Researchers address this problem by abandoning the MNL formulation to consider the correlations of overlapping paths explicitly (Bekhor et al., 2002; Frejinger and Bierlaire, 2007), or modify the MNL formulation by adding a correction term in the deterministic part of the utility (Cascetta et al., 1996; Ben-Akiva and Bierlaire, 1999; Bekhor and Prato, 2009). The latter approach benefits from the closed-form structure of MNL formulation, and is more tractable than probit and other GEV models that treat correlations explicitly. Moreover, the modified MNL formulations have been shown to be computationally efficient (Bekhor and Prato, 2009).

Despite these two streams of development, *little attention* has been given to the determinants of route choice behavior related to other trip purposes aside from work. These pressing developments appear to have overshadowed efforts in determining the factors that affect route choice behavior for different trip purposes, and the relative importance of these factors in this varied context (Ben-Akiva et al., 1984; Zhang and Levinson, 2008). Over the past decade, researchers have largely focused on analyzing route choice behavior for work trips (Ramming, 2002; Frejinger and Bierlaire, 2007; Bekhor and Prato, 2009; Kaplan and Prato, 2012). Also, the lack of input data adds to the difficulty in expanding the scope of route choice modeling from work trips to other trip purposes. Traditional route assignment models apply the same shortest path search, often

based on travel time, regardless of trip purpose and consequently oversimplify underlying decision processes behind route choice behavior. To avoid this oversimplification, route choice model specification should account for the systematic variation associated with trip purpose.

In this paper, we argue that route choice behavior varies by trip purpose and this systematic variation should be considered in route choice model specification. This argument entails two objectives: show that the utility and scale parameters for separate models of work and shop trips differ; if so, highlight the contrast in route choice behavior between work and shop trips by considering the interaction of route attributes and individual characteristics. In the process, we introduce a practical path generation algorithm that generates feasible route choice sets for route choice modeling. Following the suggestion of Ben-Akiva et al. (1984), we used a limited set of route attributes, commonly derived from transportation network data. For the route choice model, we used the PSL model because it is computationally efficient (Bekhor and Prato, 2009) and meets the assumption of a scaling estimation procedure. The sequential scaling estimation method (Swait and Louviere, 1993; Louviere et al., 2000), along with the likelihood ratio test, are used to test whether the utility and scale parameters are different in *work* and *shop* route datasets, given that a PSL model underlies both datasets (*work* dataset is used as the reference set). In general, the scaling estimation process follows the approach of combining revealed and stated preference data within the context of discrete choice

analysis (Ben-Akiva and Morikawa, 1990; Swait and Louviere, 1993; Adamowicz et al., 1994; Hensher et al., 1998).

This paper is the first, to the authors' knowledge, to compare route choice models of work and shop trips, and include the interaction of route attributes and individual characteristics in the model specification. Also, this paper is the first to use the scaling estimation method, normally applied in the estimation of revealed and stated preference data, to test whether utility and scale parameters differ in the separate models of work and shop trips. Another unique contribution of this paper is the development of an algorithm used to automatically generate alternatives for each observed route. We call this algorithm PPAG, which is short for the combination of potential path area (PPA) and gateway shortest path algorithms (GSP). The PPA is a construct in time geography (Hägerstrand, 1970) that defines the possible area that an individual can travel within a time budget, anchored around the individual's trip origin and destination. In this paper, the time budget refers to the reported travel time from origin to destination. The GSP algorithm creates an alternative route by forcing a shortest path between origin and destination to pass through a specified link or gateway (Lombard and Church, 1993). The PPA algorithm provides the constrained area wherein the GSP algorithm can extract feasible routes.

The remainder of this paper is organized as follows. Section 5.2 describes the generation of route choice data from a global positioning system (GPS)-assisted time-use survey using the PPAG algorithm, along with other geographic information system (GIS)-

based Episode Reconstruction Toolkit (GERT) modules for route choice data generation. Also, in this section, the PSL model specification is presented together with the scale factor estimation and likelihood ratio test. Section 5.3 presents the results in terms of the descriptive statistics of route attributes and selected individual characteristics, the determinants of route choice behavior as indicated by PSL model estimates, and the outcome of the likelihood ratio tests. Finally, Section 5.4 summarizes the major findings of this study and discusses future research directions.

5.2 Data and methods

To test the hypothesis that route choice preferences vary by trip purpose, we compare a route choice model for work trips with that of shop trips. In this section, we describe the data source for observed routes and introduce the path generation algorithm used to generate alternative routes. Then, we present the rationale for the use of PSL in route choice modeling, and describe the scale factor estimation and likelihood ratio tests – methods used to test the inequality of utility parameters and scale factors between models of work and shop trips.

5.2.1 Space-Time Activity Research (STAR) data

Observed or actual route choices were derived from reported work and shop trips of 708 respondents to the Space-Time Activity Research (STAR) project in Halifax, Nova Scotia, Canada. The STAR project was a time use survey that employed GPS devices to geo-reference respondent locations for two days between April 2007 and May 2008

(Millward and Spinney, 2011). Start and end times of drive to work and shop trips were used to extract travel trajectories (sequences of GPS points representing trip segments) from the STAR GPS data. We used the GIS-based Episode Reconstruction Toolkit (GERT) module on trip segment extraction to extract work and shop travel trajectories (Dalumpines and Scott, 2014). These trajectories were then input into a map-matching algorithm (Dalumpines and Scott, 2011) to extract observed routes. In turn, the observed routes were used to generate the route choice data using a route choice set generation algorithm (discussed in Section 5.2.2), and route attributes generator: all these processes were implemented through GERT's set of tools for route choice data generation. 1,462 observed routes were extracted using GERT's modules: 45% (653) for work trips and 55% (809) for shop trips.

11% (81) of the 708 respondents were included in both *work* and *shop* datasets (Table 5.1). As expected, the sample for the *work* dataset was dominantly young compared to the *shop* dataset, where roughly a third (26%) were older drivers (age > 64 years). In terms of other socio-demographic characteristics (e.g., sex, household size, personal income), both datasets are comparable. However, the two datasets differ in terms of residential and travel characteristics. For example, a large majority (81%) of the samples for the *shop* dataset had experience with public transit while only half had experience for the *work* dataset. About three out of four respondents drove to work for at least five minutes, while the ratio is even for shop trips.

Table 5.1 Individual characteristics used with route attributes to create interaction terms

Variable	Category (value)	Sample percentage ^a	
		Work (<i>n</i> = 398)	Shop (<i>n</i> = 391)
<i>Socio-demographic characteristics</i>			
Age	Other [15 to 64 years] (0)	99.1	73.6
	65 years and above (1)	0.9	26.4
Sex	Female (0)	48.6	51.2
	Male (1)	51.4	48.8
Household size	Other (0)	92.3	92.2
	≥5 (1)	7.7	7.8
Personal income	Other (0)	91.5	96.2
	C\$80,000-\$99,999 (1)	8.5	3.8
<i>Residential and travel characteristics</i>			
Frequency of public transit use	Other (0)	52.4	80.6
	Never (1) ^b	47.6	19.4
Travel time	≥ 5 minutes (0)	77.4	49.6
	< 5 minutes (1)	22.6	50.4
Residence tenure	Other (0)	40.8	29.1
	10 years and over (1)	59.2	70.9

^a Based on the number of respondents; 81 respondents were in both samples.

^b Excludes those who have not used public transit because of unavailability of transit service.

For the *shop* dataset, about 71% lived in their neighborhood for at least 10 years, while only 59% for the *work* dataset had the same length of residence. The aforementioned individual characteristics were selected because they were known to influence route choice efficiency (Papinski and Scott, 2013). These individual characteristics were incorporated into the PSL models as dummy variables. Other individual characteristics were not included because of missing data or no answers in time use diary.

The road network for the province of Nova Scotia consists of 98,132 nodes, 116,647 links, and serves a land area of around 55,000 sq. km. The study area contains

expressways as well as local roads. This network was based on the comprehensive network dataset from a geospatial data provider (DMTI CanMap[®] RouteLogistics Version 2008.3 Release; www.dmtispatial.com).

5.2.2. Path generation using the Potential Path Area - Gateway (PPAG) algorithm

Given the collection of 1,462 observed or actual routes taken by respondents, we used the PPAG algorithm, a core component of GERT's Choice Set Generator Module (CSGM), to generate route choice sets (Dalumpines and Scott, 2014). In general, the PPAG algorithm defines a PPA based on route travel time or distance and uses random gateways (links) within the PPA to generate alternative routes. In the context of path or route generation, a PPA represents an area that encloses all traversable links to reach a destination within allowable time (or distance); hence the PPA provides a sound theoretical basis, often lacking in most path generation algorithms, to constrain the selection of alternatives from the universal set.

