**Final Report**
**to the**
**CENTER FOR MULTIMODAL SOLUTIONS FOR CONGESTION MITIGATION (CMS)**

CMS Project Number: 2011-008

# Title: Route-choice Modeling using GPS-based Travel Surveys

for period May 2011 – December 2012

Sivaramakrishnan Srinivasan (PI)
Associate Professor
Department of Civil and Coastal Engineering
University of Florida
siva@ce.ufl.edu

Nagendra Singh Dhakar
Graduate Research Assistant
Department of Civil and Coastal Engineering
University of Florida
nsdhakar@gmail.com

June 25, 2013

DISCLAIMER AND ACKNOWLEDEGEMENT OF SPONSORSHIP

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

This work was sponsored by a grant from the Center for Multimodal Solutions for Congestion Mitigation, a U.S. DOT Tier-1 grant-funded University Transportation Center.

TABLE OF CONTENTS

# LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The advent of GPS-based travel surveys offers an opportunity to develop empirically-rich route-choice models. However, the GPS traces must first be mapped to the roadway network, map-matching, to identify the network-links actually traversed. For this purpose, two enhanced map-matching algorithms are implemented and compared. Next, the choice set must also be constructed by identifying possible alternate routes between the origin and destination. This is accomplished using an enhanced version of the Breath First Search Link Elimination (BFS-LE) algorithm. The data assembled from the two steps, map matching and choice set generation, are then used for developing route choice using the path-size logit structure. The GPS data from the Chicago Travel Survey are used in this analysis. In addition to travel time, the number of intersections, turns, and the circuity of the route and the proportion of route by facility type were found to be statistically-significant predictors of route choice. In addition, the sensitivity to these factors also varied depending on trip purpose, time of the day of the trip, and traveler characteristics.

# EXECUTIVE SUMMARY

The advent of GPS-based travel surveys offers an opportunity to develop empirically-rich route-choice models. However, the GPS traces must first be mapped to the roadway network, map-matching, to identify the network-links actually traversed. In the study, two enhanced map-matching algorithms are implemented and compared for their operational performance using data from a large-scale GPS survey. Once the traversed path is determined, the next step is to determine the other options (routes), choice set generation, that were available to the traveler for making the trip. For this, the enhanced version of the Breath First Search Link Elimination (BFS-LE) algorithm is implemented. The data assembled from the two steps, map matching and choice set generation, are then used for developing route choice.

The primary source of data for this study is the GPS component (in-vehicle GPS data only) of the Chicago Regional Household Travel Inventory (CRHTI). In this survey, a GPS data-logging device (GeoLogger) was used to record the date, time, latitude, longitude, speed, heading, altitude, number of satellites, and horizontal dilution of precision (HDOP) at 1-second intervals. Original data comprises of GPS streams for 9941 auto trips made by 408 household vehicles. After eliminating trips shorter than 5 minutes in duration and 2 miles in distance, the sample consists of 5294 auto trips. However, the trips with unique OD pairs were retained for the final sample. The two map-matching algorithms generated routes for 3885 trips. After mapping the person demographics and trip characteristics, the sample comprised of 2850 trips. However, with available computational resources and time, choice sets were generated for 2143 trips. Further, 1913 trips which had at-least 15 alternatives in the choice set were included for the model estimations.

The Path Size Logit (PSL) model is used for developing models for route choice. The PSL models are developed for three different choice set sizes (15 alternatives, 10 alternatives, and 5 alternatives). The utility functions are expressed in terms of route attributes, trip characteristics and traveler characteristics. The estimation results indicate intuitive effects. Specifically, free-flow travel time, left turns, right turns, intersections, and circuity were found negatively associated with the attractiveness of a route. A positive sign on the path size attribute indicates that the route with less similarity with the alternatives is more likely to be chosen. Trips going to home are the least sensitive to the travel time and right turns than the other trips. Compared to home-based trips, non-home-based trips are less sensitive to intersections and time on local roads. On average, the expected overlaps (probabilistic routes) with the chosen route are similar to the deterministic overlaps (shortest time path). Also, there is a probability of about 50% that the predicted route will outperform the shortest time path.

We envision this study as an important contribution towards the development of empirically rich route choice models. With increasing numbers of GPS surveys and benefits of using high-resolution roadway network, the availability of computationally efficient automatic procedures to generate the chosen routes and alternatives is critical. Further, the examination of route choice behavior in terms of travelers' demographics provides more insight into the route choice decisions.

# CHAPTER 1
# INTRODUCTION

Understanding the travelers' choice of routes is critical in determining the volume of traffic (level of congestion) on the different links of the roadway network. One of the important impediments to studying route-choice behavior is that the data on actual routes chosen are never collected in conventional household travel-surveys. Arguably, the primary reason for the lack of such data is that routes cannot be easily reported in the Computer-Assisted-Telephone-Interview (CATI) methods used for data collection. As a consequence of this lack of data, route-choices are predicted assuming that travelers choose the shortest-travel time paths for their trips. While travel time is a very important factor that determines the choice of route, it is reasonable to expect that it is not the only factor considered by travelers in their route-choice decisions (see for example, Li et al., 2006 and Papinski and Scott, 2011a). The advent of Geographic Position Systems (GPS)-based travel surveys now provides an approach to trace vehicle movements and, hence, collect data on the actual routes chosen for various trips.

The efforts on the empirical modeling of route choices using GPS traces are recent and growing in number (see for instance, Schussler, 2009 and Hood et al., 2010). This is probably because the number of GPS-based travel surveys has increased quite substantially only in the last decade (consequently some of the past route-choice studies have relied on synthetic data or other sources such as travelers' self-description of routes). Further, several of the methodological developments relevant to modeling route choices from GPS-based travel surveys are relatively recent.

In this context, the broad focus of this research is to combine data from GPS-based travel surveys and Geographic Information Systems (GIS)-based roadway network databases to develop models for route-choice. There are three main components to the overall approach: (1) Map Matching, (2) Choice Set Generation, and (3) Route Choice Models.

Map matching is the process of identifying the specific links of the roadway traversed by a vehicle by mapping the points from its GPS trace to an underlying GIS-based roadway network database. This step is critical as it identifies the fundamental "choice" (i.e., the route) of interest. In this study, two map-matching algorithms from the literature are enhanced, implemented, compared, and validated. Both these algorithms include systematic treatment for missing GPS points along the routes; employ efficient techniques to address computational time; and are almost entirely automated.

Once the "chosen" path has been identified, the next step is to determine the other options (routes) that were available to the traveler for making the same trip. This process is called choice-set generation. Since the surveys do not directly query the respondents on alternate options available to them, the choice set is generally constructed by considering the network topology and the trip-end locations. In this study, and enhanced version of the link-elimination approach to generating the choice set is used. The procedure is shown to generate heterogeneous alternatives which are also generally inclusive of a significant proportion of the chosen route.

After assembling the data of the chosen routes and corresponding choice sets, route choice models are developed to examine the route choice behavior. In this study, the path-size logit approach is used. The explanatory factors include route attributes (such as travel times, numbers of turns, and number of intersections), trip attributes (time of the day, day of the week, home-based versus non-home-based), and traveler attributes (gender, length of stay at current residence, etc).

The rest of this document is organized as follows. Chapter 2 presents a review of the studies related to map matching, choice set generation, route choice models. Chapter 3 describes the map-matching algorithms used for generating observed routes. Both conceptual and implementation details are presented in the chapter. Chapter 4 presents a validation of the map-matching algorithm, and subsequently, the results of the application on a large dataset are discussed. Chapter 5 provides the assembly process of the estimation dataset. The conceptual framework of the choice set generation algorithm is also presented in the chapter. At the end, data descriptive are presented. Chapter 6 presents the route choice models developed in the study and discusses the estimation results. In the end, Chapter 7 presents the summary and conclusions of the study.

# CHAPTER 2
# LITERATURE REVIEW

While GPS-based travel-surveys collect data on vehicle trajectories, these data have to be processed substantially to be transformed into a format that can be used for model estimations. There are two major steps in this processing: Map Matching and Choice Set Generation. Map matching matches a stream of GPS points to a roadway-network database to identify the traversed links in the chosen route. Once the chosen route is determined, choice-set generation methods are used to determine possible alternatives that could have been considered by the decision maker. The data on the chosen route and the choice alternatives are then merged with other available information such as trip- and traveler- characteristics for model estimations.

The next three sections provide a review of the existing studies in the areas of map matching, choice-set generation and route choice modeling. A tabular list of studies is presented in Appendix A. The reader is also referred to Dhakar (2012) for further discussions. Each section ends with a short discussion of the contribution of this study.

## 2.1 Map Matching

Based on a review of the literature, there are two broad classes of algorithms for map matching (off-line map matching of high-frequency GPS streams): The GPS-weighted shortest-path algorithms (GWSP) and the multi path algorithms (MP).

Zhou and Golledge (2006) suggested that offline algorithms use optimization techniques such as the shortest path to generate a topologically-correct route and to exploit roadway attributes, such as speed limits, one-way streets for better accuracy. These suggestions are also in line with the work undertaken by Du (2005) who proposed a method that predicts the chosen route by determining the shortest path satisfying network topology such as link location, connectivity, one-ways, and allowable u-turns. The method was implemented in ArcGIS and examined against 674 trips collected on 18 known routes of Lexington, KY. For a known OD pair, approximately 95% of the routes were constructed entirely. However, high computational times and manual interventions are also characteristic features of this study.

In a recent study, Griffin et al (2011) demonstrated another way to use shortest path for path creation. First, an initial route was obtained by inputting selected GPS points to a driving directions (DD) service, offered by web service providers such as MapQuest, Yahoo, and Google. The DD service calculates the shortest cost path for the input GPS points. After obtaining the initial route, problematic waypoints were identified using a set of rules that includes point distance, bearing, path ratio, and duplicate points. The revised set of waypoints was then resubmitted to DD services and a correct map-matched route was obtained. Algorithm was tested against a real world GPS data for 200 routes and a visual inspection of the routes confirmed an accuracy of 100%. However, the method involved spending a large amount of time in identifying problematic routes manually.

Similar drawbacks (manual methods for correcting for map matching errors) were observed in the method used by Spissu et al (2011). First, a spatial join was used to match the GPS points to the corresponding routes, and afterwards a manual inspection corrected the matching errors. The method was run for 697 trips and it found routes only for 58% of the trips. Unmatched trips were due to missing GPS points, inconsistent activity data and missing links in the roadway network.

To improve the efficiency of a map-matching process Marchal et al. (2005) proposed an algorithm that uses a multiple hypothesis technique (MHT), which was first introduced by Pyo et al. (2001) for the application of on-line map-matching. The MHT stores multiple paths during the process and in the end selects a path with the best score.

Marchal et al. (2005) calculated a “link score” by determining the distance of the link to the GPS point. The algorithm starts with finding a set of links that are closest to the first GPS point. For each link, a new path is created and links are inserted with their scores assigned to the respective paths. GPS points are processed in order (time sequence) and a path score is updated by adding the score of the last added link to its previous score. When the end of a link is reached, a copy of the path is created for each outgoing link and then the path is removed from the set of paths. In the end, the path (in the set of paths) with the lowest score is selected as the traversed path. Algorithm limits number of paths in a set to 30. Algorithm’s efficiency was not evaluated in terms of correctly identified links; instead focus was given to the operational performance using a real world data of 84 paths collected in the Zurich area. The authors argued that the accuracy and the running time are dependent on the maximum number of candidate links/ paths stored in the set. However, the algorithm couldn’t produce continuous routes in most cases, which the authors reasoned are because of irregular GPS streams caused by tunnels, tree canopies, poor signal, and so forth. As a result, a sequence of paths was generated instead of a continuous route. The authors also added that algorithm is sensitive to outliers in GPS data.

Schüssler and Axhausen (2009) modified the original algorithm by Marchal et al. (2005) to overcome their limitation of not producing a continuous route. Additionally, a modified method was used to calculate a score. First, they subdivided each trip into continuous segments depending on the gaps in GPS streams. Afterwards, they created the trip segments by using the algorithm by Marchal et al. (2005). Then, a complete trip was obtained by connecting trip segments through a shortest path search with a treatment for low quality map matching results. During the study, 3932 car trips encompassing 2.4 million GPS tracks were matched to a high resolution Swiss NAVTEQ roadway network. The results showed a smaller number of matched routes in comparison to the total routes. Further investigation of the results showed three main reasons for such low numbers of matched routes: missing links in the roadway network, off-network travel, and u-turns. Menghini et al. (2010) applied the algorithm, developed by Schüssler and Axhausen (2009), to match 320, 576 GPS points of 3387 bike trips, also extracted from the same data source. However, they encountered errors for cases of missing good point flow and a dense scatter of points.

Zhu and Golledge (2006) used the MHT with rank aggregation and proposed a three-step map-matching algorithm. Prior to the map matching, GPS data was processed for cluster reduction and density leverage. After map matching results, a Dempster belief test was used to detect the noise and off-road travel. A combination of accumulated 2-norm distance and rotational variation metric was used to decide the rankings of the candidate paths. The author did not apply the algorithm to real world data.

In summary, the GWSP algorithm directly uses the concept of shortest path in determining the route. However, the links that are not close to observed GPS points are provided higher impedances (making them less likely to be included in the shortest path) than the links that are near observed GPS points (these links have the true travel times/costs as impedances). The MP algorithm does not use the concept of shortest paths; rather, it traces though the stream of GPS points identifying all the possible routes to reach the destination from the origin.

The GWSP algorithm is more straightforward and computationally less-demanding (especially if a tool for calculating shortest-paths is available) whereas the MP algorithm is more elaborate and demanding (the need to store multiple paths can get cumbersome with dense networks). However, the latter algorithm is also free from assumptions such as preference for shortest paths and generally uses the observed data to determine the route. A comparative analysis of these approaches would therefore be of interest and this study contributes towards that end. Enhanced versions of both GWSP and MP algorithms are implemented and compared. The enhancements are aimed at achieving complete automation and better operational performance. Both algorithms are implemented in ArcObject within ArcGIS framework, using Python and Visual Basic Application (VBA).

## 2.2 Choice Set Generation

Once the chosen route has been determined, the next step is to determine the set of alternate paths available for the same trip. The universal choice set contains all possible paths between an OD pair. However, it is impossible for a traveler to be aware of all the paths in the universal choice set. Further, this universal choice set would contain high number of unattractive and unrealistic routes that a traveler would never consider during the decision-making. The inclusion of these unrealistic routes in a choice set would put an extra burden on computation time and also affect the model estimations. Therefore, a “consideration” choice set is defined as a subset of the universal choice set and contains only feasible and attractive paths. Bovy (2009) defined this choice set as the collection of travel options perceived as available by individual travelers in satisfying their travel demand. Since the surveys do not directly ask the respondents to provide information on the options available/considered, the choice sets have to be determined using the roadway network characteristics and reasonable behavioral rules. Various restrictions are applied to ensure that the number of options in the choice set is reasonable and the options themselves are somewhat different from each other.

In general, the choice-set generation approaches can be classified into three categories: (1) shortest path based methods, (2) constrained enumeration methods, and (3) probabilistic methods

**Shortest-path based methods**

Shortest-path based approaches are the most popular and commonly used methods in the literature. For a given generalized cost, this method repeatedly searches for the alternate shortest-cost path in the network. The search for the shortest path can be approached in two ways: deterministic and stochastic.

**Deterministic shortest-path based methods:** The popular algorithms in the deterministic shortest-path based methods include *k-shortest path, labeling, link elimination, and link penalty*.

*K-shortest path* algorithms extend the idea of calculating a single shortest path (e.g. Dijkstra, 1959) to determine k-shortest paths using a generalized link cost function. Recently, Papinski and Scott (2011b) generated choice sets by calculating 9 shortest-time paths for a GPS dataset of 237 home-based work trips collected for auto drivers in Halifax, Nova Scotia, Canada. Spissu et al. (2011) calculated 10 minimum-cost paths using cost functions from the existing Cagliari model. Over the years, researchers have introduced several variations to the basic approach of finding k-shortest paths while maintaining the same computational efficiency. Kuby et al. (1997) construct the choice sets by, iteratively, selecting routes from a subset of k-shortest paths that satisfy a similarity measure, whereas Van der Zijpp and Fieorenzo-Catalano (2005) find feasible paths satisfying some behavioral constraints.

Instead of calculating multiple shortest paths for one cost, the *labeling approach* finds one optimal path for each of several costs or attributes (labels). The number of paths in a choice set is equal to the number of labels considered. Ben-Akiva et al. (1984) proposed this approach and generated routes for 10 labels ( time, distance, scenic, signals, capacity, hierarchical travel pattern, quality of pavement, commercial development, highway distance, and congestion). Routes with only two labels, time and distance, replicated 70% of the chosen routes and routes with all labels together replicated 90% of the chosen routes. The study found that signals are not a significant factor and concluded that factors other than time and distance do play a significant role in route choice. Ramming (2002) used more attributes (e.g. time in secure neighborhood, tolls, left turns, free flow time), totaling 16 labels, to generate paths for 188 observations (91 OD pairs) collected through a web-based survey of faculty and staff of Massachusetts Institute of Technology (MIT), Boston. For 236 trips (182 OD pairs) from another web-based survey conducted in Turin, Italy, Prato and Bekhor (2006) (also Prato and Bekhor, 2007) determined paths by using 4 labels – distance, free-flow time, travel time, and delay. Bekhor and Prato (2009) compared the two datasets, Boston and Turin, by using labeling approach with 5 labels: distance, free-flow time, delay, traffic lights, and traffic lights. Quattrone and Vieteatta, 2011 examined data from a road-side survey of truck drivers and generated 30 routes for each of 5 costs: minimum travel time, minimum monetary cost, maximum motorway route, minimum bridges and viaducts, minimum routes with high levels of accidents. The generated routes overlap 75% of the observed routes.

*Link elimination* approaches, presented by Azevedo et al. (1993), iteratively search for the next best path by removing one or more links from the shortest cost path. Bekhor at el. (2006), Prato and Bekhor (2006), Frejinger and Bierlaire (2007) calculated multiple alternate paths by using 50, 10 and 15 iterations respectively. Using this approach, Ramming (2002) obtained up to 49 unique routes. Schussler et al. (2012) proposed a variation to the approach, called as breadth-first search link elimination (BFS-LE). It starts with calculating a shortest cost path between origin and destination and searches for the next shortest path by removing links. The resulting shortest paths are set as the starting points, nodes, for next iteration, depth, of link elimination. All nodes at a depth are processed before moving to the next depth. They also proposed two performance optimization methods: a randomization of links at a depth and a roadway topology simplification. More details of the method are provided in the methodology section. Menghini et al. (2010) implemented the BFS-LE approach for route choice of cyclists in Zurich, Switzerland.

*Link penalty* approaches, introduced by De la Barra et al. (1993), also iteratively determines multiple shortest paths. However, instead of removing links a penalty is imposed on impedances of the links in the current shortest path. Studies have used several methods of determining link penalties. Park and Rilett (1997) increased penalties only to the links that are outside a certain distance from the origin and destination of a trip. The variation resulted in less similar and more relevant routes. Scott et al. (1997) incorporated an optimization program to determining the penalty factor. Bekhor et al. (2006) defined penalty as the function of distance between origin and destination, therefore, introducing higher penalty for longer routes. Prato and Bekhor (2007) used a fixed penalty factor and iterated the process 15 times.

**Stochastic shortest-path based:** These approaches assume that the path costs are not deterministic and travelers' observe the costs with error. Also called as the simulation approach the error is represented by drawing generalized cost functions from probability distributions. Ramming (2002) analyzed home to work commute and extracted 48 draws from a Gaussian distribution with mean and standard deviations to equal to link travel times. Beirlaire and Frejinger (2005) extracted 20 draws from a truncated normal distribution with mean and variance equal to link travel times from the observed data. Average choice set size was 9.3 with maximum and minimum 22 and 2 routes respectively. Prato and Bekhor (2006), Prato and Bekhor (2007), and Bekhor and Prato (2009) implemented two simulation approaches exploiting the same procedure to draw impedances. Twenty-five and 35 draws were extracted from a truncated normal distribution with mean equal to travel time and variance equal to a percentage of the mean, 20% and 100% respectively. Left truncation limit was set equal to the free-flow time and right truncation limit was equal to the travel time calculated for a minimum speed assumed equal to 10 km/hr. Schussler et al. (2012) draw impedances from a truncated normal distribution with mean equal to travel time and standard deviations equal to different multiples of travel time. The Doubly stochastic method, proposed by Bovy and Fiorenzo-Catalano (2007), is an extension to the simulation approach where generalized cost functions are in the form of utilities with both the parameters and the attributes are stochastic. Clearly, the use of stochastic shortest path methods requires information on the (perceived and real) variability of travel times on the roadway network, which may not always be readily available.

### Constrained enumeration methods

Constrained enumeration methods construct choice sets using a set of constraints that reflect cognitive, perceptual, behavioral assumptions. Pillat et al. (2011) proposed a method with path enumeration and branch-cutting criteria. The method included a commonly-factor criteria within the generation process to keep the choice set size computable and to avoid routes with higher detour factors. The commonly factor (CF) calculates the overlap between two routes in terms of distance and is calculated as $\mathrm{CF} = \mathrm{L_{ij}}/\sqrt{\mathrm{L_i L_j}}$. Where $\mathrm{L_{ij}}$is the common distance of route $\mathrm{i}$ and route $\mathrm{j}$; $\mathrm{L_i}$is the distance of route $i$ and $L_j$ is the distance of route $j$. For the route tree, origin is the source of tree and branches are the routes in the network. Branch-cutting criteria are used to control detour factors in the route and checked for every new node added as a branch element to the route tree. Parameters of the allowed impedance functions, used to apply detour factors, were estimated using the information on known alternative routes collected during the survey. The final choice set produced reasonable routes and replicated 60% of the observed routes which authors considered promising given the uniqueness of the trip observations. Prato and Bekhor (2006) introduced Branch and Bound algorithm that enumerates paths by constructing a tree connecting origin and destination of a trip. A set of constraints, such as directional, temporal, loop, similarity, and left turn are satisfied while processing sequence of links to generate the tree. Friedrich et al. (2001) and Hoogendoorn-Lanser (2005) applied the branch and bound method for transit network and multimodal network context respectively. Bekhor and Prato (2009) also used the approach for his comparative study of two datasets: MIT, Boston and Turin, Italy.

**Probabilistic methods**

In contrast to the deterministic methods, where an alternative either belongs to a choice set or not, probabilistic methods, first proposed by Manski (1977), also represent intermediate availabilities by assigning perceived probabilities to routes. This set of approaches relies on the assumption that all routes connecting origin and destination belong to the choice set to some degree.

Cascetta and Papola (2001) construct fuzzy choice sets with the proposed Implicit Availability/ Perception (IAP) model that represents alternative's availability/membership in terms of continuous values ranging from 0 to 1. Ramming (2002) applied IAP logit with variables related to network knowledge but obtained results were not satisfactory. Frenjinger (2007) and Frejinger et al. (2009) extract a subset of paths with importance sampling approach, which selects attractive alternatives with higher probability. For an OD pair, the probability of each link in the network is calculated based on its deviation from the actual shortest path. Therefore, links on the actual shortest path have a link probability of 1 and other links between 0 and 1. Next, a repeated random walk method, starting at the origin, selects and adds links from node to node until destination is reached. The link selection process at a node is determined by the associated next links' probabilities. This approach was applied to a synthetic network with 38 nodes and 64 links and positive results were obtained confirming the superiority of models with sampling corrections than others with no correction.

The discussion thus far focused on methods to generate alternate paths. Next we focus on metrics used to assess the quality of the generate choice-sets. Different guidelines are used to assess the structure and quality of the generated choice set or the effectiveness of an algorithm. Typically, choice set size (Richardson, 1982; Prato and Bekhor, 2006) and coverage of the observed routes are used to evaluate the choice set composition. Bekhor et al. (2006) define coverage as the share of observations for which an algorithm produced a route that meets a particular threshold of overlap. Where, the overlap is usually the percentage of the observed route distance. In addition to coverage, Prato and Bekhor (2007) and Prato and Bekhor (2007) calculated a consistency index that compares a choice set generation method with the ideal algorithm that would reproduce all the observed routes. Schussler et al. (2012) used four measures to assess the structure of the choice set: choice set size, reproduction of the observed route (coverage), route diversity, and hierarchical sequence. The route diversity determines how different the routes are in the choice set. Two indicators are used for the measure: overlap among the routes and distribution of the route distances. The hierarchical sequence focuses on the shares and sequences of the road types.

A number of studies have compared choice-set generation methods in terms of computational performance and efficiency.

Papinski and Scott (2011b) compared two choice set generation algorithms using a sample of 237 home based work trips. The first algorithm is a constrained enumeration method. A potential path area (PPA) was defined for an OD pair. Nine links were randomly selected as the midpoints for each alternate route within the PPA. For each midpoint link, a shortest path based on time is found through from the origin (home) to midpoint link to destination (work). The second algorithm (k-shortest path) calculates nine shortest time paths for an OD pair. Constrained enumeration method performed better than the k-shortest path, thus emphasizing the fact that travelers' route choice do not depend only on the travel time.

Ramming (2002) (also, Bekhor et al., 2006) examined four choice set generation algorithms: labeling, link elimination, link penalty, and simulation. Computational time for link penalty approach was high considering the small number of routes generated in the choice set (also, see Prato and Bekhor, 2007). The high computation time was a result of rewriting the changes the link impedances back to the network. However, as Prato (2009) stated, in link penalty approaches, the value of penalty factor plays a crucial part. While low penalty may result in very similar routes, high penalty may produce undesirable routes. For link elimination, Ramming (2002) was not convinced that removal of "one link" at a time would produce quality choice sets. Prato (2009) also agreed with it, and pointed out that the major shortcoming of the link elimination is the network disconnection, as removing centroid connectors and major crossing does not promise a new route between origin and destination. Also, removing a link, generally, introduces only a short deviation from the previous route and the generated routes are somewhat similar. Ramming (2002) found that labeling approach was the fastest and three labels – distance, free-flow time, and time – provided enough coverage. Simulation approach, 48 draws from a Gaussian distribution, was easy to implement and showed acceptable computational performance. The final choice sets were constructed by merging routes from the labeling (distance, free-flow time and time) and simulation approach. The final choice sets consist of maximum 51 routes and a median size of 30 routes.

Prato and Bekhor (2006) (also, Bekhor and Prato, 2009, and Prato and Bekhor, 2007) expanded the comparison by Ramming (2002) with an inclusion of branch and bound approach. Among all the algorithms, branch and bound showed higher consistency index (97.1%) and coverage (up to 97%). Therefore, two choice sets were constructed during the study; first, merging routes from different algorithm, and second, routes from branch and bound algorithm. First choice set contained maximum of 55 routes with median size of 32 routes. Second choice set contained maximum of 44 routes with median size of 17 routes. Comparison of two choice sets indicated a better performance and more heterogeneous routes from branch and bound approach.

Based on the recommendations of previous studies, Schussler et al. (2012) selected three algorithms that have performed well and promised to produce realistic routes: stochastic, branch and bound (BB), and constrained random walk (CRW). Additionally, they proposed a breadth first search link elimination (BFS-LE). Before, implementing the BFS-LE approach, they compared the computational performances of the four algorithms on a smaller sample of 500 OD pairs, representative of the main sample. BB and CRW demonstrated long computational times even for small alternatives and short paths. Each of the two algorithms was run for 17 days with a processing time limit of 90 minutes per OD pair. During the time limit, BB couldn't find any route for 229 OD pairs and for 161 OD pairs it found only one route. However, CRW found at-least 5 routes for 466 OD pairs. High computation times and smaller choice set sizes discourage the use of these algorithms for the application of high-resolution networks. Similar concerns were raised by Prato (2009), who observed that the computational performance of the branch and bound approach highly depends on the depth of the tree and thus, on number of links in the paths. With this, the author anticipated that the application of the algorithm would be limited to small networks. The author also suggested that the routes generated with random walks approach may be very circuitous, contain loops and extremely long since they do not reach the destination in a reasonable number of stops, thus not making the method suitable for estimation and prediction purposes.

During the study by Schussler et al. (2012), stochastic and BFS-LE methods showed considerably better results in terms of computation time and number of alternatives. The two algorithms, stochastic and BFS-LE, constructed the choice sets for 500 OD pairs in 12 days and 7.1 days respectively. The authors further compared the computational efficiency of these two methods and found that, on average, computation time for stochastic method was 32 times higher than the BFS-LE method. As pointed out by Prato (2009), efficiency of stochastic/ simulation approach depends on the selection of probability function and number of draws. The popular distribution includes normal, log-normal, gamma etc. Perceived cost being non-negative results in truncating negative draws in the normal distribution. However, truncation may lead to biases towards certain routes (Nielsen, 2000). Log-normal and gamma distributions guarantee non-negative draws (Nielsen and Frederiksen, 2006), and thus preferred. Higher number of draws results in increasing computation cost considerably, and do not necessarily generate higher unique routes (Ramming, 2002). Prato and Bekhor (2006) experimented with number of draws and variance of the distribution. They found that both low and high variances are not efficient as former results in fewer unique paths and later produces several unrealistic paths.

Prato (2009) also highlights the shortcomings of the doubly stochastic method in the calibration of the probability function coefficients and indicated that the use of incorrect values could lead to unrealistic routes.

Schussler et al. (2012) observed that the percentage of reproduced chosen routes from BFS-LE was higher than the results of the simple link elimination method, presented in the studies by Ramming (2002) and Prato and Bekhor (2007). The authors argued that for high-resolution networks repeated shortest path search algorithms are more feasible and perform the best. The proposed BFS-LE method produced better results than the basic link elimination method and outperforms the stochastic method in terms of computational efficiency.

Based on the discussion so far and the nature of our dataset, we decided to implement an enhanced version of the BFS-LE algorithm to generate choice sets for large number of trips in a high-resolution network. The enhancements aim to address computational efficiency and the need to generate heterogeneous routes. Specifically, only those links that are likely to lead to a "different" route alternative when removed are considered for elimination. It is also ensured that any new route generated does not substantially overlap with any of the route alternatives already generated. The algorithm is applied to generate route-alternatives for over 2000 trips on the dense Chicago roadway network.

### 2.3 Route Choice Models

The random utility discrete-choice models are the most commonly used approach for analyzing route-choice decisions. Such models assume that the utility of an alternative consists of two components: deterministic and stochastic. Specifically, the utility of alternative $i$ in the choice set $C_n$ perceived by individual $n$ is give by:

$$U_{in} = V_{in} + \varepsilon_{in}$$

Where, $V_{in}$ is the deterministic or observed component, and $\varepsilon_{in}$ is the stochastic or unobserved component.

For choice modeling, logit-based models are most commonly used. Among the family of logit models, the Multinomial Logit Model (MNL) is the simplest one. For the MNL model, the probability of choosing an alternative $i$ in choice set $C_n$ is given by:

$$P(i/C_n) = \frac{\exp(V_{in})}{\sum_{j\epsilon C_n} \exp(V_{jn})}$$

The MNL model is based on the assumption of Irrelevance of Independent Alternatives (IIA), and therefore, does not consider the similarities between alternatives. The similarity of a route with other alternatives may affect the utility of choosing the route and is needed to be accounted in the choice models to have a more realistic representation of the travel behavior.

Several models have been proposed in the literature to overcome the limitation of the MNL model. The next section presents a brief description of such models (for further discussions, refer to Prato, 2009).

The computational benefits of the simple, closed-form MNL model structure have encouraged researchers to propose MNL modifications to capture the similarities among routes. The modifications are either made in the deterministic or the stochastic part of the utility.

Methods that modify the deterministic part of the utilities include the C-logit, the Path Size Logit (PSL), and the Path Size Correction Logit (PSCL) models.

**C-logit**: C-logit model, proposed by Cascetta et al. (1996), was one of the first MNL modifications. The model introduced a term, commonly factor, in the deterministic part of the utility that measures the physical overlap of a route with other routes in the choice set. The commonly factor (CF) reduces the utility of a route due to its similarity with other routes. The probability of choosing an alternative in choice set $C_n$ is given by:

$$P(i/C_n) = \frac{\exp(V_{in} + \beta_{CF} * CF_{in})}{\sum_{j\epsilon C_n} \exp(V_{jn} + \beta_{CF} * CF_{jn})}$$

Several formulations of CF are proposed in the literature (Cascetta et al., 1996, Cascetta et al., 2001):

$$CF_i = \ln \sum_{j\epsilon C_n} \left(\frac{L_{ij}}{\sqrt{L_i L_j}}\right)^{\gamma}$$

$$CF_i = \ln \sum_{l\epsilon \Gamma_i} \left(\frac{L_l}{L_i} \ln \sum_{j\epsilon C_n} \delta_{lj}\right)$$

$$CF_i = \sum_{l\epsilon \Gamma_i} \left(\frac{L_l}{L_i} \ln \sum_{j\epsilon C_n} \delta_{lj}\right)$$

$$CF_i = \ln \left[1 + \sum_{\substack{j\epsilon C_n \\ i\neq j}} \left(\frac{L_{ij}}{\sqrt{L_i L_j}}\right) \left(\frac{L_i - L_{ij}}{L_j - L_{ij}}\right)\right]$$

Where, $L_i$ is the length of route i, $L_j$ is the length of route j, $L_l$ is the length of link l, $L_{ij}$ is the common length between route i and route j, $\gamma$ is a parameter to be estimated, and $\delta_{il}$ is the link path incidence dummy, which is equal to 1 if route i uses link l and 0 otherwise.

**Path size logit (PSL):** Ben-Akiva and Bierlaire (1999) proposed the PSL model and measured the similarity using a Path Size term in the deterministic component. The path-size indicates the fraction of the path that constitutes a “full” alternative (Ben-Akiva and Bierlaire, 1999).

$$P(i/C_n) = \frac{\exp(V_{in} + \beta_{PS} * \ln PS_{in})}{\sum_{j\epsilon C_n} \exp(V_{jn} + \beta_{PS} * \ln PS_{jn})}$$

Several formulations of the Path Size have been presented in the literature.

$$PS_i = \sum_{l\epsilon \Gamma_i} \frac{L_l}{L_i} \left(\frac{1}{\sum_{j\epsilon C_n} \delta_{lj}}\right)$$

$$PS_i = \sum_{l\epsilon \Gamma_i} \frac{L_l}{L_i} \left(\frac{1}{\sum_{j\epsilon C_n} \left(\frac{L_l}{L_i}\right)^{\gamma} \delta_{lj}}\right)$$

Where, $L_i$ is the length of route i, $L_l$ is the length of link l, $\gamma$ is a parameter to be estimated, and $\delta_{lj}$ is the link path incidence dummy, which is equal to 1 if route j uses link l and 0 otherwise.

Frejinger et al. (2009) proposed Expanded PS (EPS), which includes an expansion factor to the PS attribute that corrects for the sampling:

$$PS_i = \sum_{l\epsilon \Gamma_i} \frac{L_l}{L_i} \left(\frac{1}{\sum_{j\epsilon C_n} \delta_{lj}\, \Phi_{jn}}\right)$$

$$\Phi_{jn} = \begin{cases} 1 & if\ \delta_{jc} = 1\ or\ q(j)R_n \geq 1 \\ \dfrac{1}{q(j)R_n} & otherwise \end{cases}$$

Where, $q(j)$ is the sampling probability of path j, and $R_n$ is the number of alternatives (excluding the chosen route) drawn to form the choice set.

**Path size correction logit (PSCL):** Bovy et al. (2008) argued that the PS attribute in the PSL model does not have a theoretical derivation and assumptions are not clearly stated either. In response, they proposed a Path Size Correction Logit (PSCL) model and presented a systematic derivation of the Path Size Correction term.

$$P(i/C_n) = \frac{\exp(V_{in} + \beta_{PSC} * PSC_{in})}{\sum_{j\epsilon C_n} \exp(V_{jn} + \beta_{PSC} * PSC_{jn})}$$

$$PSC_i = -\sum_{l\epsilon\Gamma_i}\left(\frac{L_l}{L_i}\ln\sum_{j\epsilon C_n}\delta_{lj}\right)$$

Where, $L_i$ is the length of route i, $L_l$ is the length of link l, and $\delta_{lj}$ is the link path incidence dummy, which is equal to 1 if route j uses link l and 0 otherwise.

Models that account for similarities in the stochastic part of the utility (error correlations) while still maintaining a closed-form formula for probabilities fall in the family of Generalized Extreme Value (GEV) models. Such models include Paired Combinatorial Logit (PCL), Cross Nested Logit (CNL), and Generalized Nested Logit (GNL).

**Paired combinatorial logit (PCL):** The model assumes that choice decisions are made from pair of alternatives in a choice set. Prashker and Bekhor (1998) provided a formulation of PCL model in route choice context:

$$P(i/C_n) = \sum_{i\neq j} P(ij)P(i|ij)$$

$$P(i|ij) = \frac{\exp(\frac{V_{in}}{1-\sigma_{ij}})}{\exp\left(\frac{V_{in}}{1-\sigma_{ij}}\right) + \exp\left(\frac{V_{jn}}{1-\sigma_{ij}}\right)}$$

$$P(ij) = \frac{(1-\sigma_{ij})\left(\exp\left(\frac{V_{in}}{1-\sigma_{ij}}\right) + \exp\left(\frac{V_{jn}}{1-\sigma_{ij}}\right)\right)^{1-\sigma_{ij}}}{\sum_{k=1}^{n-1}\sum_{m=k+1}^{n}(1-\sigma_{km})\left(\exp\left(\frac{V_{kn}}{1-\sigma_{km}}\right) + \exp\left(\frac{V_{mn}}{1-\sigma_{km}}\right)\right)^{1-\sigma_{km}}}$$

Where, $P(i|ij)$ is the conditional probability of selecting route i provided that the pair (i,j) is chosen, $P(ij)$ is the unobserved probability of selecting the pair (i,j), and $\sigma_{ij}$ is the similarity coefficient between i and j. Two different formulations of the similarity coefficients are provided in the literature: by Prashker and Bekhor (1998), and by Gliebe et al. (1999). The two formulations are presented below in respective order:

$$\sigma_{ij} = \left(\frac{L_{ij}}{\sqrt{L_i L_j}}\right)^{\gamma}$$

$$\sigma_{ij} = \frac{L_{ij}}{L_i + L_j - L_{ij}}$$

Where, $L_i$ is the length of route i, $L_j$ is the length of route j, $L_{ij}$ is the common length of route i and route j, and $\gamma$ is a parameter to be estimated.