In generating route choice sets, firstly PPAG creates a list of all traversable links given a maximum travel time or distance using service area analysis in ArcGIS[®]. PPAG automatically derives travel time or distance from the observed route, the path generated by the map-matching algorithm (Dalumpines and Scott, 2011) from the GPS trip segment. Secondly, PPAG generates alternative routes based on the observed route's endpoints (origin and destination) using a modified gateway shortest path algorithm (GSP) and the list of traversable links (created in the first step) as gateways. The modified GSP algorithm addresses the main drawbacks of the original (Lombard and Church, 1993) in

the following ways: (i) it uses a parser function that detects and discards an alternative if a loop exists (the alternative goes back over a link after reaching the gateway link), (ii) we minimized the possibility of missing alternatives through the PPA and exhaustive gateway enumeration, and (iii) alternatives are included in route choice sets if they pass a set of criteria or selection parameters. In this study, we used the following selection parameters, adopted from Prato and Bekhor (2006): the distance factor is 1.10, the loop factor is 1.50, the overlap factor is 0.80, and the maximum number of left turns is 4. These parameters were found to be effective as behavioral constraints in determining relevant route alternatives (Prato and Bekhor, 2006). The distance factor excludes routes that exceed the distance of observed route by 10 percent; this constraint rejects alternatives that require drivers to traverse routes considerably longer than their regular route. The loop factor, sometimes called the route directness index (Papinski and Scott, 2011), is based on the ratio of observed route distance over straight-line distance. It discards routes that drivers are likely to avoid because of many detours. The overlap factor removes routes with a high degree of overlap that drivers would not consider as separate alternatives.

Finally, CSGM stores each observed route and its alternatives (generated by PPAG as shapefiles) in separate folders. The RCA Variables Generator (RVGM), based on the earlier work of Papinski and Scott (2011), treats each folder generated by CSGM as a route choice set. For each route choice set, RVGM automatically processes individual routes and generates over 50 explanatory variables based on road network

attributes (Dalumpines & Scott, 2014). In addition, RVGM automatically calculates the path size, a variable that measures the degree of route overlaps. The path size variable was incorporated in the logit model and is explained in Section 5.2.3.

5.2.3. Specification of the Path-Size Logit (PSL) model

Given the choice sets generated for each observed route of the *work* and *shop* datasets, we estimated route choice models that account for the correlation structure among the alternatives in the deterministic part of the utility. Cascetta et al. (1996) were the first to introduce a correction term in a modified MNL to reduce systematic utility because of route overlaps, which they referred as the commonality factor (CF). Motivated by the lack of theoretical guidance for the CF term, Ben-Akiva and Bierlaire (1999) suggested a path size (PS) attribute instead of the CF term to correct for the overlapping routes. Between these two formulations, the PS formulation has been shown to perform better (Ramming 2002; Prato and Bekhor, 2006, 2007) and is adopted in this paper. With route length measurement assumed to be more reliable than travel time, the PS is further defined as in Bekhor and Prato (2009) as:

$$PS_k = \sum_{a \in \Gamma_k} \frac{L_a}{L_k} \frac{1}{\sum_{l \in C_n} \delta_{al}}, \quad (5.1)$$

where Γ_k is the set of links in route k , L_a is the length of link a and L_k is the length of link k , C_n is the choice set of routes generated for observation n , and δ_{al} is the link-route incidence dummy, which equals one if link a is part of route l and zero otherwise.

There are several versions of the PS formulation (e.g., Ben-Akiva and Bierlaire, 1999; Ramming, 2002); however, it was shown that the original formulation (5.1) provided intuitive results and has a theoretical motivation (Frejinger and Bierlaire, 2007). Therefore, the original formulation was adopted in this paper. The path size has values in the following range: $0 < PS \leq 1$. Hence, a unique route in the choice set (with no link overlaps) has a path size of 1, while a route with partial overlaps has a path size of less than 1. The final model, commonly known as the Path-Size Logit model (PSL), takes the following form (Bekhor and Prato, 2009; Ben-Akiva et al., 2012):

$$P_k = \frac{\exp(V_k + \ln PS_k)}{\sum_{l \in C_n} \exp(V_l + \ln PS_l)}, \quad (5.2)$$

where P_k is the probability of choosing route k , C_n is as previously defined, and V_k and V_l are the deterministic utilities of routes k and l , respectively. Equation (5.2) indicates that the systematic or deterministic utility for route k is adjusted by the $\ln PS$, where $-\infty < \ln PS \leq 0$. For a completely unique route, there is no adjustment to deterministic utility ($\ln PS = 0$). Otherwise, the deterministic utility is reduced because of link overlaps among feasible routes in the choice set.

We estimated PSL models for separate *work* and *shop* datasets, and combined datasets using generic route attributes. Another set of model estimations included individual characteristics to interact with route attributes to investigate whether route choice determinants for work trips differ in relative importance to that of shop trips. *Prior*

to these estimations that compare work against shop trips, we estimated corresponding MNL models to confirm that PSL models result in better model fit. We used the PSL models for *work* and *shop* datasets to test whether the coefficients of route choice determinants and the scale parameter are the same for work and shop trips, given that PSL model underlies both datasets. The test mentioned is discussed in Section 5.2.4.

5.2.4 Scale factor estimation and likelihood ratio tests

The estimation of PSL models for work and shop trips could result in different estimates due to differences in scale factors, utility parameters, or both. In this study, we view scale (variance) as an integral feature of route choice behavior rather than a nuisance parameter, following previous authors (Swait and Louviere, 1993; Adamowicz et al., 1994; Bradley and Daly, 1994). Hence, if datasets for work and shop trips cannot be combined due to unequal parameter vectors and scale factors, this inequality implies that the route choice preference for work trips differs from that of shop trips. To ascertain whether there is a significant difference in route choice preferences in terms of trip purposes (work versus shop), we test whether the coefficients of route choice determinants and scale factor (i.e. error variance) are equal for work and shop trips. Since the stochastic part of the utility is independent and identically Gumbel distributed (Ben-Akiva and Lerman, 1985, pp. 104-105), the PSL as shown in equation (5.2) can be rewritten as:

$$P_k = \frac{\exp[\mu(V_k + \ln PS_k)]}{\sum_{l \in C_n} \exp[\mu(V_l + \ln PS_l)]}, \quad (5.3)$$

where μ is the scale parameter for a particular dataset (μ is often normalized to unity when dealing with single dataset hence not included in equation (5.2)). In this study, we consider identical route attributes for two different datasets that represent work trips (WT) and shop trips (ST). Hence, we are interested if the scales μ_{WT} and μ_{ST} are equal and, if not, whether the parameters β_{WT} and β_{ST} differ after accounting for differences in scale. Scaling estimation follows the standard practice of combining revealed and stated preference data by pooling them under the hypothesis of equal utility parameter vectors, while controlling for the scale parameters (Ben-Akiva and Morikawa, 1990; Swait and Louviere, 1993; Adamowicz et al., 1994; Hensher et al., 1998). This procedure has been used to test the methodological transferability of path generation algorithms and model parameters in the context of route choice modeling (Bekhor and Prato, 2009). For our purposes, we used the sequential scaling estimation method (Swait and Louviere, 1993; Louviere et al., 2000) over the simultaneous scaling estimation (Ben-Akiva and Morikawa, 1990; Bradley and Daly, 1994) for two reasons: (i) we are more interested on the likelihood ratio tests rather than the estimates from the pooled datasets, and (ii) independent variables are the same for the two datasets. Moreover, Louviere et al. (2000) has shown that sequential scaling estimation produced very close estimates to that generated by simultaneous scaling estimation.

Combining the two route choice data sets allows us to estimate β , δ (PS coefficient), and μ_{ST} (*relative scale* with respect to work trips), given the vector X of observable route attributes common to *work* and *shop* datasets. Given the route choice probabilities as defined in equation (5.3), β , δ , and μ_{ST} are obtained by maximizing the following log likelihood function:

$$L_{RS} = \sum_{n \in WT} \sum_{k \in C_n^{WT}} y_{kn} \ln \left(\frac{\exp[\mu_{WT}(\beta X_k^{WT} + \delta \ln PS_k^{WT})]}{\sum_{l \in C_n^{WT}} \exp[\mu_{WT}(\beta X_l^{WT} + \delta \ln PS_l^{WT})]} \right) + \sum_{n \in ST} \sum_{k \in C_n^{ST}} y_{kn} \ln \left(\frac{\exp[\mu_{ST}(\beta X_k^{ST} + \delta \ln PS_k^{ST})]}{\sum_{l \in C_n^{ST}} \exp[\mu_{ST}(\beta X_l^{ST} + \delta \ln PS_l^{ST})]} \right), \quad (5.4)$$

where $y_{an} = 1$ if individual n chooses route k , otherwise $y_{an} = 0$. To maximize L_{RS} , the sequential scaling method graphs the log likelihood function for the combined datasets as a function of μ_{ST} under the hypothesis of equal utility parameter vectors. Since PSL has the same functional form as MNL, the global concavity of the log likelihood function of the MNL also applies in this case, which implies that there is a unique maximum. For detailed implementation of the sequential scaling method, see Swait and Louviere (1993) or Louviere et al. (2000, pp. 237-240).

Following Swait and Louviere (1993), we tested whether work and shop trips share the same utility and scale parameters by means of the following hypothesis:

$$H_1 : \beta_{WT} = \beta_{ST} \text{ and } \mu_{WT} = \mu_{ST}. \quad (5.5)$$

Firstly, we tested whether the common observed route attributes have the same parameters in *work* and *shop* datasets, that is, β_{WT} and β_{ST} are equal,

$$H_{1A} : \beta_{WT} = \beta_{ST} = \beta, \quad (5.6)$$

while allowing the scale factors to differ between the two datasets ($\mu_{WT} \neq \mu_{ST}$). Secondly, if H_{1A} is rejected, H_1 is also rejected. If H_{1A} cannot be rejected, then we test the hypothesis,

$$H_{1B} : \mu_{WT} = \mu_{ST} = \mu. \quad (5.7)$$

H_{1A} and H_{1B} can be tested by standard likelihood ratio statistics. To test whether H_{1A} can be rejected, we used the likelihood ratio test statistic,

$$\lambda_A = -2[L_{RS} - (L_{WT} + L_{ST})], \quad (5.8)$$

where L_{RS} is the log likelihood value in (5.4), which corresponds to the model estimated using the combined *work* and *shop* datasets, where μ_{WT} is normalized to unity and the *relative scale* factor μ_{ST} is estimated using the sequential scaling approach (Swait and Louviere, 1993; Louviere et al., 2000). L_{WT} is the log likelihood value corresponding to a separate model for work trips and L_{ST} the corresponding value of a separate model for shop trips. This test statistic is asymptotically chi-squared distributed with $(K + 1)$ degrees of freedom, where K is the number of parameters in each of β , β_{WT} , and β_{ST} .

The additional degree of freedom accounts for the μ_{ST} allowed to vary under the alternative hypothesis (5.6). If H_{1A} cannot be rejected, then H_{1B} is tested using the following test statistic,

$$\lambda_B = -2[L_{ES} - L_{RS}], \quad (5.9)$$

where L_{ES} is the log likelihood value for the model estimated using the combined *work* and *shop* datasets, where the scale factors are set to be *equal* under H_{1B} (5.7). This statistic is asymptotically chi-squared distributed with one degree of freedom because of the restriction on μ_{ST} .

5.3 Results and discussion

We generated route choice data from the STAR project as described in Section 5.2. The route choice data were used in the estimation of PSL models for work and shop trips. In this section, we discuss the differences of PSL models for work and shop trips and the results of likelihood ratio tests based on these models to provide evidence that utility parameters and scale factors differ in both trip purposes. We also discuss PSL models that include interaction terms to emphasize that relative importance of route choice determinants varies by trip purpose.

5.3.1 Descriptive statistics of work and shop routes generated by the PPAG algorithm

Table 5.2 shows the descriptive statistics of some variables generated by RVGM (see Section 5.2.2) for work and shop routes, based on the route choice sets generated by

Table 5.2 Selected route attribute statistics for work routes and shop routes

Variables	Work routes ^a (mean ± std.)	Shop routes ^b (mean ± std.)
Number of unique roads	11.2 ± 4.6	9.1 ± 4.1
Route travel time (min)	12.1 ± 9.9	6.9 ± 5.6
Route distance (m)	13,553 ± 12,876	6,964 ± 6,332
Route directness index (route distance over straight-line distance)	1.47 ± 0.45	1.51 ± 0.50
Link count	13.6 ± 5.7	10.8 ± 5.1
Longest leg distance (m)	5,130 ± 6,196	2,733 ± 3,122
Longest leg time (min)	4.1 ± 4.3	2.4 ± 2.6
Number of intersections	84.1 ± 54.0	51.6 ± 39.3
Path size	0.43 ± 0.17	0.46 ± 0.17
<i>Turn statistics</i>		
Left turns	2.8 ± 1.4	2.3 ± 1.4
Right turns	2.9 ± 1.9	2.6 ± 1.7
Sharp left turns	0.3 ± 0.6	0.3 ± 0.6
Sharp right turns	0.3 ± 0.6	0.3 ± 0.5
Total turns	6.3 ± 3.1	5.5 ± 2.8
<i>Speed statistics (km/h)</i>		
Minimum speed	46.6 ± 12.3	48.2 ± 10.8
Maximum speed	82.2 ± 17.1	74.0 ± 16.3
Mean speed	63.9 ± 9.1	59.6 ± 8.6
Standard deviation speed	11.6 ± 6.2	8.7 ± 6.1
10th speed percentile	51.2 ± 8	51.1 ± 7.7
20th speed percentile	54.0 ± 8.3	52.6 ± 7.7
30th speed percentile	56.9 ± 9.7	54.4 ± 8.6
40th speed percentile	60.0 ± 11.1	56.4 ± 9.8
50th speed percentile	63.5 ± 12.3	58.7 ± 10.9
60th speed percentile	66.7 ± 13.0	61.2 ± 12.0
70th speed percentile	70.1 ± 13.9	63.8 ± 13.2
80th speed percentile	74.0 ± 15.2	66.6 ± 14.4
90th speed percentile	77.9 ± 16.6	69.8 ± 15.2
<i>Percentage of trip based on road type</i>		
% distance on expressway	15.5 ± 22.6	7.4 ± 17.2
% distance on primary highway	17.3 ± 21.6	13.5 ± 21.3
% distance on secondary highway	12.3 ± 19.1	10.6 ± 20.1
% distance on major road	19.8 ± 20.4	19.0 ± 22.3
% distance on local road	34.5 ± 26.7	49.0 ± 29.8
% time on expressway	13.3 ± 19.8	6.4 ± 15.1
% time on primary highway	14.8 ± 19.2	11.2 ± 18.7
% time on secondary highway	12.1 ± 18.6	10.1 ± 19.4
% time on major road	19.6 ± 19.6	18.1 ± 21.4
% time on local road	37.8 ± 26.6	52.4 ± 29.3
<i>Percentage of longest road based on road type</i>		
% distance on expressway	7.7 ± 16.8	3.7 ± 12.7
% distance on primary highway	8.9 ± 18.2	8.6 ± 18.9
% distance on secondary highway	7.2 ± 16.6	7.0 ± 17.5
% distance on major road	6.6 ± 14.7	7.7 ± 16.9
% distance on local road	7.1 ± 16.8	12.9 ± 20.5
% time on expressway	5.3 ± 13.4	2.6 ± 10.2
% time on primary highway	7.3 ± 16.1	6.3 ± 16.3
% time on secondary highway	7.1 ± 16.1	6.5 ± 16.7
% time on major road	6.2 ± 14	7.2 ± 16.1
% time on local road	8.2 ± 17.3	14.9 ± 21.0

^a Include 653 observed or actual routes; total number of routes including alternatives is 3,938.

^b Include 809 observed or actual routes; total number of route including alternatives is 4,292.

the PPAG algorithm. The *work* dataset consists of 653 observations, and the number of alternatives ranges between 2 and 28 routes, with a mean value of 6. The *shop* dataset consists of 809 observations and the number of alternatives varies between 2 and 42 routes, with a mean value equal to 5. Consequently, the dataset for joint estimation contains 1,462 observations, with a maximum availability of 28 or 42 routes for each observation, according to trip purpose.

As expected, average work trip distance is roughly twice the average trip distance for shop trips – consistent with previous findings that shop trips are typically shorter than work trips (Zhang and Levinson, 2008). The majority of routes chosen by drivers for shop trips are around 5 km or less while work trips tend to dominate over shop trips from 10 km onward (Figure 5.1).

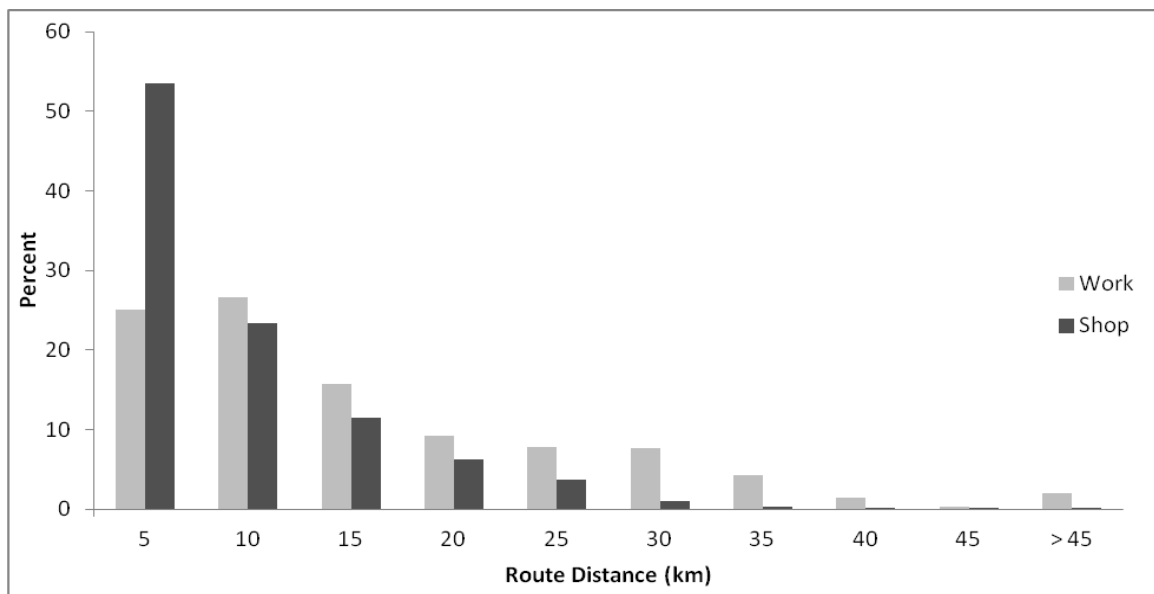


Figure 5.1 Route distance distribution for work and shop routes

Similar patterns can be observed with respect to network travel time. In general, routes for work trips tend to exhibit higher values than shop trips in terms of number of intersections, total turns, speed, and percentage of travel distance or time in expressways (Table 5.2). In addition, the results confirm that work trips tend to have more direct routes than shop trips. The path size values indicate that routes for work trips have more overlaps than those of shop trips.