**Cross nested logit (CNL):** The model assumes that choices are made within nests. In route choice context, the nests correspond to the links in the choice set (Prashker and Bekhor, 1998). Therefore, a route belongs to multiple nests. The choice probability is given by:

$$P(i) = \sum_{j} P(k)P(i|k)$$

$$P(i|k) = \frac{(\alpha_{ki}\exp(V_i))^{1/\mu_k}}{\sum_n (\alpha_{kn}\exp(V_n))^{1/\mu_k}}$$

$$P(k) = \frac{\left(\sum_i (\alpha_{ki}\exp(V_i))^{1/\mu_k}\right)^{\mu_k}}{\sum_m (\sum_n (\alpha_{mi}\exp(V_m))^{1/\mu_k})^{\mu_k}}$$

$$\alpha_{ki} = \frac{L_k}{L_i}\delta_{ki}$$

Where, $P(i|k)$ is the conditional probability of selecting route i in nest k, $P(k)$ is the unobserved probability of selecting the nest k $\alpha_{ki}$ is the inclusion coefficient, $\mu_k$ is the nesting coefficient, $L_k$ is the length of nest k, $L_i$ is the length of route i, and $\delta_{ki}$ is the link path incidence dummy, which is equal to 1 if route i uses link k and 0 otherwise.

**Generalized nested logit (GNL):** The model is a generalization of CNL model with the same formulation of inclusion coefficient but each nest has a different nesting coefficient. The formulation of the nesting coefficient is given by:

$$\mu_k = \left(1 - \frac{\sum_{l \epsilon C_n} \alpha_{kl}}{\sum_{l \epsilon C_n} \delta_{kl}}\right)^{\gamma}$$

Where, $\alpha_{ki}$ is the inclusion coefficient, $\gamma$ is a parameter to be estimated, and $\delta_{kl}$ is the link path incidence dummy, which is equal to 1 if route l uses link k and 0 otherwise.

Models that account for similarities in the stochastic part of the utility (error correlations without maintaining a closed-form formula for probabilities fall in the class of “mixed models”. In these models, the error terms is represented with two components. One part accounts for correlation and heteroscedasticity, and the other is i.i.d. extreme value. Such models include mixed logit, and logit kernel with a factor analytic.

**Mixed logit model:** The model, also known as logit kernel (LK), assumes random coefficients. The probability of choosing route i by individual n is computed by simulation and is given by:

$$P(i|C_n) = \frac{1}{D}\sum_{d=1}^{D} \frac{\exp(\beta_d' X_{in})}{\sum_{j \epsilon C_d} \exp(\beta_d' X_{in})}$$

Where, D is the number of draws, and $\beta_d'$ is $d^{th}$ draw from the distribution of $\beta$ . Different coefficient distributions used in the literature include uniform, normal, log-normal, and gamma distributions.

**LK with a factor analytic:** instead of assuming random coefficients, these models simply represent the error term with the components that accounts for similarities. Bekhor et al. (2002) provides a model, LK models with a factor analytic, which assumes that the covariance of the path utilities is proportional to the length by which paths overlap.

$$P(i|C_n) = \Lambda(i|\zeta) = \frac{\exp(\mu(X_i\beta + F_i T\zeta))}{\sum_{j\epsilon C_n} \exp(\mu(X_j\beta + F_j T\zeta))}$$

Where, F is the loading matrix, T is a diagonal matrix of covariance parameters, and $\zeta$is a vector of standard normal variables

A different approach of using LK model in route choice context was presented by Frejinger and Bierlaire (2007). They proposed an Error Component model, LK model with a Subnetwork, and captured the correlations among alternatives using a Subnetwork. The Subnetwork captures the similarities among alternatives for unobserved factors even if the alternatives are not spatially overlapped. The probability if calculated by simulation and is given by:

$$P(i|C_n) = \frac{1}{D}\sum_{d=1}^{D} \Lambda(i|\zeta^d)$$

Where, $\zeta^d$ is d$^{th}$ draw from the distribution of $\zeta$ and D is the number of draws.

Frejinger et al. (2009) compared two formulations of the PS attribute: original PS and expanded PS. The estimations, however using synthetic data, showed that the PSL model with expanded PS performs better than the original PSL model. Schussler and Axhausen (2010) compared different specifications of C-logit and PSL models for choice sets of different sizes and composition. The estimation dataset contained 1500 observations collected from an on-person GPS survey in Zurich, Switzerland. Estimation results indicated that PSL model with road type specific Path Size attribute provides the best similarity treatment. Bekhor et al. (2006) compared the PSL model with CNL model using 159 home-to-work observations collected through a questionnaire survey of faculty and staff at MIT, Boston. The estimations indicated that the CNL model is an improvement over PSL model and a better fit than the MNL model.

Bovy et al. (2008) estimated PSCL and PSL models for two datasets from different regions, Turin (228 observations) and Boston (181 observations). The estimation results suggested that the proposed PSCL model is a better-fit model than the PSL model. For the two models, predicted probabilities were also compared using a simple hypothetical network with 12 routes for a single OD pair. The results once again indicated the superiority of the PSCL model over PSL model. The authors recommended using PSL and PSCL models when all the relevant routes are present in the choice set.

Prato and Bekhor (2006) and Bliemer and Bovy (2008) compared different models from the families of logit and GEV models. Prato and Bekhor (2006) estimated MNL, C-logit, PSL, GNL, CNL, and LNL models using 236 observations (182 different ODs) of home-to-work route choice decisions collected through a web-based survey in Turin, Italy. Bliemer and Bovy (2008) estimated MNL, C-logit, PSL, PSCL, PCL, and CNL for a simple hypothetical network of a single OD pair with 12 available routes. Both studies indicated that the CNL model captures the similarity better than other models and performed the best. However, Bliemer and Bovy (2008) also examined the impact of different choice set compositions and sizes on route choice probabilities, and found that none of the estimated models was robust. The models were found sensitive to the presence of irrelevant routes in the choice sets. Also, the sensitivity was higher if the irrelevant route was more similar to the relevant routes in the choice set.

Bekhor et al. (2002) estimated LK with a factor analytic structure, MNL and PSL models using a dataset of 159 home-to-work observations collected through a questionnaire survey of faculty and staff at MIT, Boston. The LK model performed better than the MNL and PSL models. The study by Prato and Bekhor (2007) also adopted LK model with a factor analytic structure for investigating the impact of choice set composition on model estimates. In all, they estimated and compared six route choice models: MNL, C-logit, PSL, GNL, and CNL. The study recommended to use MNL modified models for large number of alternatives and nested or LK models for small alternatives in the choice sets.

Frejinger and Bierlaire (2007) showed an application of the error component model by using GPS data of 2978 observations (2170 OD pairs) collected in Borlange, Sweden and estimated five different specifications of Error Component model. The Subnetwork was constructed with 5 arbitrary chosen roads in the network. The results of the EC models were compared with the MNL and PSL models. Further, prediction performances of the models were also examined by randomly selecting 80% OD pairs for estimations and the rest for predicting choice probabilities. MNL and PSL models showed similar prediction performance, however, PSL resulted in a better fit for the estimated data. Bierlaire and Frjinger (2008) adopted the Error Component approach to demonstrate route choice modeling with network free data (GPS data, reported trips). They estimated the Error Component model for 780 observations collected via telephonic interviews in Switzerland. The Subnetwork was defined as consisting of all the main freeways in the roadway network of Switzerland. When compared with PSL model estimates, the Error Component model was found a better-fit model. Another empirical application of the Error Component Model was presented by Bekhor and Prato (2009). They also estimated MNL and PSL models to examine the methodology transferability in route choice modeling using two different datasets: Turin, and Boston. They advised to account similarities within the stochastic part of the utility, however, also warned that it would be more computationally expensive.

Distance and travel time (free-flow or estimated) are the common explanatory variables used in utility functions. Numerous specifications of the travel time are used by the researchers. Prato and Bekhor (2006) and Prato and Bekhor (2007) specified travel times for experienced and inexperienced drivers and found that the experienced drivers are concerned about the travel time more than the in-experienced drivers. Schussler and Axhausen (2010) and Bierlaire and Frejinger (2008) used travel times on each road type, defined according to an existing hierarchy of roadway links. Schussler and Axhausen (2010) used the time-of-day dependent travel times, whereas Bierlaire and Frejinger (2008) created piecewise linear specifications of free-flow travel time. The corresponding proportions of the total travel time were also included in the utility functions. In addition to the free-flow travel time, Bekhor et al. (2002) and Bekhor et al. (2006) created a variable for time spent on government numbered routes.

Some studies examined the effect of travel time delay on path utilities. For example, Bovy et al. (2008) and Bekhor and Prato (2009) included percent delay with respect to the free-flow travel time and Bekhor et al. (2002) and Bekhor et al. (2006) included delays for three different income categories in the utility functions.

Path utilities are also specified with dummy variables to capture the effect of landmarks (Prato and Bekhor, 2006; Prato and Bekhor, 2007), maximum average speed (Bovy et al., 2008; and Bekhor and Prato, 2009), least-distance and least time path (Bekhor et al. 2002; and Bekhor et al. 2006) on choosing a path. Prato and Bekhor (2006) and Prato and Bekhor (2007) believed that the behavioral variables (habit, time-saving skills, and navigation abilities) are also determinant factors in path utilities. Frejinger and Bierlaire (2007) specified utility functions with some route attributes (effect of number of speed bumps and left turns on uncontrolled signals) as the explanatory variables.

Based on the discussion so far, the PSL, CNL, and EC with Subnetwork models have shown good empirical performances in route choice modeling. The CNL and EC models consist of complex probability structures and require high computation time for model estimations (Bovy et al., 2008). Moreover, estimations with large number of observations generated from a high-resolution network would increase the computation time even more. Prato and Bekhor (2007) suggested using MNL modified models if large number of alternatives is present in the choice set. Among MNL modified models, PSL and PSCL models have produced good estimates. Although, PSL and PSCL models do not differ very much in terms of results (Bovy et al., 2008). The only difference between the two models is availability of a systematic derivation of PSCL model.

To our knowledge, only a few empirical studies have used demographic characteristics as predictor variables (e.g., household income by Bekhor et al., 2002 and Bekhor et al., 2006). Moreover, the use of a large-scale GPS dataset for route choice modeling is also recent and limited (only Bierlaire and Frejinger, 2008, and Schussler and Axhausen, 2010). Except for the study by Bierlaire and Frejinger (2008), predictive assessments of the model on non-synthetic data are also limited.

In our study, we adopt the PSL model with the original PS formulation proposed by Ben-Akiva and Bierlaire (1999). The original PS formulation has shown the best empirical performance (Frejinger and Bierlaire, 2007). The model estimations are performed for a large GPS dataset containing 1913 observations. In addition to the several route attributes, trip- and traveler attributes are also included in the utility functions. The models are estimated using three choice-set sizes (5, 10, and 15 alternatives).

# CHAPTER 3
# MAP MATCHING ALGORITHMS

The focus of this chapter is on Map Matching Algorithms. As already defined, map matching is the process of identifying the specific links of the roadway traversed by a vehicle by mapping the points from its GPS trace to an underlying GIS-based roadway network database. This step is critical as it identifies the fundamental “choice” (i.e., the route) of interest.

Based on a review of the literature, there are two broad classes of algorithms for map matching: The GPS-weighted shortest-path algorithm (GWSP) and the multi path algorithm (MP). The first uses the concept of shortest path in determining the route and is a computationally enhanced version of the approach proposed by Du (2005). The second is new multipath map-matching algorithm, which uses the concept of multiple paths proposed by Marchal et al. (2005) and also implemented by Schüssler and Axhausen (2009).

The GWSP algorithm is more straightforward and computationally less-demanding (especially if a tool for calculating shortest-paths is available) whereas the MP algorithm is more elaborate and demanding (the need to store multiple paths can get cumbersome with dense networks). However, the latter algorithm is also free from assumptions such as preference for shortest paths and generally uses the observed data to determine the route. A comparative analysis of these approaches would therefore be of interest and this study contributes towards that end. Further, we enhance each of the two algorithms computationally and to achieve automation to the greatest extent possible. Both algorithms are implemented in ArcObject within ArcGIS framework, using Python and Visual Basic Application (VBA).

Prior to the discussion of each of our implementations of the two map-matching algorithms, it is useful to outline a generic procedure employed to treat missing GPS points. Factors such as loss of signal while traveling in dense urban areas (canyon effect) can cause GPS points to be missing over parts of a trip. During the map-matching procedure, such missing GPS points can lead to incompleteness in the final predicted routes and/or premature termination of the algorithm (for e.g. see Chung and Shalaby, 2005, Spissu et al., 2011, Marchal et al., 2005). In order to overcome such implementation issues, points are artificially added to the GPS traces at times when the true GPS recordings are missing. Given that the recording frequency of the GPS devices is known, the occurrence of missing points can be detected by simply comparing the time stamps of consecutive points. Whenever missing data are detected, additional points spaced 75 feet apart are added using a simple extrapolation from the previous known/extrapolated point. This treatment of missing GPS points is called “Trip Smoothing” and implemented in Python. Overall the trip-smoothing procedure provides a definitive direction for the algorithm to proceed at locations with missing GPS points thereby reducing the possibility of a breakdown in the algorithm. The algorithms are applied on the processed GPS streams. It is useful to acknowledge that the trip-smoothening based only on spatial extrapolation is a naïve procedure and that more-sophisticated methods have also been proposed. However, we found that our simple procedure was adequate to ensure that the subsequent steps of the map-matching algorithms were implemented fully and effectively.

### 3.1 The GPS-Weighted Shortest-Path (GWSP) Algorithm

The algorithm begins by extracting out a “sub-network” for each trip, which comprises of links in the general vicinity of the GPS points. A “buffer zone” (of size 200 meters in this study) is created around each GPS point and all the links within this area are identified. The set of all links within the buffer zones of at least one of the GPS points in the trip comprises the sub-network. All subsequent processing is done on this sub-network instead of the entire roadway network.

The next step identifies the links in the sub-network with high GPS counts. For this purpose, another “buffer zone” of a smaller size (75 feet or 23 meters) is created around each GPS point. The buffers are consolidated to form a polygon for the entire trip. After this, the links in the sub-roadway network that falls completely within the buffer polygon are selected. Since these represent the link likely to have been traversed by the trips, the impedances are retained to be the true travel times. For the remaining links in the sub-roadway network, the link costs are set to a high value (5000 times the link travel times).

Finally, the shortest-path algorithm built into the ArcGIS software is called to determine the shortest-path for each trip. The use of the sub-roadway network and the assignment of high impedances to links without any GPS points, generally forces the shortest path algorithm to pick a route with links that have more GPS points.

The implementation details of this algorithm are available in Dhakar (2012).

### 3.2 The Multi Path (MP) Algorithm

As in the case of the GWSP algorithm, the MP algorithm also begins by extracting out a “sub-network” for each trip, which comprises of links in the general vicinity of the GPS points. All subsequent processing is done on this sub-network instead of the entire roadway network.

The next step aims to identify the (sequential) set of links within the sub network that could have been potentially traversed by the trip. For this purpose, each GPS point is mapped to the nearest roadway link within the sub-network. In this process, every link in the sub-network could have been mapped to zero, one, or more GPS points. The set of links with at least one GPS point mapped to it constitutes the initial chosen route (ICR). The time stamp on the earliest (first) GPS point mapped to each link in the initial chosen route is determined. The links in the ICR, called the initial chosen links (ICL), are then sorted based on this time stamp so that the sequence of links in the initial chosen route reflects the general temporal trajectory of the trip.

The next step aims to identify the (sequential) set of nodes within the sub network that could have been potentially traversed by the trip. The set of nodes at the end of the ICL identified from the previous step constitutes the “segment nodes” (SN). Each segment node is then mapped to the nearest GPS point and its time stamp is extracted. The segment nodes are then sorted by the time stamp so that the sequence of nodes reflects the general temporal trajectory of the trip. The set of end nodes of the links in the sub-roadway network but not in the ICL, represents the “local nodes”

(LN).The next step is the creation of the link-to-nodes and node-to-links incidence matrices separately for segment- and link- nodes.

The final (iterative) step identifies multiple possible paths between the origin and destination. The algorithm starts at the first segment node, also the one closest to the origin, and sequentially iterates through all the segment nodes (SN) in the list.
At any segment node, the links originating from that node are found using the segment node-to-link matrix. The “eligible” links are then identified (the links that do not have a dead-end or do not re-trace the path; note that not all links are connected in the “forward” direction to other links because of the use of a sub-roadway network). The remaining links are disqualified from further analysis. Now the following process is applied iteratively for all eligible links at the segment node.

For each eligible link, a copy of the current route leading up to this node is created and subsequently the eligible link is added to this. The end node of the newly added link is then obtained from the link-to-nodes matrices. This new end node can either be a segment node or a local node. If it is a segment node, it is labeled as a “right node” (something to be processed later on) and the algorithm proceeds to examine other eligible links from the segment node under consideration. However, if a local node is encountered at the other end of an eligible link, the algorithm processes it similar to the processing of the original segment node. Specifically, all the links at the local node are found by using the local node-to-links matrix. The links are then classified as eligible and disqualified. A new route is created for each eligible link. The algorithm continues to create more routes until all the new routes at the segment node are met with either a segment node or a dead end. With this, all possible paths from the current segment-node have been identified and saved. The algorithm then proceeds to the next segment node. This is one of the “right nodes” identified on a path leading from a previous segment node.

Once all segment nodes in the list of SN have been processed the algorithm has generated multiple possible routes for the trip. Among these, the one that has maximum number of GPS points is considered as the “chosen” route.

The implementation details of this algorithm are available in Dhakar (2012).

## 3.3 Summary

It is fairly evident that the MP algorithm is significantly more computationally intensive than the GWSP approach. Two innovations help enhance the operational performance. First, the multi-paths are constructed over a sub-network that comprises of roadway links in the general vicinity of the GPS points. This prevents the needs to store an excessive number of paths and ensures that, at any point in the algorithm, the paths being considered are fairly close to the overall chosen path. The use of “segment-nodes” is the second enhancement. The earlier multipath algorithms (e.g. Marchal et al. 2005; Schüssler and Axhausen, 2009) iterate through every GPS point in a stream (i.e., the possibility of alternate paths are explored at every GPS point). A high-frequency GPS stream can easily contain thousands of GPS points, thus making the process computationally intensive. The proposed algorithm iterates over segment nodes. Hence, the possibility of alternate paths is explored at only nodes along the roadway links. Such nodes are much fewer compared to the number of GPS points, thereby improving computational efficiency.