5.3.2 Route choice behavior for shop trips versus that for work trips

Following the suggestion of Ben-Akiva et al. (1984), we used a limited set of route attributes commonly derived from transportation network data; that is, variables common to both datasets: route travel time, route distance, percentage of travel time on expressways, percentage of travel time on local roads, total turns, number of intersections, and log of path size (LNPS), which is the path size term. In this study, the number of intersections was used as a proxy for the number of stops. Also we focused on *total turns* because *left turns* have been used as a threshold to determine relevant alternative routes (Prato and Bekhor, 2006). Since right and left turns are both disutilities, we tested and found that there is a strong statistical basis to combine the two variables into a single variable, *total turns*.

Estimates for the PSL model are found in Table 5.3 for separate and combined *work* and *shop* datasets. All estimation processes are performed with Stata[®]. We tested MNL models for both datasets and compared them with PSL models. As expected, we found noticeable improvements in model fit in favor of PSL models, both statistically

significant at the 95% confidence level, which confirms similar findings from previous studies (Frejinger and Bierlaire, 2007; Bekhor and Prato, 2009).

Table 5.3 PSL estimates for separate and combined datasets: work versus shop trips

Variable	Work		Shop		Combined	
	coef.	z	coef.	z	coef.	z
Route travel time	-0.1925	-2.99 ^a	-0.4957	-3.87 ^a	-0.2728	-4.59 ^a
Route distance	0.0003	4.53 ^a	0.0006	4.35 ^a	0.0004	6.16 ^a
% of travel time on expressways	3.0565	5.34 ^a	0.6033	1.06	1.5787	4.36 ^a
% of travel time on local roads	-1.8514	-4.40 ^a	-1.2303	-3.80 ^a	-1.3056	-5.81 ^a
Total turns	-0.1119	-3.81 ^a	-0.3222	-10.81 ^a	-0.2104	-11.15 ^a
Number of intersections	-0.0410	-6.61 ^a	-0.0277	-3.69 ^a	-0.0338	-7.87 ^a
Ln of path size (LNPS)	-0.8006	-3.61 ^a	-0.4767	-2.41 ^b	-0.5508	-4.16 ^a
Scale parameter Work (μ_{WT})					1.000	Fixed
Scale parameter Shop (μ_{ST})					1.189	NA ^c
Parameters estimated	7		7		8	
Observations	653		809		1462	
Log likelihood intercept only	-1014.85		-1153.46		-2168.31	
Log likelihood full model	-759.23		-889.18		-1668.45	
Adjusted Rho-bar squared	0.245		0.223		0.227	

^a Statistically significant at level 0.01.

^b Statistically significant at level 0.05.

^c Not available.

In general, the model estimates based on generic route attributes have expected signs as reported in previous studies (Ben-Akiva et al., 1984; Ramming, 2002; Frejinger and Bierlaire, 2007; Zhang and Levinson, 2008; Bekhor and Prato, 2009; Kaplan and Prato, 2012). However, these attributes have different effects (magnitude) on route choice decisions by purpose. These findings imply that trip purpose should be explicitly considered in the model specification, which supports the same idea as suggested earlier by Ben-Akiva et al. (1984).

In both trip purposes, model estimates illustrate that drivers tend to minimize travel time, prefer to travel on expressways or major highways, avoid local roads, and minimize the number of turns and intersections. The only exception is that the variable *route distance* has positive signs (Frejinger and Bierlaire, 2007; Zhang and Levinson, 2008). Drivers seeking to minimize travel times would prefer higher functional type roadways such as expressways, eventually leading them to take longer distances (Ramming, 2002). Although drivers prefer to travel on expressways to reduce travel time for work trips, this consideration has no significant effect on shop trips ($p = 0.289$) since most of the shop trips were shorter in terms of distance compared to work trips (Figure 5.1), and often involved local roads (for work trips, the disutility of travel time on local roads is about 9.6 times the disutility of total travel time; only about 2.5 times for shop trips). The differences in the influence of route attributes between work and shop trips (based on separate models) were not detected in the combined or joint model. In particular, the estimate for percentage of travel time on expressways indicates a significant influence on route choice decisions based on the joint model but not significant in the separate model for shop trips. Neglect of this variation attributed to trip purpose may over-generalize the influence of route choice determinants.

For the most part, the model results are consistent with those of Frejinger and Bierlaire (2007). Although the LNPS estimate is conceived as a correction factor for link overlaps and expected to be positive, we agree with Frejinger and Bierlaire (2007) that the term has vague interpretation as shown by the negative estimates. Rather seen as a

correction factor, instead it is seen as a utility probably associated with attractive but hidden features of overlapping routes. We assume that route overlaps provide some advantages to drivers such as better access to other routes (route switching), faster routes (expressways), and access to shopping and other services. Interestingly enough, most shopping destinations are located along congested parts of the network where most of the routes overlap. In Table 5.3, negative LNPS estimates suggest that drivers tend to prefer routes that share links with other alternative routes – a higher preference in work trips than shop trips (based on LNPS estimate relative to route travel time).

5.3.3 Statistical test of utility parameters and scale equality

The comparison of separate models for work and shop trips revealed obvious differences in parameter estimates as described in Section 5.3.2. To provide statistical basis for this argument, we estimated separate 7-parameter PSL models for each dataset by trip purpose (Table 5.3), obtaining log likelihood values of -759.23 (L_{WT}) for *work* dataset and -889.18 (L_{ST}) for the *shop* dataset. We then tested hypothesis H_{1A} by assuming the parameters are the same in both datasets but that scale factors differed: the value of log likelihood of the *combined* datasets was -1668.45 (L_{RS}), corresponding to a relative scale factor estimate of 1.189. The chi-squared statistic for the H_{1A} hypothesis is $-2[-1668.45 - (-759.23 - 889.18)] = 40.07$, with 8 degrees of freedom, and the corresponding critical value of the chi-squared distribution at the 95% confidence level is 15.51; therefore, we reject the hypothesis of parameter equality. Because H_{1A} is rejected,

we rejected the equality of the utility and scale parameters for the separate models of work and shop trips (5.8). The rejection means that our estimate for the relative scale factor (1.189) is not valid, since it was premised on the equality of the parameter estimates obtained from the two datasets (5.9). This result implies that work and shop trips have underlying models with different parameters, which suggests that drivers attach value on route attributes relative to trip purpose. However, we should note that the difference may be the result of some hidden factors not fully captured by our model specifications such as perception errors associated with observed route attributes (Zhang and Levinson, 2008).

5.3.4 Difference between shop and work route choice highlighted by interaction variables

Aside from the statistical tests, we included interaction variables to provide further evidence that individual characteristics influence route choice behavior for work trips in different ways than shop trips (the relative importance of route choice determinants varies by trip purpose). The inclusion of interaction variables also allows us to compare the relative importance of route choice determinants because work and shop trips' models have different scale factors as shown in Section 5.3.3. Table 5.4 shows separate model estimates for work and shop trips, with interaction terms included in the specification. Due to the small number of individuals (those from large households or in the high-income category) and identical individuals in both datasets, there is a systematic loss of significance for some interaction variables. In spite of this, there is an increase of model

fit and stark contrasts in utilities between work and shop trips (Table 5.4); these contrasts are discussed in the following paragraphs by individual characteristics.

Table 5.4 PSL estimates for shop and work, with interaction terms

Variable	Work		Shop	
	coef.	z	coef.	z
Route travel time	-0.2270	-3.30 ^a	-0.6911	-4.86 ^a
× Personal income (C\$80,000-\$99,999)	-0.1706	-1.16	0.2912	1.19
× Age (65 years and above)	-0.1001	-0.12	0.3084	2.55 ^b
Route distance	0.0004	4.81 ^a	0.0006	4.79 ^a
% of travel time on expressways	3.0538	5.32 ^a	0.4559	0.80
% of travel time on local roads	-1.1382	-1.97 ^b	-2.0208	-3.39 ^a
× Residence tenure (10 years and over)	-1.0696	-1.45	1.1687	1.76 ^c
Total turns	-0.0809	-2.68 ^a	-0.2549	-6.85 ^a
× Travel time (< 5 minutes)	-0.2032	-3.76 ^a	-0.1479	-3.54 ^a
× Age (65 years and above)	-0.2840	-0.70	-0.0141	-0.22
Number of intersections	-0.0352	-4.62 ^a	-0.0405	-3.92 ^a
× Sex (male)	-0.0129	-1.28	0.0273	2.18 ^b
Ln of path size (LNPS)	-1.0208	-2.63 ^a	-0.4633	-1.22
× Residence tenure (10 years and over)	0.3434	0.85	-0.4344	-1.07
× Frequency of public transit use (never)	0.1786	0.46	1.1418	2.58 ^a
× Household size (≥ 5)	-0.9161	-1.19	0.3443	0.51
Parameters estimated	16		16	
Observations	653		809	
Log likelihood intercept only	-1014.85		-1153.46	
Log likelihood full model	-747.00		-871.05	
Adjusted Rho-bar squared	0.248		0.231	

^a Statistically significant at level 0.01.

^b Statistically significant at level 0.05.

^c Statistically significant at level 0.1.

High income versus low income. It is known that high-income drivers put more value on travel time savings when it comes to work trips (e.g., Pitombo et al., 2011; Papinski and Scott, 2013), but we found that this is not the case for shop trips. In fact, for

shop trips, high-income drivers put less weight on the disutility of travel time than low-income drivers. Specifically, the disutility of travel time for high-income drivers (personal income \$80,000-\$99,999 CAD) for work trips is $(-0.2270 + -0.1706 =) -0.3976$, about 1.8 times the disutility of travel time for low-income drivers. Quite the reverse for shop trips where the disutility of travel time for low-income drivers is $(-0.6911 + 0.2912 =) -0.3999$, about 1.7 times the disutility of high-income drivers for travel time. This comparison clearly illustrates the stark contrast between work and shop trips in terms of route choice decisions, given the income category of respondents.

Old versus young. For work trips, the effective coefficient of travel time for older drivers is $(-0.2270 + -0.1001 =) -0.3271$, about 1.4 times the disutility of younger drivers; for shop trips, the travel time coefficient for younger drivers $(-0.6911 + 0.3084 =) -0.3827$, about 1.8 times the disutility of travel time for older drivers. The travel time disutility for older drivers is not significant for work trips due to the very small number of working seniors (Table 5.1). In spite of this, the estimates reveal that older drivers are less sensitive to travel time when it comes to shop trips than younger cohorts. This is consistent with the observation that seniors have less mandatory activities than their counterparts (Scott et al., 2009); therefore, they can afford to choose inefficient routes.

In both trip purposes, older drivers are more likely to minimize the total number of turns than younger drivers. Moreover, older drivers have higher propensity to minimize total turns for work trips than shop trips - to minimize travel delay and consequently reduce total travel time. Turns, particularly left turns, have been identified as one of the

risky maneuvers for older drivers (Chandraratna and Stamatiadis, 2003); no wonder that they want to minimize turns as much as possible.

Long-term versus new resident. Longer residency (10 years and over) have opposite effects on route choice decisions between work and shop trips. Because work schedule is typically tighter than that of shopping, those with good knowledge of surrounding routes - commonly drivers who lived in the neighborhood for at least ten years - are more likely to avoid local roads for work trips than new-resident drivers. In this case, drivers perceived local roads to cause travel delay and need to be avoided. But with respect to shop trips, drivers who are long-term residents perceived local roads to offer more shopping opportunities hence they are less deterred to travel in local roads than newer residents.

The effects of residence tenure to LNPS estimates strengthen our previous assumption that LNPS is more of a utility than a correction factor. Thus, LNPS somehow indicates a more behavioral interpretation, at least in this case, than what was intended originally in choice theory (Ben-Akiva and Lerman, 1985). There is a direct relationship between LNPS and route overlaps; with negative LNPS coefficients for both work and shop trips, route overlaps become a utility rather than just a correction factor. But if we take into account length of residency, some intuitive differences between the two models emerge. Specifically, for work trips, long-term resident drivers with good familiarity of surrounding routes are more likely to avoid common links (links that overlap with other routes), than newer residents. On the other hand, they behave differently when navigating

the network to shopping destinations; they have more preference for routes with common links than their counterparts. Hence, common links are perceived to provide more shopping opportunities as shopping centers are typically located along these links. Also these links provide better access to routes toward other shopping destinations or activity locations.

Short-duration versus long-duration travel. For short-duration travel (< 5 minutes), drivers are more likely to minimize turns for work trips than shop trips. This implies that drivers prefer a more direct route for work trips to meet work schedule requirements as opposed to shop trips. Maximization of route directness was also reported in previous studies that focused on work trips (Papinski et al., 2009; Prato et al., 2012).

Male versus female. Male drivers are more likely to avoid intersections for work trips than shop trips. Conversely, male drivers have higher propensity to avoid intersections over female drivers for work trips, while the opposite is true for shop trips. Perhaps female drivers prefer to minimize travel delays at intersections to maximize their time for shopping.

Public transit experience versus without. Drivers who have not used public transit are more likely to avoid common links (overlap routes) for shop trips than work trips. Since these drivers were car dependent (never used public transit), they tend to enjoy more mobility than their counterparts - probably leading them to explore more unique routes. This behavior appears to be significant in the model for shop trips.

Large versus small household. Because work schedule puts more pressure on other activities, drivers living in large households are more inclined to choose routes for work trips with common links than those in small households. Drivers, in this case, may perceived route overlaps to provide them with more opportunities to perform other activities (e.g., buy breakfast, drop kids to school) along the way to work. However, for shop trips, drivers from large households tend to choose unique routes to minimize travel time for more time for shopping, especially to visit more shopping destinations.

In summary, the addition of interaction variables improved the model fit and highlighted the contrast in route choice determinants between work and shop trips. Overall, the model results suggest a *restrictive* route choice behavior for work trips, in contrast to *nonrestrictive* route choice behavior for shop trips - both consistent with the mandatory and discretionary nature of trips involved.

5.4 Conclusion

The empirical results clearly indicate that route choice behavior varies by trip purpose as suggested by the inequality of utility parameters and scale factors for separate models of work and shop trips (found to be statistically significant at level 0.05), given generic route attributes commonly employed in previous studies (Ramming, 2002; Frejinger and Bierlaire, 2007; Kaplan and Prato, 2012). Moreover, the comparison of PSL models incorporating the interaction between individual characteristics and route attributes have demonstrated the stark contrasts in the relative importance of route choice

determinants for work versus shop trips. In general, work route choice behavior tends to be *restrictive* while shop route choice behavior tends to be *nonrestrictive* – a generalization that is consistent with the mandatory and discretionary nature of work and shop trips, respectively. For work trips, route choice decisions are restricted by work schedule. Hence, the choice of routes is strongly influenced by route attributes that makes the total travel experience faster and easier in order to arrive on time at work destinations. Consequently, drivers tend to select routes with short travel time, prefer to travel on expressways, avoid local roads, and minimize number of turns and intersections. In contrast, these determinants have lax effects on route choice behavior for shop trips relative to work trips. Sometimes the effect is entirely the opposite; for example, high-income individuals are more likely to avoid routes with longer duration when driving to work but less likely to avoid the same routes when driving to shop. These results suggest that trip purpose should be considered in route choice model specification.

We deduce from the model results that path size (represented as LNPS in the PSL models) took on a more behavioral meaning than originally intended (Ben-Akiva and Bierlaire, 1999). Along with Frejinger and Bierlaire (2007), we observed that LNPS captures the attractiveness of latent factors associated with overlapping routes. PSL models of work and shop trips suggest that drivers perceived LNPS as an indicator of common links. Their route choice behaviors vary accordingly depending on the mandatory or flexible nature of trip purpose in relation to the perceived benefits from choosing common links. Along these lines, it would be interesting to consider the latent

factors explicitly (e.g., number of shopping opportunities along route links) in model specification to find out if these reduce the effects of LNPS. Also, future developments should consider different route choice sets to test the sensitivity of LNPS on path generation algorithms.

The PPAG algorithm offers an efficient yet theoretically sound alternative among path generation algorithms. The algorithm uses the PPA approach in constraining feasible routes, which lends it theoretical validity and GSP lends it computational efficiency. Moreover, the entire set of GERT's modules on route choice data generation (TGEM, CSGM, RVGM) proved to be successful in automatically generating relevant inputs for the route choice analyses, also particularly suited to the requirements of PSL and related route choice models. Descriptive analysis of the route choice sets generated by the PPAG algorithm provided intuitive results concerning route alternatives for work and shop trips, an outcome that suggests the reasonableness of the PPAG algorithm in generating route choice sets. GERT's modules, particularly the PPAG algorithm, are useful in providing relevant inputs for further in-depth investigations into route choice decision processes.

Traditional route assignment models may have oversimplified the complex nature of route choice behavior. Aside from confirming the fact that route choice determinants are not only limited to travel time, this paper has also demonstrated that route choice behavior varies by trip purpose. Moreover, it was shown that the relative importance of route attributes varies as well in relation to individual characteristics though the findings need larger sample to be more conclusive. Future research should consider the use of

larger samples taken from different locations, and more efficient scaling estimation procedures (e.g., Full Information Maximum Likelihood) to provide comparable estimation results and identify more determinants that significantly differentiates between work and shop trips.

5.5 Acknowledgements

Financial support for this project was provided by a grant awarded to Darren M. Scott from the Natural Sciences and Engineering Research Council of Canada (261850-2009). The authors acknowledge the Halifax STAR Project for providing the data for this research.