# CHAPTER 4
# EXPLORATORY ANALYSIS OF CHOSEN ROUTES

The chapter presents the analysis that validates the two (multipath and shortest path-based) map-matching algorithms presented in the previous chapter. This is accomplished by performing local data collection and comparing the generated routes against "true" routes. Subsequent to the small-scale validation exercise, the two algorithms were applied to data from a larger-scale GPS-based travel survey. The relative performances of the two algorithms are compared. Further, the routes determined from each of these algorithms are compared against the shortest-distance, and shortest-time paths for the same trip-end locations.

## 4.1 Validation

Prior to a large-scale application of the map-matching algorithms, it is important to validate these against true routes. Data collected from GPS-based travel surveys are inadequate for this purpose as the "true" routes are not directly elicited in these surveys (only the GPS traces are passively obtained). Therefore, we performed our own in-vehicle data collection in Orlando, Florida, USA, using the Geostats' Geologger (the vehicle-based device used in the Chicago and several other GPS-based travel surveys nationwide). The reader is referred to the Chapter on Data for a brief description of the device and the data collection procedure.

The GPS data were collected for about 33 trips (37,214 GPS points) with each trip being at-least 5 minutes in duration and 2 miles long. As the vehicle was driven on known routes, the routes from the map matching process were first verified manually and algorithm efficiency was subsequently evaluated in terms of the proportion of the true nodes, distance, and time replicated by the algorithms.

For 26 out of the 33 trips, both methods generated routes (In the rest, the shortest-path based algorithm did not generate a route – these are discussed later). Table 4-1 presents a summary of overlap measures for these trips and for each algorithm. The results indicate that both algorithms are able to replicate the true routes to a very large extent (overall, over 98% of the distance and time are replicated and over 95% of the nodes are replicated). The freeway only trips (>98% on freeway) are replicated almost 100% by both algorithms. However, for trips with arterials and local streets over 96% of the distance and time are replicated.

The primary reason for the inaccuracy with the MP algorithm was because of routes containing loops; the algorithm struggles to include a loop in the route. Additionally, the algorithm occasionally selects links that are closer to true links because they have higher GPS points. Also, both algorithms find a way if some links are missing in the network, thus selecting other links in the route.

As already mentioned, there were 6 cases in which the shortest path did not generate a route (The multi-path algorithm generates the routes for all trips). Further, one trip had a significantly different route generated by the shortest path algorithm relative to the true route. A summary of these seven trips is presented in Table 4-2.

Table 4-1. Validation of the map-matching algorithms

| | True route | | | | | | | Overlap from MP (%) | | | Overlap from GWSP (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nodes | Distance (Miles) | Time (Minutes) | Avg. Speed | % Distance on expressway | % Distance on arterial | % Distance on local road | Nodes | Distance | Time | Nodes | Distance | Time |
| (a) Freeway Trips | 74 | 29.7 | 34.38 | 74.47 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 40 | 20.61 | 22.49 | 72.72 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 68 | 26.21 | 30.78 | 69.82 | 98.51 | 1.49 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 63 | 25.60 | 28.76 | 75.87 | 98.24 | 1.76 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 37 | 24.63 | 26.87 | 78.25 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 25 | 18.90 | 20.62 | 74.75 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 60 | 15.62 | 17.04 | 73.84 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 36 | 16.27 | 19.74 | 73.61 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 26 | 13.57 | 14.8 | 64.8 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 121 | 15.80 | 17.23 | 71.05 | 100 | 0 | 0 | 97.52 | 98.6 | 98.6 | 100 | 100 | 100 |
| | 40 | 17.18 | 18.74 | 68.69 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 46 | 20.66 | 24.02 | 72.19 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 48 | 17.28 | 18.85 | 73.18 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 28 | 16.47 | 17.97 | 72.62 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | | | | | | | **Total** | **99.58** | **99.92** | **99.92** | **100** | **100** | **100** |
| (b) Arterial and Local Streets Trips | 77 | 27 | 31.46 | 66.36 | 91.55 | 8.19 | 0.26 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 104 | 5.12 | 6.81 | 37.06 | 14.98 | 78.91 | 6.11 | 100 | 100 | 100 | 77.66 | 79.51 | 82.98 |
| | 90 | 4.4 | 8.7 | 25.78 | 19.75 | 43.22 | 37.03 | 67.78 | 66.63 | 57.86 | 100 | 100 | 100 |
| | 145 | 9.65 | 13.69 | 36.04 | 70.79 | 7.25 | 21.96 | 96.97 | 96.58 | 97.55 | 100 | 100 | 100 |
| | 80 | 5.05 | 9.59 | 14.12 | 14.02 | 15.14 | 70.84 | 85.53 | 95.21 | 95.38 | 100 | 100 | 100 |
| | 166 | 10.48 | 17.23 | 32.36 | 35.22 | 32.90 | 31.88 | 89.16 | 90.04 | 89.4 | 100 | 100 | 100 |
| | 124 | 7.12 | 11.22 | 26.85 | 29.49 | 48.98 | 21.53 | 97.58 | 99.26 | 99.17 | 100 | 100 | 100 |
| | 76 | 11.52 | 15.45 | 52.42 | 66.77 | 32.27 | 0.96 | 98.68 | 99.24 | 98.3 | 100 | 100 | 100 |
| | 98 | 16.65 | 21.28 | 55.07 | 76.22 | 23.17 | 0.60 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 34 | 17 | 26.64 | 67.13 | 95.64 | 0.00 | 4.36 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 176 | 26.02 | 30.91 | 66.58 | 89.15 | 10.84 | 0.00 | 92.61 | 97.3 | 95.2 | 100 | 100 | 100 |
| | 78 | 6.90 | 9.52 | 44.26 | 0 | 72.46 | 27.54 | 89.47 | 90.13 | 82.46 | 89.47 | 90.13 | 82.46 |
| | | | | | | | **Total** | **92.93** | **96.86** | **95.17** | **97.48** | **98.82** | **98.60** |

Table 4-2. Troublesome trips for the GWSP method

| True route | | | | | | | Overlap from MP (%) | | | Overlap from GWSP (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nodes | Distance (miles) | Time (minutes) | Avg. speed (mph) | % Distance on expressway | % Distance on arterial | % Distance on local Roads | Nodes | Distance | Time | Nodes | Distance | Time |
| 55 | 13.82 | 15.09 | 81.2 | 99.94 | 0.00 | 0.06 | 100 | 100 | 100 | Stop on the wrong side | | |
| 99 | 12.56 | 18.61 | 80.4 | 75.53 | 21.72 | 2.75 | 100 | 100 | 100 | Missing links in network | | |
| 256 | 16.09 | 22.19 | 68.9 | 29.52 | 63.27 | 7.21 | 91.34 | 95.02 | 94.43 | Missing links in network | | |
| 157 | 21.32 | 24.22 | 84.1 | 84.01 | 13.34 | 2.64 | 97.47 | 99.25 | 99.11 | Missing links in network | | |
| 60 | 6.53 | 9.46 | 65.3 | 0.00 | 73.64 | 26.36 | 100 | 100 | 100 | Missing links in network | | |
| 52 | 16.26 | 19.12 | 83.8 | 91.07 | 8.79 | 0.14 | 100 | 100 | 100 | Missing links in network | | |
| 53 | 2.45 | 4.29 | 38.8 | 0.00 | 0.00 | 100.00 | 90.2 | 93.71 | 91.36 | 9.8 | 6.29 | 8.64 |

One of the major reasons for the failure of the GWSP algorithm is the absence of certain links in the roadway. In one of such cases, as shown in Figure 4-1, the GWSP algorithm cannot find a route but the MP algorithm selects off-path links to complete a route. This suggests that the MP algorithm is less sensitive than the GWSP algorithm to the errors in the underlying roadway network in terms of generating a complete, feasible route for the trips. To overcome this issue, it would be necessary to pre-process the network GIS data and add any missing links.

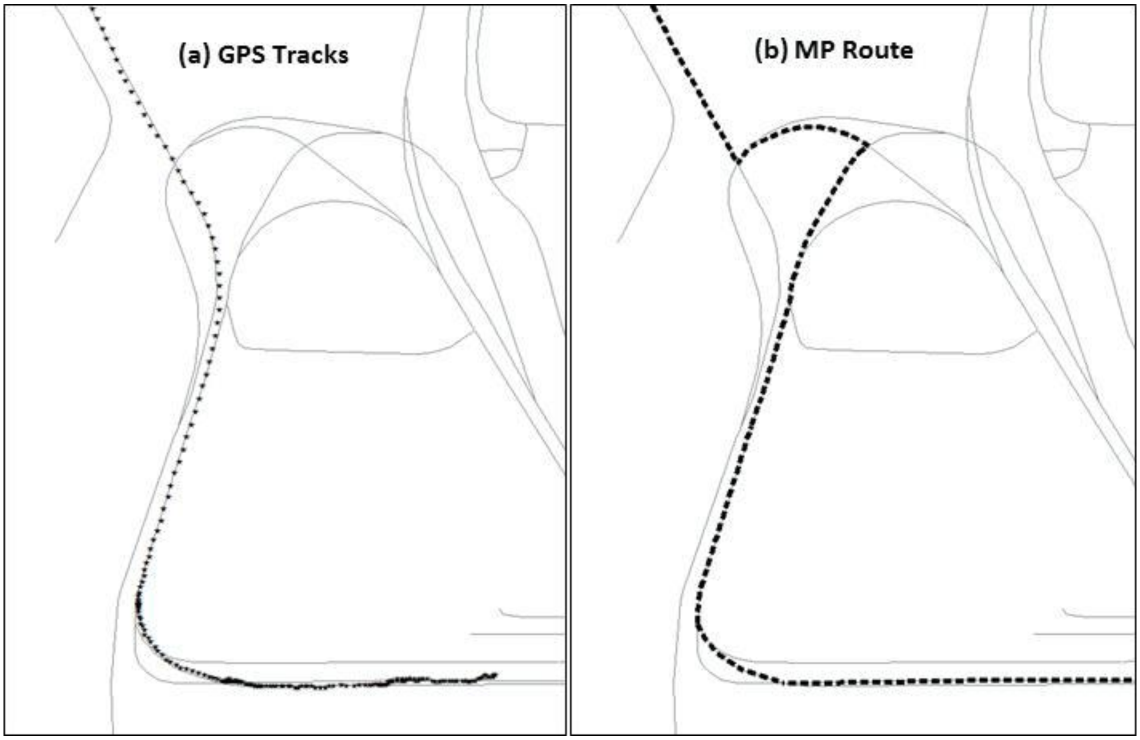


Figure 4-1. Missing links in the network – example 1 (a) GPS tracks, and (b) MP route

A second example with missing links in the roadway is presented in Figure 4-2. In this case, the GWSP method does find a route; however, this is much farther away from the true route relative to the route determined by the MP method.

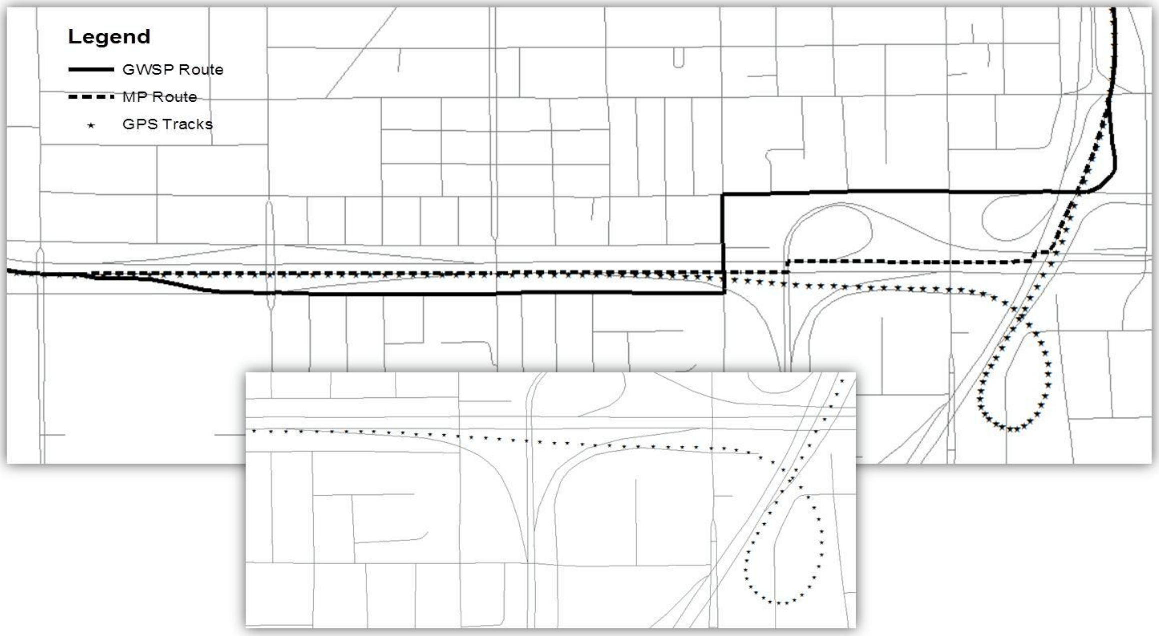


Figure 4-2. Missing links in the network – example 2

A second reason leading to the failure of the GWSP method occurs when trip origin/destination is close to a divided roadway represented by two separate links in the GIS network. In this case, if the GPS points gets "snapped" to the link with the wrong directionality (i.e., the vehicle is moving one way but the points are "closer" to the opposite link), the algorithm fails. To overcome this issue, the algorithm has to be expanded to ensure that the GPS points are snapped to not just the nearest link but also to the one in which the direction of flow is consistent with the bearing of the GPS points. This extension is currently not accommodated in our algorithm.

Finally, it is also interesting to note that the GWSP algorithm for trip 9 is very different from the true path which is also reasonably well detected by the MP algorithm. The schematic of this trip represented in Figure 4-3 quickly highlights why this is the case. For this trip, the two trip-ends are fairly close to each other spatially but the actual travel path included a much longer distance and a "U" turn. Even though links that do not have a GPS point are assigned a very high impedance to prevent the algorithm from choosing these links, it is possible (as in this example) that the total impedance of the set of links with GPS points is still larger than the large artificial impedance assigned to the link without GPS points. This problem could be potentially overcome by assigning even higher impedances to the links without GPS points. The MP algorithm is less affected by the relative locations of origin and destination and, hence, correctly finds the links traversed during the trip.

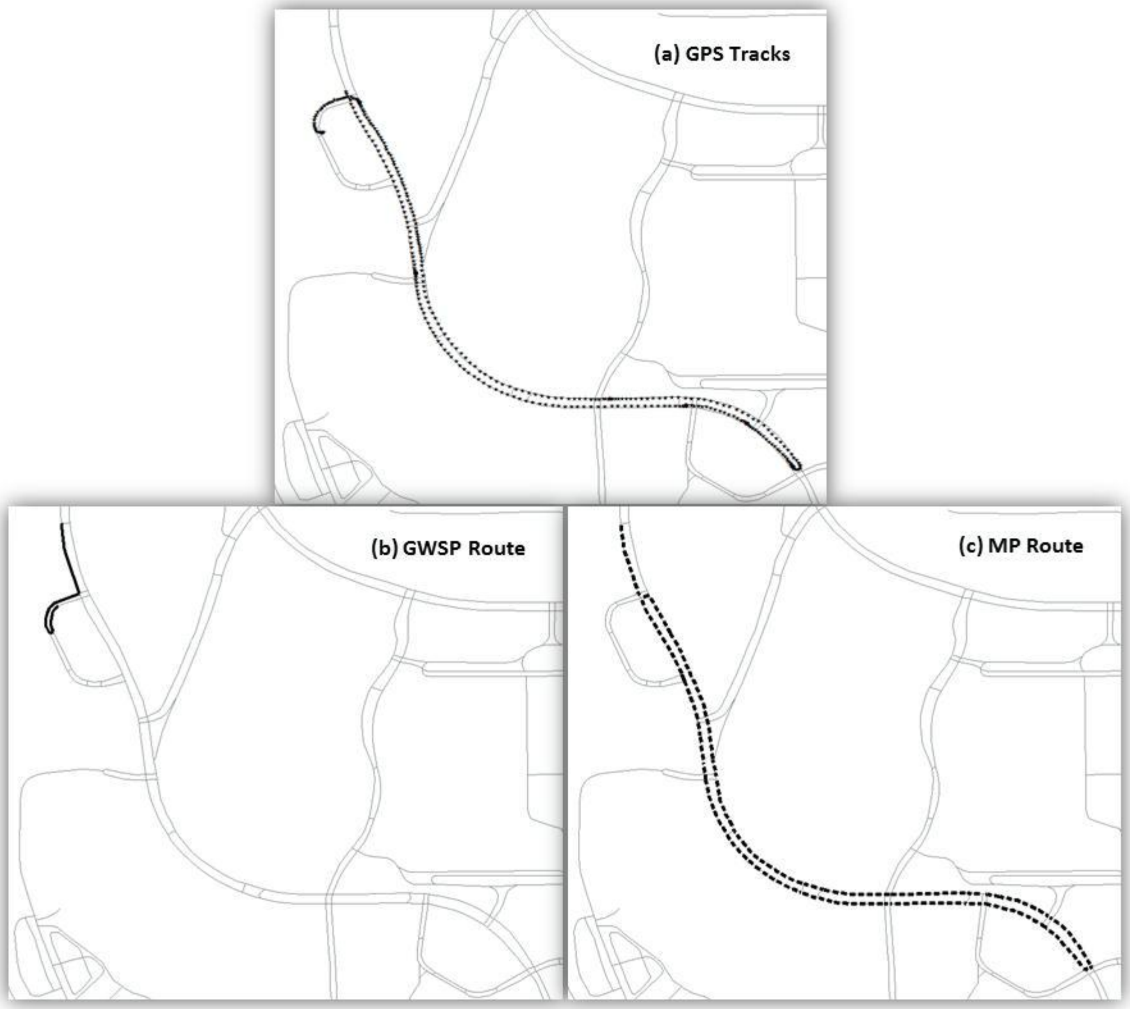


Figure 4-3. A round trip (a) GPS tracks, (b) GWSP algorithm, and (c) MP algorithm

Overall, this small-scale validation exercise demonstrates the ability of the proposed algorithms to generate reasonable routes that are fairly close to the true routes. Issues such as incomplete roadway networks were identified to cause the GWSP algorithm to fail in certain cases. It is also useful to note here that this study relies on the algorithms built-into ArcGIS for shortest-path calculations. Developing customized algorithms that are sensitive to missing links can help (in addition to the other methods discussed above) in ensuring that a complete route is always generated.