5.6 References

- Adamowicz, W., Louviere, J., Williams, M., 1994. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management* 26 (3), 271-292.
- Bekhor, S., Ben-Akiva, M., Ramming, M.S., 2002. Adaptation of logit kernel to route choice situation. *Transportation Research Record: Journal of the Transportation Research Board* 1805, 78-85.
- Bekhor, S., Prato, C.G., 2009. Methodological transferability in route choice modeling. *Transportation Research Part B: Methodological* 43(4), 422-437.
- Ben-Akiva, M.E., Bergman, M.J., Daly, A.J., Ramaswamy, R., 1984. Modeling interurban route choice behavior. In: Volmuller, J., Hamerslag, R. (Eds.),

Proceedings of the Ninth International Symposium on Transportation and Traffic Theory, VNU Science Press, Utrecht, pp. 299-330.

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. In Hall, R. (Ed.), *Handbook of Transportation Science* Vol. 23, Springer, US, pp. 5-33.

Ben-Akiva, M.E., Gao, S., Wei, Z., Wen, Y., 2012. A dynamic traffic assignment model for highly congested urban networks. *Transportation Research Part C: Emerging Technologies* 24 (0), 62-82.

Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.

Ben-Akiva, M., Morikawa, T., 1990. Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A: General* 24 (6), 485-495.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies* 16 (2), 187-198.

Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation* 21 (2), 167-184.

Carpenter, S.M., 1979. Drivers' route choice project - pilot study (Research Report): Transport Studies Unit, Oxford University.

Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., 1996. A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In Lesort, J.B. (Ed.), *Transportation and Traffic*

theory: Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Pergamon, Lyons, France.

Chandraratna, S., Stamatiadis, N., 2003. Problem driving maneuvers of elderly drivers. *Transportation Research Record: Journal of the Transportation Research Board* 1843, 89-95.

Dalumpines, R., Scott, D.M., 2011. GIS-based map-matching: development and demonstration of a postprocessing map-matching algorithm for transportation research. In: Geertman, S., Reinhardt, W., Toppen, F. (Eds.), *Advancing Geoinformation Science for a Changing World Vol. 1*, Springer, Berlin, pp. 101-120.

Dalumpines, R., Scott, D.M., 2014. GIS-based episode reconstruction toolkit (GERT): a transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data. Manuscript in preparation.

Frejinger, E., Bierlaire, M., 2007. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological* 41 (3), 363-378.

Frejinger, E., Bierlaire, M., 2010. On path generation algorithms for route choice models. In: Hess, S. (Ed.), *Choice Modelling: The State-of-the-art and The State-of-practice*, Emerald Group Publishing Limited, Bradford.

Hägerstrand, T., 1970. What about people in regional science. *Papers of the Regional Science Association* 24 (1), 6-21.

Hanson, S., 2004. The context of urban travel: concepts and recent trends. In: Hanson, S., Giuliano, G. (Eds.), *The Geography of Urban Transportation* (3rd ed.), Guilford Press, New York, pp. 3-29.

- Hensher, D., Louviere, J., Swait, J., 1998. Combining sources of preference data. *Journal of Econometrics* 89 (1-2), 197-221.
- Kaplan, S., Prato, C.G., 2012. Closing the gap between behavior and models in route choice: the role of spatiotemporal constraints and latent traits in choice set formation. *Transportation Research Part F: Traffic Psychology and Behaviour* 15 (1), 9-24.
- Lombard, K., Church, R.L., 1993. The gateway shortest path problem: generating alternative routes for a corridor location problem. *Geographical Systems* 1, 25-45.
- Louviere, J.J., Hensher, D.A., Swait, J.D., 2000. *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, UK.
- Millward, H., Spinney, J., 2011. Time use, travel behavior, and the rural-urban continuum: results from the Halifax STAR Project. *Journal of Transport Geography* 19 (1), 51-58.
- Nelson, D., Niles, J., 2000. Observations on the causes of nonwork travel growth. In: *79th Annual Meeting of the Transportation Research Board: Compendium of Papers CDROM*, Washington D.C., January 9-13.
- Papinski, D., 2010. *Investigating Route Choice Decisions Using GPS and Prompted-recall Diary Data*, Ph.D. Thesis, McMaster University, Hamilton.
- Papinski, D., Scott, D.M., 2011. A GIS-based toolkit for route choice analysis. *Journal of Transport Geography* 19 (3), 434-442.
- Papinski, D., Scott, D.M., 2013. Route choice efficiency: an investigation of home-to-work trips using GPS data. *Environment and Planning A* 45 (2), 263-275.

- Papinski, D., Scott, D.M., Doherty, S.T., 2009. Exploring the route choice decision-making process: a comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour* 12 (4), 347-358.
- Pitombo, C. S., Kawamoto, E., Sousa, A.J., 2011. An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. *Transport Policy* 18 (2), 347-357.
- Prashker, J.N., Bekhor, S., 2004. Route choice models used in the stochastic user equilibrium problem: a review. *Transport Reviews: A Transnational Transdisciplinary Journal* 24 (4), 437-463.
- Prato, C.G., 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modeling* 2 (1), 65-100.
- Prato, C.G., Bekhor, S., 2006. Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board* 1985, 19-28.
- Prato, C., Bekhor, S., 2007. Modeling route choice behavior: how relevant is the composition of choice set? *Transportation Research Record: Journal of the Transportation Research Board* 2003, 64-73.
- Prato, C., Bekhor, S., Pronello, C., 2012. Latent variables and route choice behavior. *Transportation* 39 (2), 299-319.
- Ramming, S., 2002. *Network Knowledge and Route Choice*, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.

- Scott, D.M., 2006. Constrained destination choice set generation: a comparison of GIS-based approaches. In: 85th Annual Meeting of the Transportation Research Board: Compendium of Papers CD-ROM, Washington, D.C., 2006, January 22-26.
- Scott, D.M., Newbold, K.B., Spinney, J.E.L., Mercado, R., Páez, A., Kanaroglou, P.S., 2009. New insights into senior travel behavior: the Canadian experience. *Growth and Change* 40 (1), 140-168.
- Swait, J., Louviere, J., 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research* 30 (3), 305-314.
- Wen, Y., Balakrishna, R., Ben-Akiva, M., Smith, S., 2006. Online deployment of dynamic traffic assignment: architecture and run-time management. *Intelligent Transport Systems IEE Proceedings* 153 (1), 76-84.
- Zhang, L., Levinson, D., 2008. Determinants of route choice and value of traveler information: a field experiment. *Transportation Research Record: Journal of the Transportation Research Board* 2086, 81-92.

Chapter 6

Conclusion

This dissertation set out to advance the current tools and methods of automatically generating information from global positioning system (GPS) data to support travel behavior research at the individual level, specifically to provide inputs for activity analysis in general and route choice modeling in particular. Existing tools and methods tend to suffer from several limitations: lack of transferability, lack of a complete set of integrated tools, and lack of capability in processing large GPS data. Collectively, these limitations impede leveraging GPS data to provide inputs for studies on individual travel behavior. Consequently, this dissertation proposed a geographic information system (GIS)-based episode reconstruction toolkit (GERT), which was demonstrated to be promising in automatically extracting activity episodes from GPS data and in deriving information related to these episodes from additional data such as a road network and land use. Specifically, this dissertation introduced and demonstrated the utility of GERT's core components such as map-matching and mode detection algorithms, emphasized the importance of a framework for the development of GERT's integrated set of tools, and demonstrated the utility of GERT in providing inputs for route choice modeling. In this final chapter, the contributions and implications of this dissertation are presented, as well as the directions for future research.

6.1 Contributions to activity analysis and route choice modeling

The four substantive chapters (*Chapters 2 to 4*) of this dissertation, taken together, focus on the development of a transferable, modular, and scalable toolkit (GERT) useful in extracting stationary activity and travel episodes from GPS data, and in appending more information to these episodes from additional data such as a road network and land use. These substantive chapters correspond to the four specific objectives laid out in Chapter 1. In general, the substantive chapters highlighted the potential of GERT to provide useful inputs to activity analysis and route choice modeling, given the increasing availability of GPS data. Specifically, each substantive chapter demonstrated the potential of GERT and some of its key components as summarized below.

Chapter 2 presented the development and demonstration of a GIS-based map-matching (GMM) algorithm. The GMM algorithm produced accurate results in a reasonable amount of time. In addition, the GMM algorithm generated relevant route attributes such as travel time, travel distance, and number of left and right turns that serve as explanatory variables in route choice models. The ease and flexibility of the Python[®] scripting language used in developing the GMM algorithm make this tool easy to develop and implement. It can be improved to suit data inputs and specific fields of application in transportation research. As GIS increasingly becomes a popular tool in transportation and other disciplines, the GMM algorithm serves as a practical tool that GIS users can easily use to automatically extract routes from GPS trajectories. For example, in extracting

bicycle routes from the data collected by GPS-enabled smartphones, Hudson et al. (2012) had chosen the GMM algorithm over other post-processing map-matching algorithms (e.g., Schuessler & Axhausen, 2009a; Newsom & Krumm, 2009; Hood, 2010) for it was easy to use for ArcGIS[®] users, and unlike other map-matching algorithms they reviewed, requires very minimal user-specified values.

Chapter 3 introduced a transferable and efficient method of extracting and classifying activity episodes from GPS data without additional information. The proposed method, which was referred as GERT's GPS Episodes Extraction and Mode Detection (MDM) component, was found to be promising with 90% overall accuracy, despite using only three variables derived from extracted episodes from GPS data: median speed (m/s), maximum change in heading (degrees), and total duration (minutes). The calculated kappa statistic is 0.91 ($p < 0.001$), which indicated almost perfect agreement between observed episodes and those predicted by the multinomial logit (MNL) model. However, the model was ineffective in predicting *bus* and *other travel* episodes due to the limited samples for these episode types. In spite of this, the proposed method showed potential as a more transferable and efficient alternative among mode detection algorithms (e.g., Gonzalez et al., 2008; Schuessler & Axhausen, 2009b; Gong et al., 2012), given few input requirements directly derived from GPS data and the efficiency provided by the MNL model. Moreover, this chapter is the first to introduce the use of the MNL model in detecting the most likely mode for each extracted episode from GPS data. The straightforward procedures in extracting episodes, along with their descriptive statistics,

provide researchers with rich information to analyze episode characteristics and to develop more effective and efficient algorithms for mode detection.

Chapter 4 presented the entire toolkit, with particular emphasis on its main components, and demonstrated its capability in extracting activity episodes from GPS data. About 26,000 episodes were automatically reconstructed using GERT from 47.3 million GPS points. A comparison of the episode and duration distributions reveal similar patterns between time-use diary (TUD) and GPS episodes, a similarity that confirms that GERT's components work properly, at the aggregate level, in reconstructing episodes from GPS data. From the standpoint of the three challenges faced by existing methods (Section 1.1.1), GERT's overall performance can be considered impressive because of its scalability – GERT can scale up to large GPS data (aside from its ability to accommodate additional information); modularity – GERT has a complete set of tools to support analyses and model estimations; and transferability – GERT's reliance on generic variables (latitude, longitude, time) makes it applicable to other places. Overall, GERT's modules provide transportation researchers with a set of practical tools in extracting rich datasets (e.g., stationary activity and travel episodes, activity locations, travel segments, route choice sets, route attributes) from GPS datasets to advance our understanding of activity/travel patterns in general and route choice decision processes in particular.

Chapter 5 compared the separate route choice models of work and shop trips to test whether route choice decision processes vary by trip purpose, and in the process, demonstrated the utility of Potential Path Area-Gateway (PPAG) algorithm and other

GERT modules in generating inputs for route choice modeling. The results showed that, indeed, route choice behavior varies by trip purpose because utility and scale parameters were statistically different in separate models of work and shop trips. The inequality suggested that drivers attach value to route choice determinants in relation to trip purpose. The inclusion of interaction terms in model specifications further indicated that work route choice behavior tends to be *restrictive* compared to the *nonrestrictive* route choice for shop trips, a generalization consistent with the mandatory and discretionary nature of work and shop trips, respectively. Moreover, these results and the descriptive analysis of route choice sets demonstrated that the PPAG algorithm generates reasonable alternatives and showed potential as a practical alternative among path generation algorithms (e.g., Ben-Akiva et al., 1984; Ramming, 2002; Frejinger & Bierlaire, 2010). GERT's modules interlinked with the PPAG algorithm helped to fully automate the procedures involved in extracting inputs ready for route choice modeling. With the increasing availability of GPS data, GERT's ability to generate route choice data lessens the burden on researchers in collecting and processing these data (note the tedious process of manually tracing travelled routes, e.g., Ben-Akiva et al., 1984; Ramming, 2002; Papinski et al., 2009; Winters et al., 2010), and allow them to focus on other important aspects of route choice modeling.

6.2 Practical and theoretical implications

The findings from this dissertation, taken together, suggest the importance of a framework for the development of tools and methods in the utilization of GPS data for research. This dissertation also bolsters the important role of GIS as an ideal platform for toolkit development, and encourages the development of simple but effective techniques to further leverage GPS data for transportation research. From the theoretical front, this dissertation suggests that route choice modeling should consider the influence of trip purpose on route preferences. These practical and theoretical implications are further discussed as follows.

Importance of a framework in developing a toolkit to effectively utilize GPS data for research. In the course of the design and development of GERT's components, it has been observed, in general, that basic structures tend to be lacking in the development of existing tools and methods for the extraction of information from GPS data. Basic structures, in the form of guiding principles, were not explicitly considered; instead, localized and immediate objectives (e.g., to derive inputs for modeling) tend to be the guiding principles in the development of techniques applied to GPS data utilization (see for example Section 3.1). This dissertation suggests the importance of frameworks that consider a broader perspective – taking into account the potentials and limitations of GPS data across different sources and across different problem domains. Moreover, this broader perspective should be rooted in an implementation framework that covers

problem identification, need and feasibility analyses, and other considerations similar to strategies employed in the adoption of GPS technology in the commercial sector (Theiss et al., 2005). As a starting point, this dissertation demonstrated the use of a framework based on the software design principles of transferability, modularity, and scalability. This framework guided the development of GERT's components, which is potentially applicable across different GPS data sources, but also useful to other problem domains besides transportation.

Bolster the important role of GIS as an ideal platform for the development of tools and methods for extraction of information from GPS data. GIS has long ago permeated the confines of transportation research, and researchers in this field are often adept in the use of GIS in spatial data management, analysis, and reporting. GERT's main components were interfaced with ArcGIS® as additional tools, an integration that allows researchers already familiar with the GIS software ease and flexibility in using GERT's components – helping them to be more productive. Since GERT's spatial outputs are in native ArcGIS® shapefile format (e.g., outputs of TUD-GPS Trip Segments Extraction Module (TGEM) and Activity Locations Identification Module (ALIM)), they can be easily viewed and manipulated in ArcGIS® or other GIS applications. With the proliferation of GPS data, GIS becomes naturally a preferred tool in handling these locational data (e.g., Shaw & Wang, 2000; Miller, 2003); in this context, GERT's integration with GIS extends existing GIS functionalities and further strengthens the utility of GIS for transportation research.

Simple techniques can be as effective as complicated ones. In the development of tools and methods for extracting information from GPS data, this dissertation showed that simple techniques (e.g., use of GPS-derived generic inputs such as location and time stamp as the core inputs for GERT's algorithms, MNL for episode classification, basic GIS functionalities in map-matching and extraction of additional information from other sources) proved to be effective in extracting and classifying episodes (Chapters 3 and 4), in retrieving actual routes and attributes for travel episodes (Chapters 2 and 5), and in extracting more information for activity locations from additional data (Chapter 4). This observation agrees with what Karlaftis and Vlahogianni (2011, p. 396) had opined – that “frequently simpler models give as good results as complex ones”. Simple techniques, as demonstrated in this dissertation, require minimal inputs and tend to be more efficient than advanced procedures developed so far – considering the effort and time in tool development, set-up, and processing.

GERT to leverage GPS data for transportation research. This dissertation proposed a toolkit and demonstrated the potential of this toolkit in the automatic generation of inputs for activity analysis and route choice modeling. With this toolkit, transportation researchers can easily and essentially reconstruct stationary activity and travel episodes from GPS data, and with these reconstructed episodes, conduct various analyses and modeling. Aside from route choice modeling, GERT provides useful functionalities for other research fields that increasingly use GPS to track down individual movements such as physical activity (e.g., Handy et al., 2002; Krenn et al., 2011; Clark et

al., 2014), tourism (e.g., Shoval & Isaacson, 2007), health research (e.g., Kerr et al., 2011), traffic congestion (e.g., Taylor et al., 2010), among others.

Route choice models should consider the influence of trip purpose. The conventional traffic assignment uses shortest path, often based on estimated travel time, in loading trips onto the road network without regard to trip purpose. Chapter 5 of this dissertation has shown that route choice preference varies by trip purpose, and the relative importance of route choice determinants (travel time among one of them) varies as well. These findings are consistent with that of earlier studies (e.g., Wachs, 1967; Carpenter, 1979; Ben-Akiva et al., 1984; Zhang & Levinson, 2008), and suggest the need to consider trip purpose in route choice model specification.

6.3 Directions for future research

Future work should focus on the sensitivity and comparative analyses, methodological expansion, case studies that employ GERT's components, tuning tests to improve the toolkit's performance, and exploration of GERT's potential for web deployment. These future research directions are discussed below.

Sensitivity analysis and detailed validation. This dissertation used data captured by person-based GPS devices with high temporal resolution (at least one reading per second) and horizontal accuracy of 10 m or better in the development and validation of GERT's components. Future research would benefit from the application of GERT to GPS data with different spatial and temporal resolutions, particularly those collected by

GPS devices with lower positional accuracy, typical of GPS-enabled smartphones, tablets, and other mobile devices equipped with GPS receivers. Also, it would be interesting to test the sensitivity of GERT's components (e.g., GMM, ALIM) and their parameters to the spatial resolution of road network and additional data such as land use and points of interest. Similarly, the PPAG algorithm's behavioral thresholds (adopted from Prato and Bekhor, 2006) should be further tested to determine the effects of these thresholds to realistic generation of alternative routes, given the observed routes extracted from GPS data. Finally, a detailed episode-to-episode validation test should be conducted along the lines of the framework suggested by Stopher and Shen (2011) to evaluate GERT's capability in extracting episodes from GPS data – with emphasis on sequential accuracy (i.e., if GERT's episodes closely replicate the sequence of episodes as reported in TUDs).