## 4.2 Application to a Large Scale GPS-based Travel Survey

### 4.2.1 Data

The GPS data comes from the in-vehicle GPS-survey component of the Chicago Regional Household Travel Inventory (CRHTI). A high-resolution, GIS-compatible roadway network for the study area was obtained from ArcGIS Data and Maps from ESRI. This GIS layer has information on speed, functional classification, and distance of most links in the roadway (including local streets). Additionally, a GIS compatible sub-zone layer for the area was also obtained from the Chicago Department of Transportation (DOT).

Chicago Regional Household Travel Inventory (CRHTI), NuStats (2008) is a recently conducted, between January 2007 and February 2008, comprehensive study of the demographics and travel behavior characteristics of a large number of households (10,552 households) in the greater Chicago area. Travel information for either 1-day or 2-day of travel by household members was collected in conjunction with the socio-demographics of the households and its members. Additionally, the study also had a GPS data collection component, collected in four stages for both in-vehicle and on-person, for the trips made by the selected households.

In stage 1, in-vehicle GPS data was collected for a day's duration at households that travel a lot (travel throughout the region as part of their job or who traveled into Chicago for personal or business reasons at least three times per week). Stage 2 had in-vehicle GPS data for households with at least one member making more than 10 trips per day by auto or who traveled more than 75 miles per day as part of their job. In stage 3 in-vehicle GPS data was collected for heavy travelers for the period of 7 days. In stage 4, wearable GPS data was also collected for the period of 7 days. (NuStats and GeoStats, 2008)

For the collection of in-vehicle travel data, a simple GPS data-logging device GeoLogger, Figure 4-4, was used.


Figure 4-4. In-vehicle GeoLogger (source: GeoStats)

The GeoLogger recorded date, time, latitude, longitude, speed, heading, altitude, number of satellites, and HDOP for a vehicle at 1-seocnd interval. The Original data comprises of 6089852 GPS points from 9941 trips made by 408 vehicles (259 households).

The original data comprises of 6,089,852 GPS points from 9941 trips made by 408 vehicles (from 259 households). From these, trips shorter than 5 minutes in duration or 2 miles in distance were removed. The origin and destination of the remaining 5290 trips were mapped to the sub-zones and the trips with unique subzones at both ends were identified. The resulting 4406 trips have unique origin-destination pairs.

The two map matching algorithms were applied to all these 4406 trips. The MP and GWSP methods generated complete routes for 4093 and 3889 trips respectively. It is useful to note that these were generated without any manual interventions such as visual inspections. There were 3886 trips for which both map-matching algorithms generated a route. For these cases, the shortest-distance (SD) and shortest (free-flow) travel-time (ST) routes were also generated. After further cleaning to remove outliers and inconsistent cases, the final sample consists of 3513 trips each of which has four routes; two from the map-matching algorithms (MP and GWSP) and two from normative route-choice algorithms (SD and ST).

It is useful to recognize that the GPS traces do not necessarily start/end at nodes on the roadway network whereas the “routes” generated from various algorithms do. To ensure consistency in the start- and end- locations of all four routes generated for any trip, the origin- and destination- nodes determined from the MP algorithm were used as the terminal nodes in the determination of the other three routes (all of which use the ArcGIS-based shortest-path methods).

For the final sample, based on the trip start times (available from the GPS data), six discrete time periods are defined for the time of day: early morning (midnight – 6:30 AM), AM peak (6:30 AM – 9:00 AM), AM off peak (9:00 AM – noon), PM off peak (noon – 4:00 PM), PM peak (4:00 PM – 6:30 PM), and evening (6:30 PM – midnight). The frequency distribution of trips for the time of day, Figure 4-5, shows that the travelers preferred PM off peak period, noon – 4 pm, for their travel as 30% of the trips, the most, were undertaken during the time period. Also, only 3% of the trips, the least, started in the early morning suggesting this period as the least preferred. Further, the examination of the trip lengths, Figure 4-6, indicates that the dataset consists of trips with considerable lengths. More than 50% of the trips are longer than 5 miles in length.

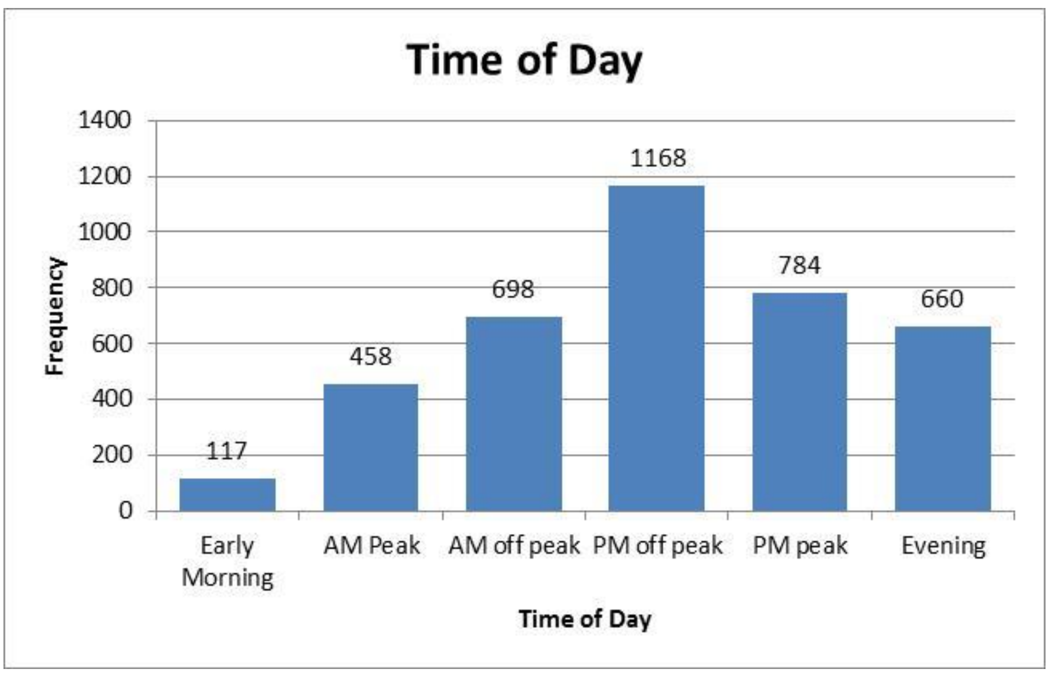

Figure 4-5. Time of day frequency distribution

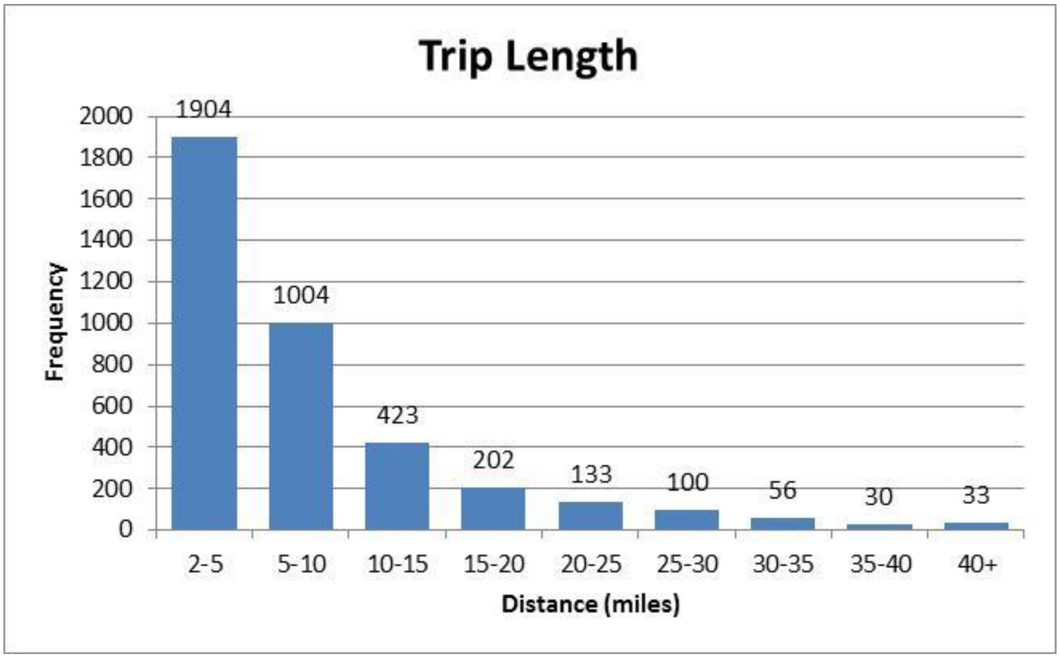


Figure 4-6. Trip length frequency distribution

### 4.2.2 Aggregate Comparisons

The useable data from Chicago comprise a total of 3513 trips each representing a unique origin-destination pair. For each OD pair, four routes are constructed: two chosen routes from the two map-matching algorithms (MP and GWSP), a shortest-distance (SD) route, and a shortest-time (free-flow) (ST) route.

Table 4-3 shows aggregate statistics of the routes produced by the four methods. Both the MP and GWSP algorithms produced similar routes as the summary measures (averages and deviances) for link count, distance and time are quite comparable.

The median distance of the “chosen” (i.e. determined from MP or DWSP algorithms) path is higher than median distance of the shortest-distance (and shortest-time) path between the same sets of origin-destination pairs. Similarly, median (free-flow) travel time of the “chosen” (i.e. determined from MP or DWSP algorithms) path is higher than the median (free-flow) travel time of the shortest-distance (and shortest-time) path between the same sets of origin-destination pairs.

Table 4-1. Aggregate statistics of routes from the four methods

| | MP | GWSP | SD | ST |
|---|---|---|---|---|
| **Links (Count)** | | | | |
| Mean | 112.10 | 111.02 | 108.57 | 110.58 |
| Median | 76 | 76 | 72 | 75 |
| Std. Deviation | 97.63 | 96.83 | 100.55 | 98.32 |
| Min | 8 | 8 | 5 | 5 |
| Max | 662 | 678 | 711 | 668 |
| Q1 | 48 | 47 | 44 | 46 |

| | | | | |
|---|---|---|---|---|
| Q3 | 138 | 135 | 133 | 137 |
| **Distance (Miles)** | | | | |
| Mean | 8.27 | 8.23 | 7.36 | 7.76 |
| Median | 5.23 | 5.21 | 4.72 | 4.89 |
| Std. Deviation | 7.90 | 7.90 | 6.81 | 7.31 |
| Min | 2.01 | 2.01 | 0.67 | 0.67 |
| Max | 57.12 | 57.12 | 46.59 | 52.72 |
| Q1 | 3.15 | 3.12 | 2.89 | 3.01 |
| Q3 | 10.08 | 10.01 | 9.14 | 9.68 |
| **Time (Minutes)** | | | | |
| Mean | 14.03 | 13.86 | 13.97 | 12.10 |
| Median | 9.64 | 9.48 | 9.16 | 8.28 |
| Std. Deviation | 11.64 | 11.56 | 12.39 | 10.02 |
| Min | 2.78 | 2.78 | 1.60 | 1.60 |
| Max | 80.84 | 83.30 | 87.40 | 64.59 |
| Q1 | 6.07 | 5.97 | 5.72 | 5.28 |
| Q3 | 17.68 | 17.46 | 17.46 | 15.28 |

### 4.2.3 Measures of Similarity of Pairs of Routes

The above analysis presented an overall (aggregate) summary across all the trips in the sample. Next, we examine the extent to which the routes generated between any OD pair are similar. The overlap (i.e., the set of common links) between any pair of routes is determined in ArcGIS by using the intersect tool. Once the common set of links have been identified, it is possible to calculate the number of common links, the total distance across the common links (overlap distance), and the total (free-flow) travel time across the common links (overlap time).

From the three values identified above, four route-level measures of similarity can be constructed. These are (1) overlapping index (OI), (2) commonly ratio (CR), (3) distance deviation index (DDI), and (4) time-deviation index (TDI).

The overlapping index (OI) is the ratio of number of links common to the two routes to the total number of unique links in both routes (Spissu et al., 2011). An OI value of 0 indicates that the two routes are completely disjoint (don’t share any common links) and a value of 1 indicates that both routes overlap perfectly (all links are the same).

Commonly ratio (CR) of two routes is calculated as: $\mathrm{CR} = \frac{L_{ij}}{\sqrt{L_i}\sqrt{L_j}}$, where $L_{ij}$ is the distance of common links in route $i$ and $j$; $L_i$ is the distance of the route $i$ and $L_j$ is the length of the route $j$ (Pilat et al., 2011). As in the case of OI, the CR measure also takes values between 0 and 1 with 0 representing no overlap and 1 representing perfect overlap.

The previous two metrics can be applied to any pair of routes. The next two metrics, DDI and TDI (Spissu et al., 2011) are used to compare an algorithm-generated route to the SD and ST routes respectively.

The Distance Deviation Index (DDI) determines the extent to which the chosen route is longer (in distance) than the shortest-distance path between the same OD pair and is calculated as: $\mathrm{DDI} = \frac{\mathrm{d_{CR}} - \mathrm{d_{SD}}}{\mathrm{d_{SD}}}$, where $\mathrm{d_{CR}}$ is the distance of the chosen route and $\mathrm{d_{SD}}$ is the distance of the SD route.

Similarly, Time Deviation Index (TDI) determines the extent to which the chosen route is longer (in time) than the shortest-time path between the same OD pair and is calculated as: $\mathrm{TDI} = \frac{\mathrm{t_{CR}} - \mathrm{t_{ST}}}{\mathrm{t_{ST}}}$, where $\mathrm{t_{CR}}$ is the time of the chosen route and $\mathrm{t_{ST}}$ is the time of the ST route.

Using the similarity measures as described, the routes are compared in a pair-wise manner. First, the routes generated by the two map-matching algorithms are compared to determine the extent to which they are similar. Next, the routes from each of the algorithms are compared to the SD path. Finally, the routes from each of the algorithms are compared to the ST path.

#### 4.2.4 Extent of Similarity of the Routes Generated by the Two MM Algorithms

The routes generated by the two map-matching algorithms (MP and GWSP) are compared using the overlap index and commonly ratio measures.

The frequency distributions of OI and CR are presented in Figure 4-7. Almost 88% of the routes have a commonly ratio of higher than 0.90 and 70% of the routes have an OI of 0.9 or greater. In fact, 49.44 % of the trips have a perfect overlap (OI = 1 and CR = 1). About 95% of the routes have a CR value of greater than 0.8 and 85% of the routes have an OI value of greater than 0.8. Thus, in most (~ 85%) of cases, both algorithms generate routes that overlap significantly (80% or more) in terms of links and distance. Further, the overlaps are greater in terms of distance than in terms of links. This indicates that even if slightly different links are determined by two map-matching algorithms, these are likely to be short-distance links.

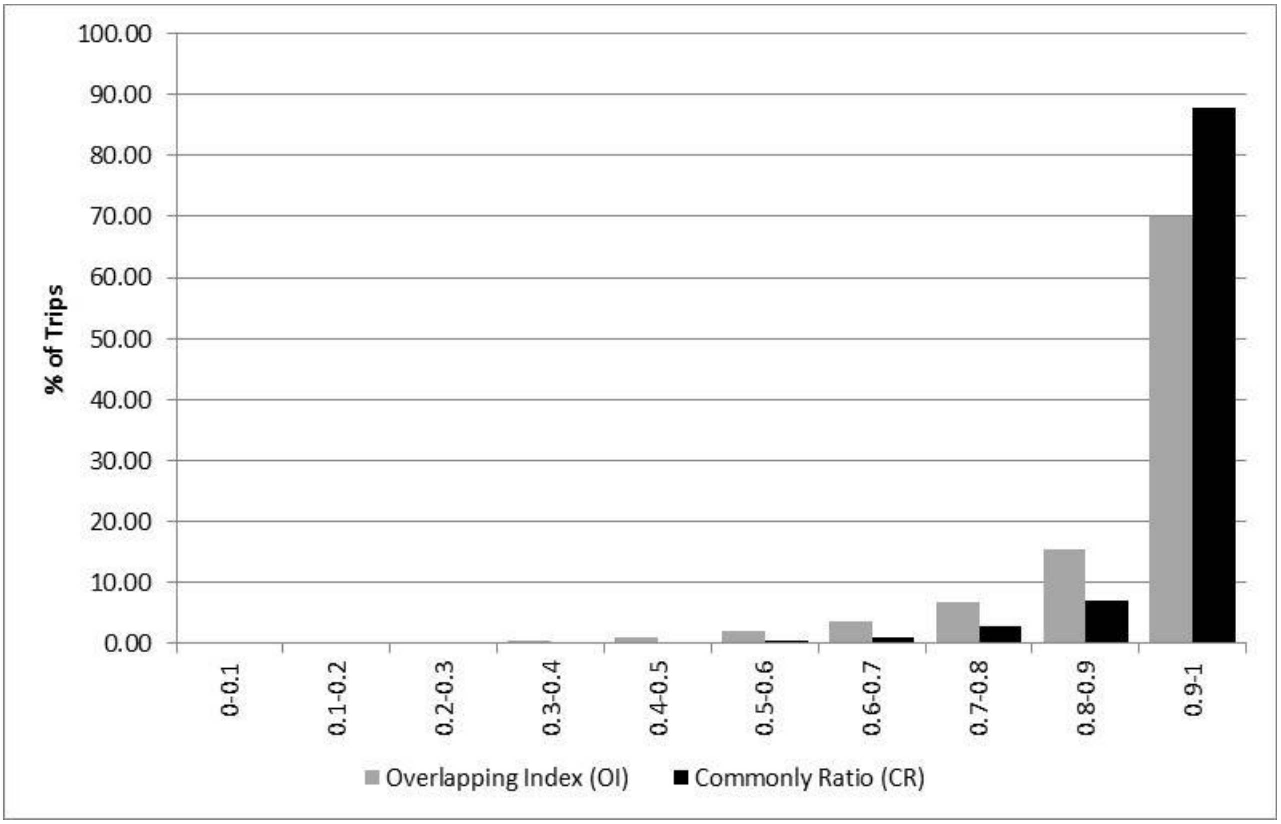


Figure 4-7. Comparison of routes from the two map-matching algorithms

The routes are further examined by classifying the trips based on the trip length and time-of-day. First, the trip length is categorized into three groups: short (2-5 miles), medium (5-10 miles), and long (more than 10 miles). The time of day was reclassified into two periods: peak period (AM peak and PM peak) and off-peak period (early morning, AM off-peak, PM off-peak, and evening). After this, six discrete categories are created: peak short trips, peak medium trips, peak long trips, off-peak short trips, off-peak medium trips, and off-peak long trips.

From the Table 4-4, the OI and CR values marginally decrease with increasing length of the trips (for both peak and off-peak trips). This seems reasonable as, for longer trips, the algorithms “see” a much larger sub-network thereby increasing the probability of generating different paths. However, the median OI and CR values are significantly large even for the longest trips. For trips of any length, there is not an appreciable difference between peak- and off-peak trips.

Table 4-2. Overlapping index (OI) for routes from MP and GWSP methods

| | | Peak | | | Off-peak | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Short | Medium | Long | |
| | Sample size | 499 | 301 | 330 | 1219 | 623 | 541 | 3513 |
| ing Index | Mean | 0.93 | 0.91 | 0.89 | 0.92 | 0.91 | 0.89 | 0.91 |
| | Median | 1.00 | 0.96 | 0.93 | 1.00 | 0.96 | 0.93 | 0.97 |
| | Std. Dev. | 0.14 | 0.13 | 0.12 | 0.14 | 0.13 | 0.12 | 0.13 |
| | Min | 0.18 | 0.24 | 0.36 | 0.14 | 0.06 | 0.32 | 0.06 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.90 | 0.87 | 0.85 | 0.90 | 0.88 | 0.85 | 0.88 |
| | Q3 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.97 | 1.00 |
| Commonly Ratio (CR) | Mean | 0.97 | 0.96 | 0.95 | 0.96 | 0.96 | 0.95 | 0.96 |
| | Median | 1.00 | 0.99 | 0.98 | 1.00 | 0.99 | 0.98 | 0.99 |
| | Std. Dev. | 0.08 | 0.08 | 0.07 | 0.08 | 0.08 | 0.07 | 0.08 |
| | Min | 0.26 | 0.48 | 0.52 | 0.34 | 0.17 | 0.40 | 0.17 |
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.97 | 0.96 | 0.94 | 0.96 | 0.96 | 0.94 | 0.96 |
| | Q3 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

### 4.2.5 Comparing the Chosen Routes Against the Shortest-Distance Routes

The comparison of the chosen routes against the shortest-distance routes was performed using the commonly ratio (CR) and the distance deviation index (DDI).

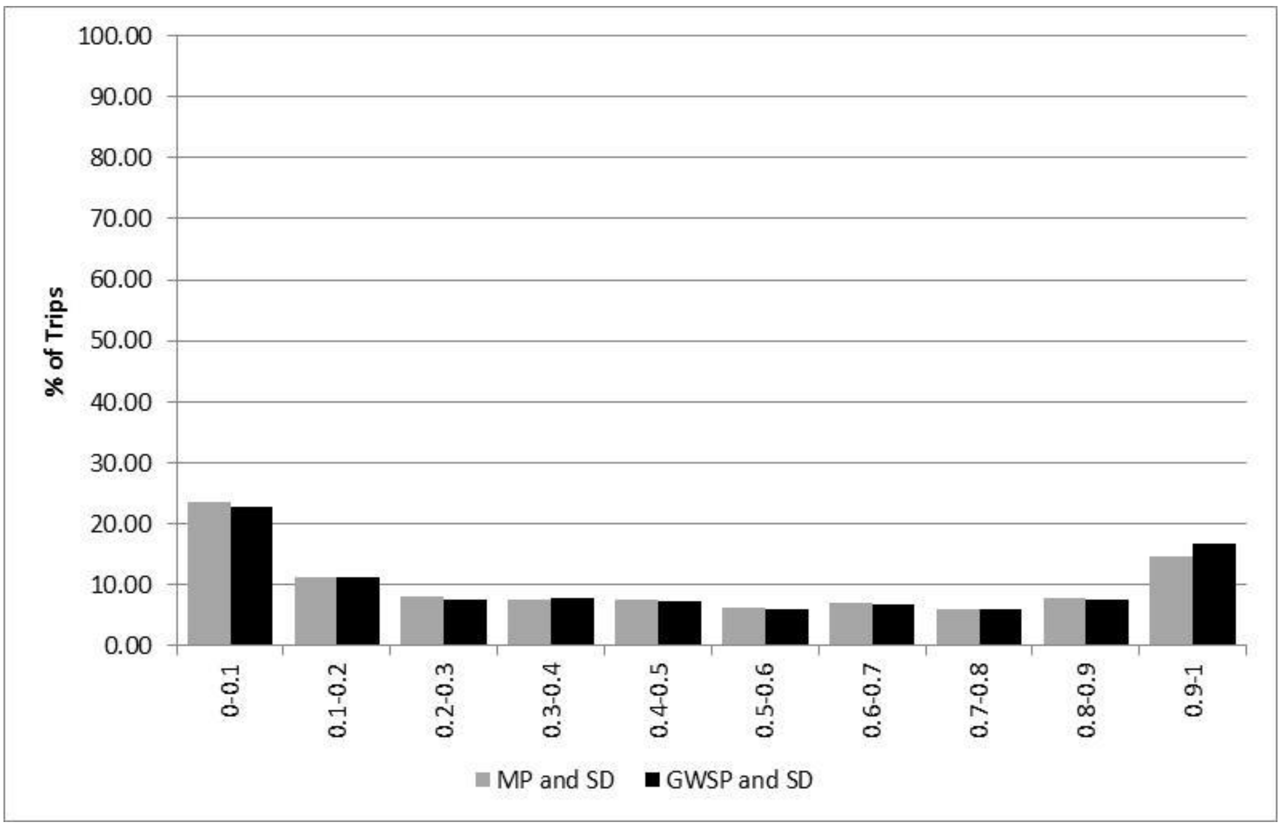


Figure 4-8. Commonly ratio (CR) of routes from the MM and SD methods

The Figure 4-8 presents the CR values calculated between the chosen routes and the shortest-distance paths. It is evident that the chosen routes are considerably different from the shortest-distance routes. On an average (median), the CR value is about 0.40 with a significantly large deviation across the trips. Less than 15% of the trips have a CR value of over 0.9 and about 22% of the trips have a value less than

0.1.The trends are quite similar irrespective of the algorithm used to generate the chosen route. This is not surprising given that we have already established that the two map-matching algorithms generate fairly similar routes.

Commonly ratios of the chosen routes and the SD routes are presented in Table 4-5. The CR of the chosen routes decreases sharply with the trip length (0.56 for short, 0.4 for medium, and 0.16 for long) for both peak and off-peak conditions. This suggests that chosen path is likely to be more different from the SD path for longer trips.

Table 4-3. CR of the chosen routes and the SD routes

| | | Peak | | | Off-peak | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Short | Medium | Long | |
| | Sample size | 499 | 301 | 330 | 1219 | 623 | 541 | 3513 |
| CR of MP and SD Routes | Mean | 0.54 | 0.43 | 0.28 | 0.52 | 0.44 | 0.28 | 0.44 |
| | Median | 0.56 | 0.38 | 0.16 | 0.53 | 0.37 | 0.15 | 0.39 |
| | Std. Dev. | 0.34 | 0.34 | 0.29 | 0.34 | 0.34 | 0.29 | 0.34 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.23 | 0.12 | 0.04 | 0.19 | 0.11 | 0.04 | 0.11 |
| | Q3 | 0.85 | 0.75 | 0.46 | 0.85 | 0.75 | 0.45 | 0.76 |
| CR of GWSP and SD Routes | Mean | 0.56 | 0.45 | 0.29 | 0.54 | 0.45 | 0.29 | 0.46 |
| | Median | 0.58 | 0.40 | 0.16 | 0.55 | 0.39 | 0.16 | 0.41 |
| | Std. Dev. | 0.35 | 0.34 | 0.30 | 0.35 | 0.35 | 0.29 | 0.35 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.23 | 0.13 | 0.04 | 0.20 | 0.12 | 0.04 | 0.12 |
| | Q3 | 0.89 | 0.76 | 0.48 | 0.87 | 0.78 | 0.46 | 0.79 |

From the Table 4-6, on average, DDI increases with the trip length, indicating that the chosen route is generally longer than the shortest-distance route and the disparity is greater for longer trips. On an average (median), a distance along the chosen route for a short distance trips is about 4% longer than the shortest-distance path for the same route. The corresponding values are 6% and 10% for medium- and long- distance trips respectively.

Table 4-4. DDI of the chosen routes with the SD routes

| | | Peak | | | Off-peak | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Short | Medium | Long | |
| | Sample size | 499 | 301 | 330 | 1219 | 623 | 541 | 3513 |
| DDI of MP routes with SD Routes | Mean | 0.11 | 0.12 | 0.15 | 0.12 | 0.12 | 0.15 | 0.13 |
| | Median | 0.04 | 0.06 | 0.10 | 0.05 | 0.07 | 0.11 | 0.07 |
| | Std. Dev. | 0.21 | 0.27 | 0.18 | 0.24 | 0.23 | 0.13 | 0.22 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 2.48 | 3.89 | 2.21 | 2.97 | 3.62 | 0.82 | 3.89 |
| | Q1 | 0.01 | 0.02 | 0.06 | 0.01 | 0.02 | 0.06 | 0.02 |
| | Q3 | 0.12 | 0.15 | 0.18 | 0.14 | 0.16 | 0.19 | 0.15 |
| DDI of GWSP routes with SD Routes | Mean | 0.10 | 0.12 | 0.14 | 0.11 | 0.11 | 0.14 | 0.12 |
| | Median | 0.04 | 0.05 | 0.10 | 0.04 | 0.06 | 0.11 | 0.06 |
| | Std. Dev. | 0.21 | 0.25 | 0.18 | 0.23 | 0.21 | 0.13 | 0.21 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 2.48 | 3.51 | 2.31 | 2.97 | 3.10 | 0.82 | 3.51 |
| | Q1 | 0.00 | 0.01 | 0.05 | 0.00 | 0.02 | 0.06 | 0.01 |
| | Q3 | 0.10 | 0.13 | 0.18 | 0.12 | 0.15 | 0.19 | 0.14 |

Looking at the results for the CR and the DDI measures simultaneously is quite illuminating. It is clear that even though the chosen routes share only about 40% (overall median) of the distance in common with the shortest-distance path, the distance along the chosen route is only about 6% (overall median) longer than the distance along the shortest-distance path. This is possibly because of the presence of alternate (possibly parallel) links/paths in the network that are very comparable in terms of distances.

### 4.2.6 Comparing the Chosen Routes Against the Shortest-Time Routes

Prior to comparing the chosen routes against the shortest-time routes, it is useful to demonstrate that the SD and the ST paths are generally not the same in many cases. A comparison of the SD and ST routes, Figure 4-9, showed that only about 24% routes depicted OI of 0.9 or higher. Moreover, about 21% routes showed OI of less than 0.1. This confirms that for an OD pair the SD path is not necessarily the ST path. Also, the comparison of the chosen routes with the SD path, previous section, showed that as the trip length increases, the traveler tends to avoid the SD path. Therefore, to further understand the chosen routes, they are compared against the ST routes.

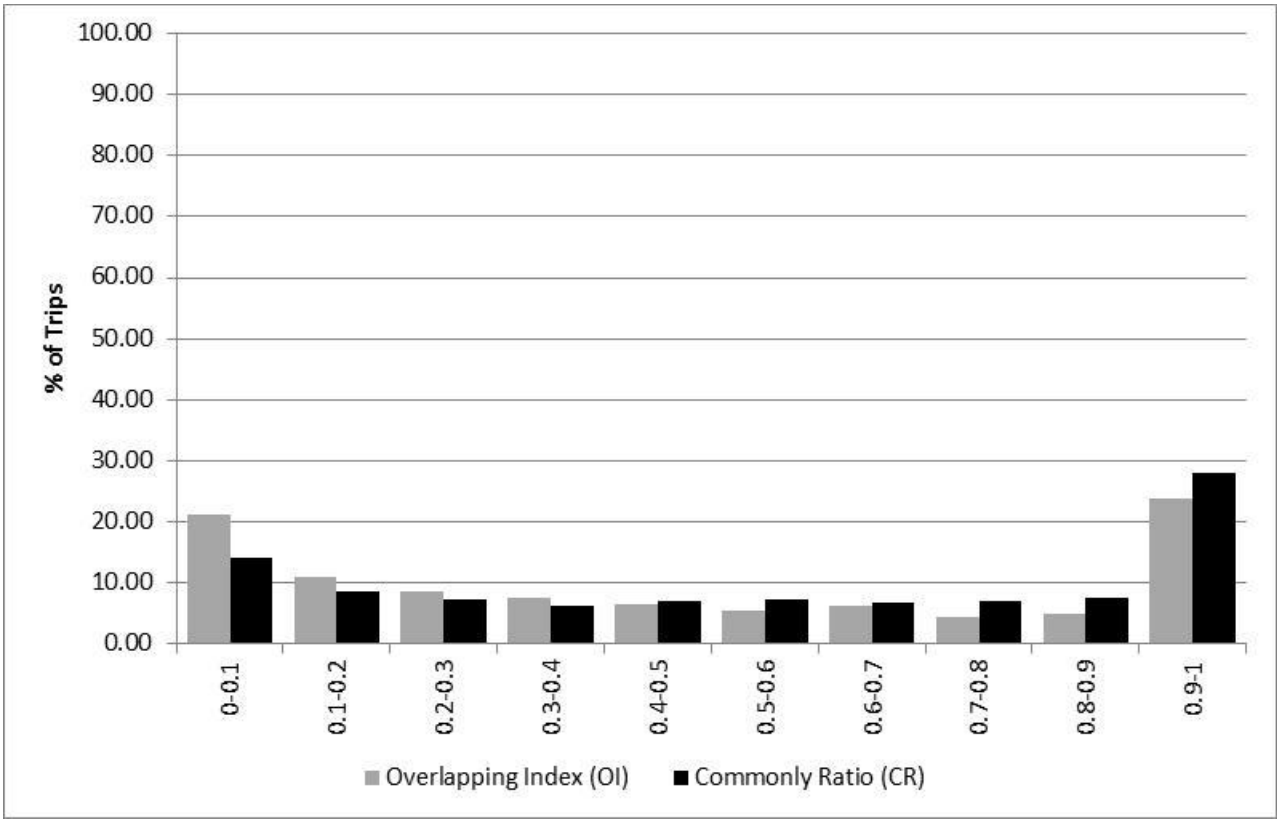


Figure 4-9. Comparison of routes from SD and ST methods

The comparison of the chosen routes against the shortest-time routes was performed using the commonly ratio (CR) and the time deviation index (TDI).

Figure 4-10 and Table 4-7 present the CR values calculated between the chosen routes and the shortest-time paths. Similar to the shortest-distance routes, the chosen routes are considerably different from the shortest-time routes too. On an average (median), the CR value is about 0.55 with a significantly large deviation across the trips. Less than 23% of the trips have a CR value of over 0.9 and about 17% of the trips have a value less than 0.1.The trends are quite similar irrespective of the algorithm used to generate the chosen route.

The CR of the chosen routes decreases with the trip length (0.58 for short, 0.51 for medium, and 0.42 for long) for both peak and off-peak conditions. This suggests that chosen path is likely to be more different from the ST path for longer trips.

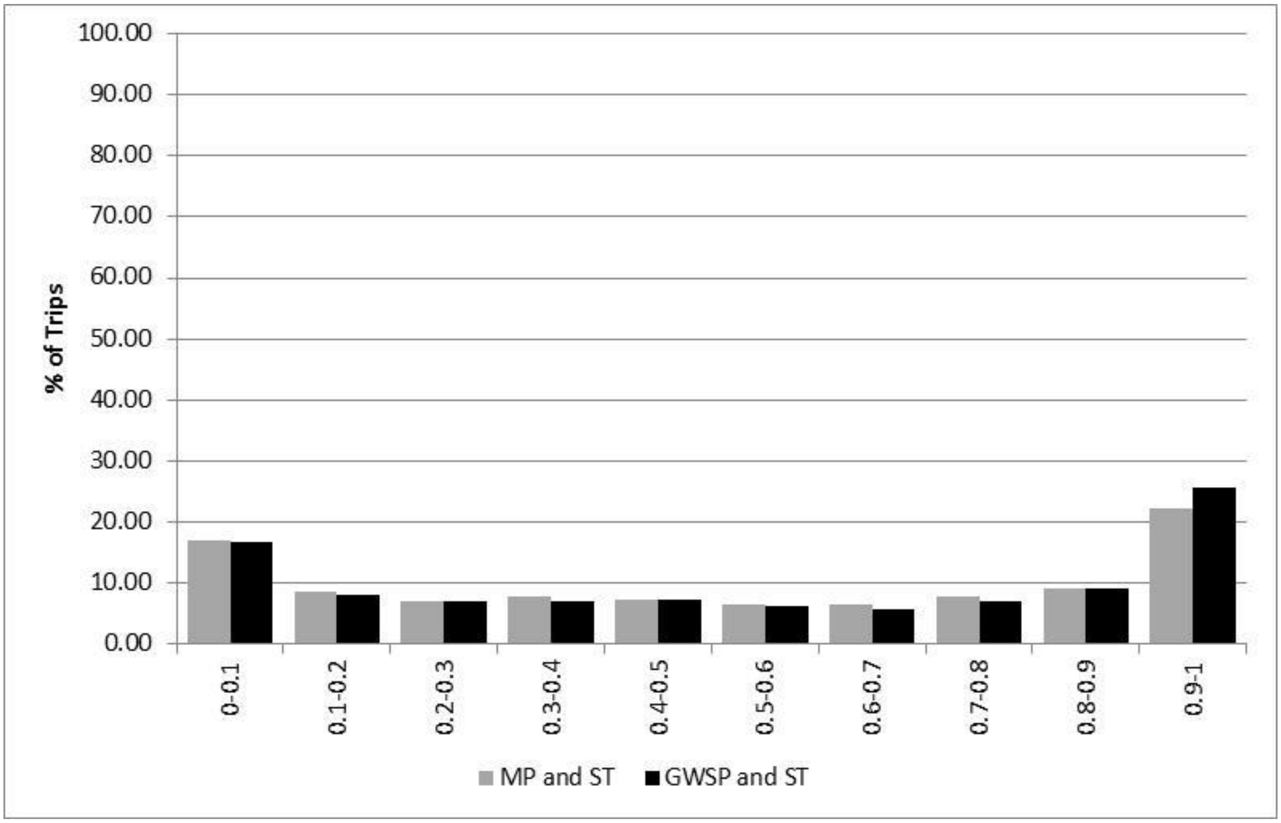


Figure 4-10. Commonly ratio (CR) of routes from the MM and ST methods

Table 4-5. CR for the chosen routes and the ST routes

| | | Peak | | | Off-peak | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Short | Medium | Long | |
| | Sample size | 499 | 301 | 330 | 1219 | 623 | 541 | 3513 |
| CR of MP and ST Routes | Mean | 0.58 | 0.51 | 0.42 | 0.59 | 0.51 | 0.45 | 0.53 |
| | Median | 0.63 | 0.51 | 0.39 | 0.67 | 0.49 | 0.41 | 0.53 |
| | Std. Dev. | 0.35 | 0.35 | 0.33 | 0.35 | 0.35 | 0.32 | 0.35 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.26 | 0.16 | 0.10 | 0.26 | 0.19 | 0.14 | 0.19 |
| | Q3 | 0.93 | 0.85 | 0.69 | 0.93 | 0.86 | 0.74 | 0.88 |
| CR of GWSP and ST Routes | Mean | 0.60 | 0.53 | 0.45 | 0.61 | 0.52 | 0.47 | 0.55 |
| | Median | 0.67 | 0.52 | 0.42 | 0.69 | 0.51 | 0.44 | 0.57 |
| | Std. Dev. | 0.36 | 0.36 | 0.34 | 0.36 | 0.35 | 0.34 | 0.36 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Q1 | 0.28 | 0.17 | 0.10 | 0.28 | 0.19 | 0.15 | 0.20 |
| | Q3 | 0.98 | 0.88 | 0.74 | 0.99 | 0.88 | 0.79 | 0.91 |

From the Table 4-8, for peak travel, TDI increases with the trip length, suggesting that the more time is spend on the roads compared to the shortest-time between the OD pair and disparity is greater for longer trips. Therefore, with more distance travelled during peak period, more time is spent in the congestion. However, for the off-peak travel less disparity is observed for longer trips.