Comparative analysis. This analysis should focus on identifying the strengths and limitations of different preprocessing (i.e., data cleaning and smoothing) and episode classification (mode detection) algorithms, including GERT's equivalent modules. This is a tricky task because of the different assumptions, input requirements, and run settings of the algorithms to be compared (Lawson et al., 2010). In this case, GERT's emphasis on the use of generic variables derived directly from GPS data (e.g., latitude, longitude, time, speed, distance, duration, heading, and acceleration) may serve as a common ground in establishing a framework for comparison. The idea behind this comparison is to make GERT more flexible in handling different scenarios, taking advantage of the strengths of

other algorithms and consequently giving users the option to use an algorithm appropriate for given scenario. Furthermore, there is some value in incorporating other algorithms that fits well within GERT's framework.

Methodological expansion. Future work may also extend the functionalities of GERT's components. In general, GERT's flexible framework (Figure 4.2) allows for expansion in terms of additional modules that can be incorporated at the Stage level, or functional enhancements (e.g., new methods and properties) within the core modules. For example, other mode detection algorithms such as rule-based methods (e.g., Bohte & Maat, 2009; Gong et al., 2012), fuzzy logic (e.g., Tsui & Shalaby, 2006; Schuessler & Axhausen, 2009b), neural networks (e.g., Gonzalez et al., 2008), and other machine learning algorithms (e.g., Zheng et al., 2008; Bolbol et al., 2012) can be adapted and plugged into GERT at the Stage 2 level. This effort would require the new modules to adapt to the outputs generated at the Stage 1 level, and generate outputs in the format accessible to GERT's components at Stages 3 and 4. An example for functional enhancements would be to extend ALIM's functionality to automatically assign activity type to stationary episodes based on spatio-temporal characteristics of episodes and other information (Huang et al., 2010). These are not easy tasks, but these efforts would provide GERT's users the option to choose algorithms appropriate for a particular task, and further leverage the use of GPS data for research purposes.

Case studies (application). GERT's ability to reconstruct stationary activity and travel episodes from GPS data could provide the inputs for case studies. For example,

activity locations can be generated easily from GPS data using GERT; these locations can be used for exploratory analysis and modeling to determine activity patterns over time and space, and determine the factors that influence these patterns. Using the travel episodes and the observed routes based on these episodes, it is possible to conduct comparative studies of route choice efficiency across different geographic regions to advance knowledge in this particular issue (Papinski & Scott, 2013). Many case studies that employ GERT's outputs could help in identifying other data needs that can be derived directly or indirectly from GPS data, and subsequently would inform future enhancements to GERT's components.

Performance tuning. Because GERT consists of several components, detailed assessment of the computing performance of its components was not fully implemented (a rough assessment was presented in Section 4.4.2). Although Python[®] programming language excelled in rapid prototyping and was chosen for GERT's development, the need for computing speed requires detailed analysis of the code to determine bottlenecks, and if practical, recoding of some of the components in C++ or other compiled languages. Future work on performance tuning of GERT's modules should also consider the use of efficient algorithms to boost performance, without the loss of accuracy.

Web deployment and sharing. GERT's components may be deployed in a web application environment in order to share its functionalities to a wider audience. Through client applications that send requests to a GERT-enabled web application, users can upload GPS data for processing with the option of displaying activity locations and travel

routes in an interactive map or download the extracted episodes for further data analysis and modeling (e.g., Bohte & Maat, 2009). The web-based application environment may also serve as a platform for collaboration among researchers interested in the development of tools and methods for the utilization of GPS data – with GERT’s components providing the server-side processes. Taking advantage of GERT’s modularity, research collaborators may enhance existing modules, and add new modules to expand GERT’s capabilities or provide better alternatives to current modules. Future work along this line should consider the experience gained from similar undertakings such as the work of Macal and North (2009) in the development of a free and open-source agent-based modeling and simulation (ABMS) toolkits.

6.4 References

- Ben-Akiva, M. E., Bergman, M. J., Daly, A. J., & Ramaswamy, R. (1984). Modeling interurban route choice behavior. In J. Volmuller & R. Hamerslag (Eds.), *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory: 11-13 July, 1984, Delft, The Netherlands* (pp. 299-330). Utrecht, The Netherlands: VNU Science Press.
- Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285-297.
- Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36(6), 526-537.

Carpenter, S. M. (1979). Drivers' route choice project - pilot study. Transport Studies Unit, Oxford University.

Clark, A.F., Scott, D.M. & Yiannakoulias, N. (2014). Examining the relationship between active travel, weather, and the built environment: a multilevel approach using GPS-enhanced data. *Transportation*, 41(2), 325-338.

Frejinger, E., & Bierlaire, M. (2010). On path generation algorithms for route choice models. In S. Hess (Ed.), *Choice modelling: the state-of-the-art and the state-of-practice*. Bradford: Emerald Group Publishing Limited.

Gong, H., Chen, C., Bialostozky, E., Lawson, C.T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36(2), 131-139.

Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P. L., Georggi, N. L., & Perez, R. (2008). Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. Paper presented at the 15th World Congress on Intelligent Transportation Systems.

Handy, S. L., Boarnet, M. G., Ewing, R., & Killingsworth, R. E. (2002). How the built environment affects physical activity: views from urban planning. *American Journal of Preventive Medicine*, 23(2, Supplement 1), 64-73.

Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters*, 3(1), 63-75.

Huang, L., Li, Q., & Yue, Y. (2010). Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. Paper presented at the Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, San Jose, California.

- Hudson, J. G., Duthie, J. C., Rathod, Y. K., Larsen, K. A., & Meyer, J. L. (2012). Using smartphones to collect bicycle travel data in Texas (Final Report No. UTCM 11-35-69). College Station, Texas: Texas Transportation Institute, The Texas A&M University System.
- Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3), 387-399.
- Kerr, J., Duncan, S., & Schipperjin, J. (2011). Using global positioning systems in health research: a practical approach to data collection and processing. *American Journal of Preventive Medicine*, 41(5), 532-540.
- Krenn, P. J., Titze, S., Oja, P., Jones, A., & Ogilvie, D. (2011). Use of global positioning systems to study physical activity and the environment: a systematic review. *American Journal of Preventive Medicine*, 41(5), 508-515.
- Lawson, C. T., Chen, C., & Gong, H. (2010). Advanced applications of person-based GPS in an urban environment. New York: New York University at Albany. Retrieved from http://www.utrc2.org/sites/default/files/pubs/advanced-applications-gps1-final_2.pdf
- Macal, C. M., & North, M. J. (2009). Agent-based modeling and simulation. In *Winter Simulation Conference* (pp. 86-98). Winter Simulation Conference.
- Miller, H. J. (2003). What about people in geographic information science? *Computers, Environment and Urban Systems*, 27(5), 447-453.
- Newsom, P., & Krumm, J. (2009). Hidden Markov map matching through noise and sparseness. *ACM GIS 2009*. Seattle, WA.

- Papinski, D., & Scott, D. M. (2013). Route choice efficiency: an investigation of home-to-work trips using GPS data. *Environment and Planning A*, 45(2), 263-275.
- Papinski, D., Scott, D. M., & Doherty, S. T. (2009). Exploring the route choice decision-making process: a comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4), 347-358.
- Prato, C. G., & Bekhor, S. (2006). Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board*, 1985, 19-28.
- Ramming, S. (2002). *Network knowledge and route choice* (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, USA.
- Schuessler, N., & Axhausen, K.W. (2009a). Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT). Swiss Federal Institute of Technology, Zurich.
- Schuessler, N., Axhausen, K.W., (2009b). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105, 28-36.
- Shaw, S. L., & Wang, D. (2000). Handling disaggregate spatiotemporal travel data in GIS. *GeoInformatica*, 4(2), 161-178.
- Shoval, N., & Isaacson, M. (2007). Tracking tourists in the digital age. *Annals of Tourism Research*, 34(1), 141-159.

Stopher, P., & Shen, L. (2011). In-depth comparison of global positioning system and diary records. *Transportation Research Record: Journal of the Transportation Research Board*, 2246, 32-37.

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

Theiss, A., Yen, D. C., & Ku, C.-Y. (2005). Global positioning systems: an analysis of applications, current development and future implementations. *Computer Standards & Interfaces*, 27(2), 89-100.

Tsui, A., Shalaby, A., 2006. Enhanced system for link and mode identification for personal travel surveys based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 38–45.

Wachs, M. (1967). Relationships between drivers' attitudes toward alternate routes and driver and route characteristics. *Highway Research Record*, 197, 70-87.

Winters, M., Teschke, K., Grant, M., Setton, E., & Brauer, M. (2010). How far out of the way will we travel? *Transportation Research Record: Journal of the Transportation Research Board*, 2190, 1-10.

Zhang, L., & Levinson, D. (2008). Determinants of route choice and value of traveler information: a field experiment. *Transportation Research Record: Journal of the Transportation Research Board*, 2086, 81-92.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the web. Paper presented at the 17th World Wide Web Conference, Beijing, China.