On an average (median), a time along the chosen route for a short distance trips is about 8% longer than the shortest-time path for the same OD pair. The corresponding values are 11% and 14% for medium- and long- distance trips respectively.

Table 4-6. TDI of the chosen routes with the ST routes

| | | Peak | | | Off-peak | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Short | Medium | Long | |
| | Sample size | 499 | 301 | 330 | 1219 | 623 | 541 | 3513 |
| TDI of MP routes with ST Routes | Mean | 0.18 | 0.19 | 0.19 | 0.18 | 0.18 | 0.16 | 0.18 |
| | Median | 0.08 | 0.11 | 0.14 | 0.08 | 0.11 | 0.12 | 0.10 |
| | Std. Dev. | 0.35 | 0.26 | 0.22 | 0.32 | 0.28 | 0.15 | 0.28 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 4.45 | 2.69 | 2.36 | 4.82 | 3.24 | 0.86 | 4.82 |
| | Q1 | 0.01 | 0.04 | 0.05 | 0.01 | 0.03 | 0.05 | 0.03 |
| | Q3 | 0.22 | 0.26 | 0.25 | 0.22 | 0.23 | 0.23 | 0.23 |
| TDI of GWSP routes with ST Routes | Mean | 0.16 | 0.17 | 0.17 | 0.16 | 0.16 | 0.15 | 0.16 |
| | Median | 0.06 | 0.09 | 0.13 | 0.07 | 0.09 | 0.11 | 0.09 |
| | Std. Dev. | 0.35 | 0.24 | 0.21 | 0.31 | 0.25 | 0.14 | 0.27 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 4.45 | 2.38 | 2.42 | 4.82 | 3.02 | 0.85 | 4.82 |
| | Q1 | 0.00 | 0.01 | 0.05 | 0.00 | 0.02 | 0.04 | 0.01 |
| | Q3 | 0.19 | 0.23 | 0.23 | 0.20 | 0.22 | 0.20 | 0.21 |

Results for the CR and the TDI measures clearly show that even though the chosen routes share only about 53% (overall median) of the time in common with the shortest-time path, the time along the chosen route is only about 10% (overall median) longer than the time along the shortest-time path. This suggests that the travel is made on links with comparable travel times to the links in the shortest-time path.

### 4.2.7 Simultaneously Comparing the Chosen Routes Against the SD and ST Routes

Further, to understand whether the traveler followed the SD route more or the ST route, a simultaneous comparison of the chosen routes against the SD and ST routes was conducted. For this, distance overlap, commonly ratio, of the chosen route with the SD route and the ST route were compared, Figure 4-11. Specifically, if the CR of the chosen route with the SD route is higher than the CR of the chosen route with the ST route then it can be deduced that the traveler followed the SD route more than the ST route. Similarly, if the CR of the chosen route with the ST route is higher than the CR of the chosen route with the SD route then it can be deduced that the traveler followed the ST route more than the SD route.

To avoid the comparison of fairly similar shortest-cost routes, the trips with CR of the SD and ST routes higher than 0.9 were not considered for the comparison. Also, if the observed route have no overlap with either of the shortest-cost routes, it was removed from the analysis.

It is important to note that as the observed route from MP and GWSP are mostly similar, the similar results were obtained for the either route, therefore results are shown only for one MM method.

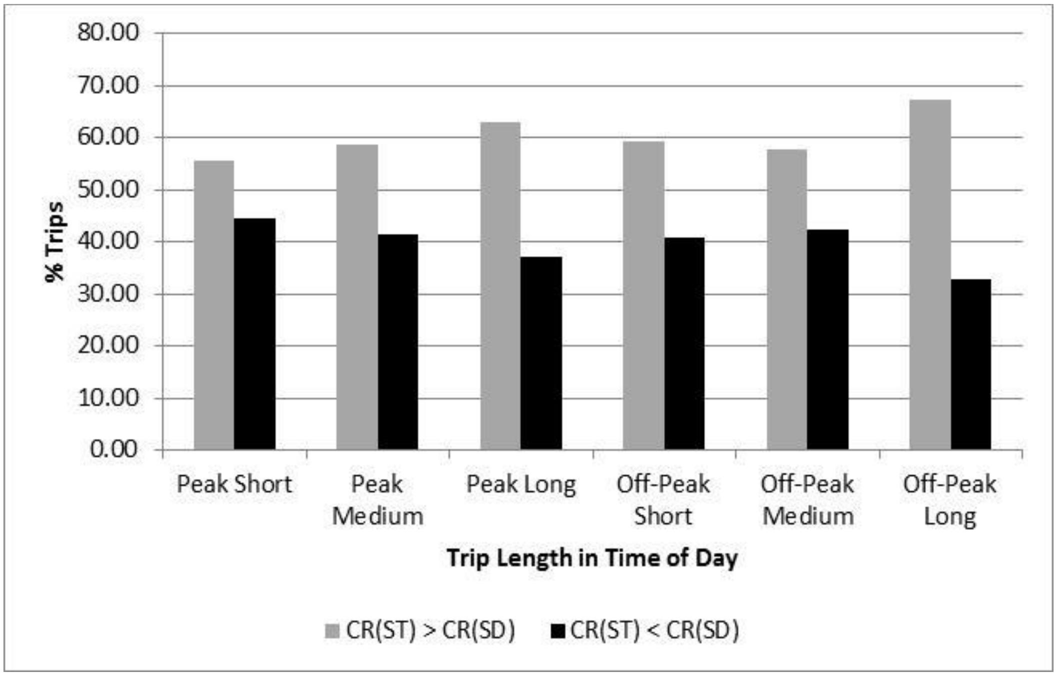


Figure 4-11. Comparison for trip length in time of day

The comparison of the chosen routes with the SD and ST routes for trips length in different time periods is shown in Figure 4-11. The graph clearly shows that trips prefer to travel on ST route more than the corresponding SD route. Moreover, as the trip length increases ST route is followed more. However, ST route is travelled less during the peak period indicating that during congestion people prefer try to follow the SD path more.

## 4.3 Summary

The algorithms were first validated on locally collected GPS data on known routes. While both algorithms performed well, it is observed that the GWSP was more sensitive to missing links in the network than the MP approach. On applying the algorithms to large dataset comprising almost 4000 trips, both algorithms generated complete trips for a vast majority of the cases (again the MP produced mode complete trips than the GWSP).

On comparing the 3500 trips for which both algorithms produced results, we find a substantial similarity between the routes generated in terms of both common links and the extent of distance overlap. This holds irrespective of trip distance and time-of-day of travel. Broadly, this result suggests that the simpler GWSP algorithm might be appropriate to generate chosen routes from GPS traces if a good (complete) roadway network database is available.

On comparing the predicted “chosen” routes to the shortest-distance and shortest-path routes between the same locations, we find that the extent of overlap between the “chosen” and “shortest” paths in terms of the common links is small. The disparity between these routes generally increases with distance but is generally the same across the peak- and off-peak periods. Finally, the “chosen” paths follow the shortest-(free-flow)-time paths more than they follow the shortest distance paths, especially for longer trips. This clearly reflects a preference for using higher-speed facilities for longer trips.

Even though the chosen paths are quite deviant from the shortest paths in terms of the actual links traversed, the overall distance/time along the chosen path is fairly close to the distance/time along the shortest paths. This is possibly because of the presence of alternate (possibly parallel) links/paths in the network that are very comparable in terms of distances and times.

# CHAPTER 5
# DATA ASSEMBLY FOR MODEL ESTIMATIONS

The chapter describes the procedure to assemble the data for model estimations. This process involves two major steps. First, several characteristics of the trip- and trip-maker are determined for use as explanatory variables in the models. This is discussed in Section 5.1. Next, alternative routes are constructed and characterized for each trip leading to the formation of the choice set for each route. An enhanced version of the BFS-LE algorithm proposed by Schussler et al. (2012) is used for choice-set generation. This procedure and the various attributes determined for each of the routes in the choice set are discussed in Section 5.2. This section also presents a descriptive summary of the analysis sample to be used for building the route-choice models.

## 5.1 Trip- And Traveler- Characteristics

The two map-matching algorithms generated observed routes for 3513 routes. Of these trip- and traveler- characteristics were completely determined for 2850 trips. The loss of samples was mainly because of our inability to determine the trip-maker characteristics for over 600 trips. The reader will note that the GPS streams do not directly identify the trip-maker and, therefore, secondary data had to be used to determine the drivers of the vehicles being tracked by the GPS devices. Section 5.1.1 discusses the trip-characteristics and Section 5.1.2 presents the procedure for determining the vehicles' primary drivers, and, hence, the characteristics of the trip makers.

### 5.1.1 Trip Characteristics

Certain characteristics of a trip such as the total travel duration and the time-of-day of travel are directly obtained from the GPS streams. One of the major trip attributes that is not obtainable purely from the traces is the trip purpose. The travel survey did include the residential location (latitude and longitude) of the respondents. If a trip-end fell within a buffer zone of radius 0.5 mile around the home location of the respondent, the trip-end was classified as "home". By this procedure, 1597 out of the 2850 trips are classified as HB trips.

Table 5.1 presents descriptive of the trip characteristics for the sample. 56% of the trips are home-based and among them, about 30% originated at home. Further, majority of the trips (61.4%) are performed during weekdays (excluding Friday) with about 26% during weekends.32.5% of the trips are made during peak period. Also, trips shorter than 5 miles in length are prominent (48.8%) in the sample.

Table 5-1. Trip characteristics

| Characteristic | Share (%) |
|---|---|
| *Trip Type* | |
| From Home | 30.4 |
| To Home | 25.6 |
| NHB | 44.0 |
| *Travel Day* | |
| Weekday (Monday - Thursday) | 61.4 |
| Friday | 12.7 |
| Weekend | 26.0 |
| *Travel Time* | |
| Peak | 32.5 |
| Off-Peak | 67.5 |
| *Trip Length* | |
| Short (2-5 miles) | 48.8 |
| Medium (5-10 miles) | 26.6 |
| Length (>10 miles) | 24.6 |

### 5.1.2 Traveler Characteristics

Traveler characteristics include both household-attributes (such as composition and income) and person-attributes (such as age and gender).

The determination of the household-attributes of the trip-maker was rather straightforward given that there were explicit identifiers in the data linking the GPS traces to specific vehicles, and the vehicles to the households surveyed. The household–attributes determined include size, vehicles, licensed drivers, type, home ownership, and income.

The determination of the person-attributes was not as straightforward as there was no explicit identifier linking the GPS traces to a specific person in the household. As a consequence, a primary driver was identified for every household vehicle using the self-reported (CATI) component of the overall travel survey. The primary driver of a vehicle is the household member who used the vehicle the most during the survey day (for the one-day CATI survey). To a large extent (over 75%), a unique household member was identified as the primary diver for each vehicle. This primary driver was assumed to the trip-maker for all trips associated with the driver's vehicle. The person characteristics determined include age, gender, ethnicity, and employment/student status. Table 5-3 presents descriptive of the travelers' characteristics.

On average, a household in the sample is consist of about 3 members and owns 2 vehicles. The sample is dominated by the households that live in a 1-family detached home (81.1%) and owns the home (91%). Also, about 72% of the households have annual income of higher than $60k.

The sample has a mix of males and females and majority (78%) of the individuals is older than 35 years. Most of the individuals (49.9%) in the sample are of white ethnicity and about 45% refused to reveal their ethnicity. 71.4% of the individuals are employed and only 11% are students.

Table 5-3. Travelers' characteristics of the final sample

| Household characteristics | | Person characteristics | |
|---|---|---|---|
| Attribute | Share (%) | Attribute | Share (%) |
| *Household Size* | 2.84 (1.29)* | *Gender* | |
| *Household Vehicles* | 2.17 (0.97)* | Male | 46.7 |
| *Home Type* | | Female | 53.3 |
| 1-family detached | 81.1 | *Age (years)* | 48.08 (18.45)* |
| 1-family attached to other houses | 11.8 | < 16 | 5.7 |
| Building with multiple apartments | 5.9 | 16 – 25 | 6.1 |
| Refused | 1.3 | 25 – 35 | 10.1 |
| *Home Ownership* | | 35 – 45 | 17.2 |
| Owned | 91.0 | 45 – 55 | 27.3 |
| Rented | 7.6 | >55 | 33.5 |
| Refused | 1.4 | *Ethnicity* | |
| *Household Income(per year)* | | White | 49.9 |
| Less than $20,000 | 1.2 | Black/ African American | 3.7 |
| $20,000 - $34,999 | 6.4 | Other | 1.2 |
| $35,000 - $49,999 | 8.8 | Refused | 45.1 |
| $50,000 - $59,999 | 4.9 | *Employment* | |
| $60,000 to $74,999 | 8.7 | Yes-Full Time | 52.8 |
| $75,000 to $99,999 | 27.5 | Yes-Part Time | 18.6 |
| More than $100,000 | 35.8 | Not | 23.5 |
| Refused | 6.7 | Refused | 5.1 |
| *Time at current home location* | | *Student* | |
| New (<2 years) | 9.1 | Yes-Full Time | 7.2 |
| Medium (2-10 years) | 44.7 | Yes-Part Time | 3.8 |
| Experienced (>10 years) | 46.2 | No | 89.0 |

Note: *presents mean (std. deviation)

## 5.2 Determination And Characterization Of Alternate Routes

A major step in the data assembly process is to determine a set of alternate routes for each OD pair (i.e. the choice set) and to characterize the routes in the set using several attributes. The Breadth First Search Link Elimination (BFS-LE) procedure to determine the alternate routes is discussed in 5.2.1 and the characterization of the routes is described in Section 5.2.2

### 5.2.1 Breadth First Search Link Elimination (BFS-LE)

The fundamental BFS-LE algorithm from Schussler et al. (2012) was implemented using Visual Basic Application (VBA) in ArcGIS including some operational enhancements. The algorithm begins by constructing the shortest-cost path considering the full network. Subsequently, a BFS-LE tree (i.e. the set of alternate routes for the OD pair) is developed by repeatedly constructing shortest-cost paths after removing a link from a previously constructed shortest-path. In this study, free-flow travel time on links is used as the generalized cost and the built-in shortest-path calculation tools from ArcGIS are used.

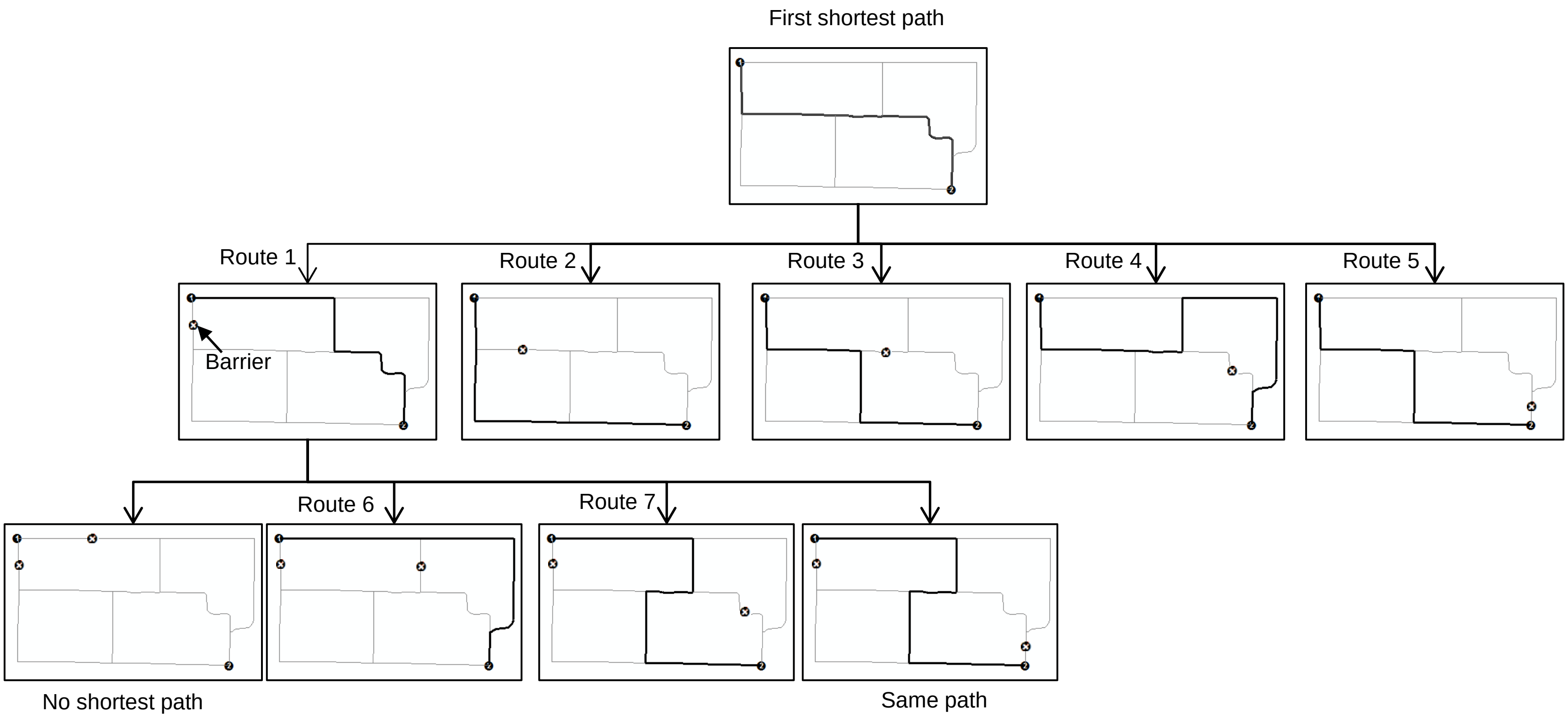


Figure 5-1. Basic BFS-LE tree

Figure 5-1 presents a conceptual summary of the approach. Each “box” is a node of the BFS-LE tree and comprises a sub-roadway network (The root node or the top-most node has the full roadway network). All nodes at the same level (or depth) represent sub-networks that were obtained by removing links from the same previously-generated shortest path. In the current implementation, the “removal of a link” is implemented in ArcGIS by using a point barrier at the mid-point of the link. All the nodes (sub networks) at the same depth are examined before proceeding to nodes at a lower depth implying a breadth-first search.

Some iteration may result in no feasible shortest path. Moreover, similar paths could be calculated at different nodes. In that case, only the first path is saved into the choice set.

The following implementation details are of interest.

In a high-resolution network, a least cost path between an OD pair can easily contain a large number of nodes and links as even the local streets and the corresponding intersections are represented. Further, even a continuous roadway segment could be represented by several links of small lengths (in other words not all nodes are intersections) to capture differences in one or more attributes along the roadway. Thus, the number of possible links that could be removed is too large making the BFS-LE tree quite complex. At the same time, removing links that are simply a part of a contiguous segment would not lead to truly alternate paths thereby affecting the overall computational efficiency of the algorithm. To address this issue, only links that end on intersections are considered as possible candidates for “elimination”.

It is useful to note that Schussler et al. 2012 addressed this issue by adjusting the roadway topology by “dissolving” adjacent links that are not connected by an intersection into bigger segments. Essentially, this procedure ensures that all the nodes in the corrected network are intersections. However, one could lose the detailed roadway-attribute information during this process as these get averaged to represent the longer segment. Our approach of eliminating only links ending in intersections addresses the computational issues while still preserving the detailed roadway characteristics.

Further links that are very close to the origin and/or destination (within 0.15 miles) are also not candidates for elimination as these can quickly lead to sub networks with infeasible paths. This procedure was also adopted by Park and Rilett (1997).

Once a link has been eliminated and a sub-network has been developed, the algorithm first checks for the uniqueness of the sub-network. The reader can see that it is possible to arrive at the same sub network (node) by removing the same set of links but in a different order. Ensuring uniqueness prior to re-running the shortest-path within the sub network is another approach to reduce processing time.

Once a shortest path has been identified within a sub network, its “uniqueness” is examined. For this purpose, the commonly factor is calculated between the newly identified route and all the previously identified routes in the choice set. If this factor is less than 0.95, the new route is introduced into the choice set as another alternative. If not, the new route is deemed to be very similar to one of the routes already in the choice set and is therefore excluded.

Finally, it is also very useful to set termination conditions. The algorithm is set to terminate after generating a minimum set of alternatives for each route (20 in this study). However, sometimes it might take an excessively long time to identify 20 alternatives. To take care of this issue, a maximum run time threshold is also imposed with this time being dependent on the speed of the computer being used.

The algorithm was run for 2692 of the 2850 trips (the trips longer than 25 miles were excluded considering run time issues). As shown in Figure 5-2 and Table 5-4, for over 72% of these trips, 14 or more alternative routes were generated within stipulated run times. Less than 15% of the cases had fewer than 10 routes and less than 2% had fewer than 5 routes.


Figure 5-2. Frequency distribution of choice set size

Table 5-4. Choice set size

| Choice set size | Count | % Share |
|---|---|---|
| >= 14 | 1941 | 72.10 |
| >= 10 | 2303 | 85.55 |
| >= 5 | 2653 | 98.55 |

Further, the observations with fewer than 14 alternatives in the choice set were examined for not having enough routes. As the path generation algorithm starts with the true shortest time route and generates more alternatives by removing links from it, the comparison of link counts in these routes may provide an insight into the behavior. The comparison of the link count in the shortest time route is presented in Table 5-5 and Figure 5-3. The comparisons indicate that the average link count in the shortest time route of the observations with fewer than 14 alternatives is significantly higher than the observations with at-least 14 alternatives. Also, as mentioned before, a route is included in the choice set only if it meets the commonly factor criteria of 0.95. Therefore, with longer trips it is more difficult to keep generating highly diverse routes that meet the set CF criteria. This is due to the high trip lengths in the denominator of the CF calculation and would, compare to shorter trips, need a larger portion of non-overlapped distance to meet the CF criteria of 0.95.

Table 5-5. Comparison of link counts in the first shortest time routes

| | With at-least 14 alternatives | | Fewer than 14 alternatives | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Link Count | 74.87 | 48.11 | 147.71 | 81.54 |
| Distance | 5.13 | 3.52 | 10.24 | 5.81 |
| Time | 8.83 | 5.63 | 15.80 | 8.19 |

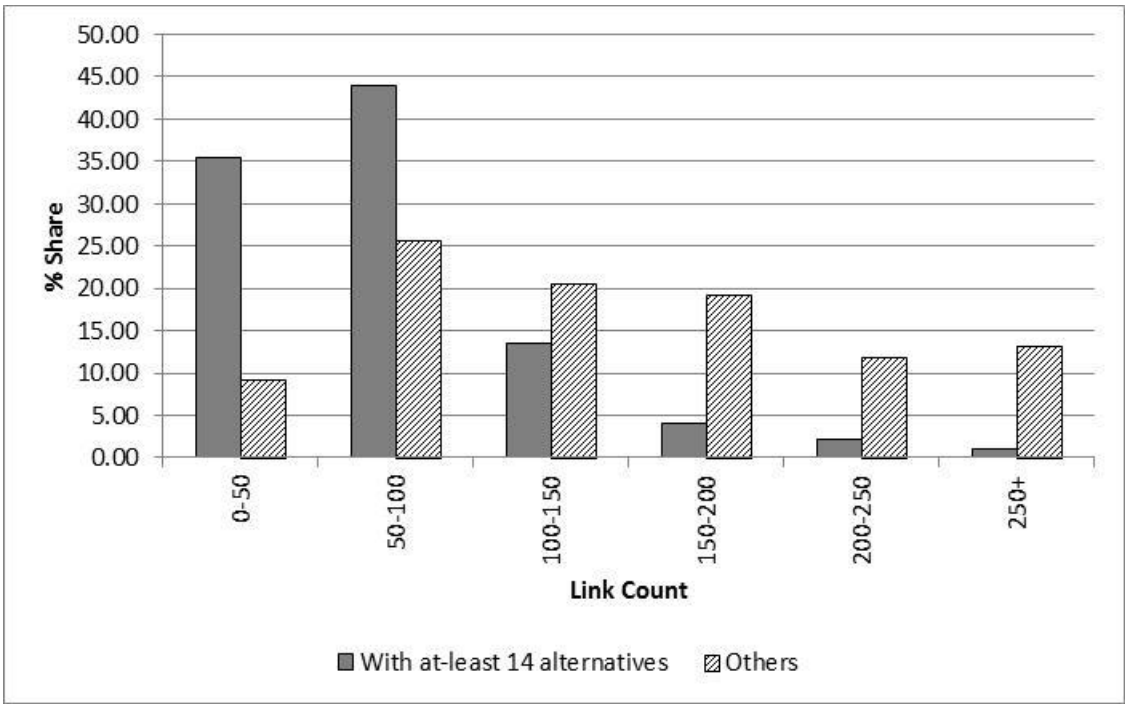


Figure 5-3. Link count comparison

Again, about 72% (1941) of the choice sets have 14 alternatives or higher. With this, we decided to use the choice set size of 15 for model estimations. Therefore, the observations with less than 14 alternatives in the choice set were filtered out. Further, 28 observations out of the remaining 1941 observations had the choice set of 14 alternatives and the chosen route was one of them. Again, such observations were also eliminated. Finally, the estimation sample (CS15) consists of 1913 observations with each having 15 alternatives in the choice set. For the same observations, two more samples with choice set sizes of 10 (CS10) and 15 (CS10) were also constructed. It is important to note that to construct choice sets with size 10, first 10 alternatives generated from the path generation algorithm were considered. A similar process was used for constructing the sample with choice set size 5.

### 5.2.2 Route Attributes

Several attributes were generated for the chosen route and for each of the alternatives in the choice set. A summary of these attributes is presented in table 5-6. The procedures employed to generate these attributes are discussed subsequently.

Table 5-6. Route attributes

| Attribute name | Definition |
|---|---|
| Link Count | Number of links in the route |
| TotalDistance (miles) | Total length of the route |
| TotalTime (minutes) | Total (free flow) time spent in traversing the route |
| IntersectionCount | Number of intersections along the route |
| LongestLegDistance (miles) | Distance of the longest continuous stretch between two intersections |
| LongestLegTime (minutes) | Travel time of the longest continuous stretch between two intersections |
| LeftTurns | Number of left turns made traversing the route (include all left turns and does not distinguish based on the left turn type) |
| RightTurns | Number of right turns made traversing the route |
| ExpresswayDistance (miles) | Distance of the expressway segments in the route |
| ExpresswayTime (minutes) | Travel time on the expressway segments in the route |
| LongestExpresswayDistanceLeg (miles) | Distance of the longest continuous expressway segments |
| LongestExpresswayTimeLeg (minutes) | Travel time on the longest continuous express segments in the route |
| ArterialDistance (miles) | Distance of the arterial segments in the route |
| ArterialTime (minutes) | Travel time on the arterial segments in the route |
| LongestArterialDistanceLeg (miles) | Distance of the longest continuous arterial road segment in the route |
| LongestArterialTimeLeg (minutes) | Travel time on the longest continuous arterial segments in the route |
| LocalRoadDistance (miles) | Distance of the local road segments in the route |
| LocalRoadTime (minutes) | Travel time on the local road segments in the route |
| LongestLocalRoadDistanceLeg (miles) | Distance of the longest continuous local road segments in the route |
| LongestLocalRoadTimeLeg (minutes) | Travel time on the longest continuous local road segments in the route |
| MaxSpeed (mph) | Average speed during the trip |
| MeanSpeed (mph) | Maximum speed attained during the trip |
| Circuity | Deviation in terms of total length from the straight line distance between the origin and destination |

A node is considered as an intersection if there is three or more segments meet on that node. Hence, the number of intersections is calculated by determining the number of nodes with three or more segments.

A leg is defined as the stretch of the route between two intersections. Therefore, the longest leg by distance and time is calculated as the maximum leg distance and leg time respectively for a route.

Number of turns in a route are determined by reading the directions output provided by the route solver in ArcGIS. The directions window explicitly specifies the types of turn, if required, along a route. The output also distinguishes the turns in terms of sharp and normal turns. The text in the output is read to determine the number of turns.

The roads in the network are classified into three categories: freeways, arterials, and local roads. The total distance and time on each road types is calculated and then the corresponding proportions are determined. The longest continuous travel (distance and time) made on each road type is also estimated.

Two measures of speed are calculated for a route: average speed, and maximum speed. The average speed is calculated by taking the time weighted average of the posted speeds on the segments of a route.

Circuity is used as a measure of the route distance deviation from the network-free straight line distance between the origin and destination. The straight line distance (SLD) is calculated using the Haversine formula of calculating distance between two points:

$$\begin{aligned}\mathrm{SLD\ (miles)} = \mathrm{ArcCos}[&\sin(\mathrm{lat1}) * \sin(\mathrm{lat2}) + \cos(\mathrm{lat1}) * \cos(\mathrm{lat2}) \\ &* \cos(\mathrm{long2} - \mathrm{long1})] * \mathrm{R}\end{aligned}$$

Where, lat1 and long1 are the latitude and longitude of a point, and R is the earth radius (3949.99 miles).

The circuity is then calculated by taking the ratio of the route length with the straight line distance. The circuity is always greater than or equal to 1.

$$\mathrm{Circuity} = \frac{\mathrm{Route\ Length}}{\mathrm{SLD}}$$

Finally, it is useful to acknowledge that good estimates of congested travel times were not available for use in this study.

Table 5-7. Descriptive of route attributes

| Attributes | Chosen route | | Choice set alternatives | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | CS15 | | CS10 | | CS5 | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Link Count | 75.03 | 47.86 | 79.75 | 49.66 | 79.07 | 49.59 | 78.29 | 49.50 |
| TotalDistance (miles) | 5.31 | 3.69 | 5.48 | 3.67 | 5.44 | 3.66 | 5.37 | 3.63 |
| TotalTime (minutes) | 9.75 | 6.44 | 9.83 | 6.16 | 9.69 | 6.04 | 9.52 | 5.95 |
| Intersections | 16.28 | 11.21 | 17.75 | 11.40 | 17.60 | 11.43 | 17.36 | 11.36 |
| LongestLegDistance (miles) | 1.47 | 1.77 | 1.35 | 1.49 | 1.35 | 1.52 | 1.34 | 1.50 |
| LongestLegTime (minutes) | 2.70 | 2.83 | 2.39 | 2.49 | 2.37 | 2.29 | 2.34 | 2.25 |
| LeftTurns | 1.31 | 1.21 | 3.18 | 1.55 | 3.04 | 1.53 | 2.81 | 1.51 |
| RightTurns | 1.36 | 1.25 | 3.23 | 1.58 | 3.09 | 1.56 | 2.84 | 1.55 |
| Distance Proportion on Expressways | 0.59 | 5.82 | 0.58 | 5.41 | 0.62 | 5.60 | 0.64 | 5.71 |
| Time Proportion on Expressways | 0.46 | 4.74 | 0.42 | 4.09 | 0.46 | 4.28 | 0.47 | 4.37 |
| LongestExpresswayDistanceLeg (miles) | 0.08 | 0.81 | 0.07 | 0.69 | 0.07 | 0.70 | 0.07 | 0.67 |
| LongestExpresswayTimeLeg (minutes) | 0.08 | 0.90 | 0.07 | 0.77 | 0.08 | 0.78 | 0.07 | 0.75 |
| Distance Proportion on Arterials | 33.14 | 33.52 | 36.50 | 31.07 | 37.01 | 31.35 | 38.00 | 32.09 |
| Time Proportion on Arterials | 29.08 | 31.11 | 31.15 | 27.92 | 31.72 | 28.30 | 32.80 | 29.17 |
| LongestArterialDistanceLeg (miles) | 1.75 | 2.81 | 1.80 | 2.59 | 1.79 | 2.58 | 1.85 | 2.75 |
| LongestArterialTimeLeg (minutes) | 2.57 | 3.99 | 2.60 | 3.65 | 2.58 | 3.62 | 2.67 | 3.86 |
| Distance Proportion on Local Roads | 66.27 | 33.80 | 62.92 | 31.28 | 62.37 | 31.58 | 61.37 | 32.30 |
| Time Proportion on Local Roads | 70.46 | 31.36 | 68.43 | 28.12 | 67.83 | 28.51 | 66.74 | 29.37 |
| LongestLocalRoadDistanceLeg (miles) | 1.08 | 1.94 | 1.19 | 1.61 | 1.16 | 1.59 | 1.13 | 1.59 |
| LongestLocalRoadTimeLeg (minutes) | 2.26 | 3.96 | 2.51 | 3.48 | 2.43 | 3.29 | 2.38 | 3.28 |
| MaxSpeed (mph) | 40.09 | 6.37 | 41.13 | 5.87 | 41.14 | 5.90 | 41.11 | 5.90 |
| MeanSpeed (mph) | 34.19 | 4.98 | 33.17 | 4.64 | 33.32 | 4.68 | 33.55 | 4.76 |
| Circuity | 1.39 | 0.36 | 1.44 | 0.34 | 1.42 | 0.31 | 1.40 | 0.28 |

Distance and time on expressway is really low compared to arterials or local roads, Table 5-7. As most of the trips (64.6%) are shorter than 10 miles, it is reasonable to expect low expressway usage. Among choice sets with different sizes, average distance and average time increase with more alternatives in the choice sets. As the path generation algorithm starts with the true shortest time route, and determines the next shortest time route, next alternative, by eliminating a link in the route, with more alternatives the routes get longer. As expected, the average intersections, left turns and right turns in the chosen route are less than the alternatives in the choice sets. The average mean speed on the chosen routes is higher than the alternatives in the choice sets.

Choice set composition is examined in terms of the maximum overlap of alternatives with the chosen route. The overlaps are calculated using a python script and the results are presented in Figure 5-4 for three datasets.

For CS15, about 36% of the observations already include the chosen route in the choice set. Moreover, more than 51% of the observations replicate at-least 90% of the chosen route in the choice set. These numbers are lower for sample with lesser alternatives; nevertheless more than 40% of the observations include at-least 90% of the observed route. This indicates that the choice sets are composed of reasonable alternatives.

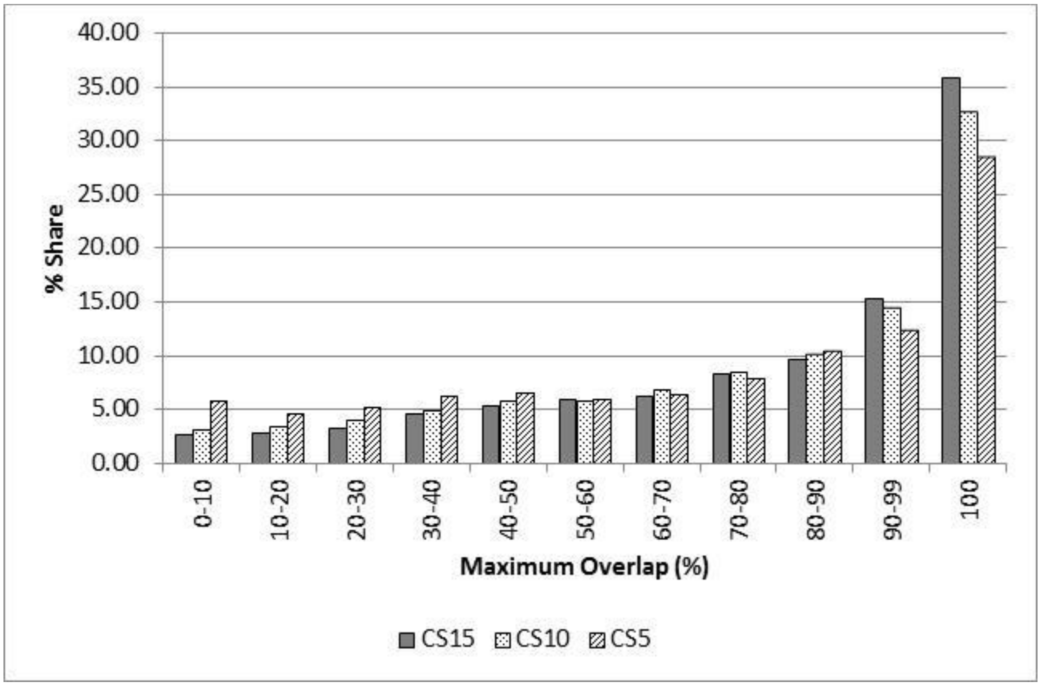

Figure 5-4. Frequency distribution of overlap with the chosen routes

Further, to assess the heterogeneity of alternatives in the choice sets, the variation of the commonly factor (CF) within a choice set is calculated and its distribution over all the observations is examined, Figure 5-5. As expected, the variation in the variation of CF within a choice set grows as the choice set size decreases. Precisely, CS5 contains more diverse routes (higher CF) in the choice sets than the other two datasets. The inclusion of more alternatives in the choice sets results in lesser variation in CF, thus suggesting the presence of more similar routes in the choice sets. The behavior seems reasonable with the fact that fewer alternatives are more likely to have higher diversity than high alternatives. In general, all datasets contain routes with varied commonly factor within the choice sets which suggests that the choice set contains dissimilar routes.

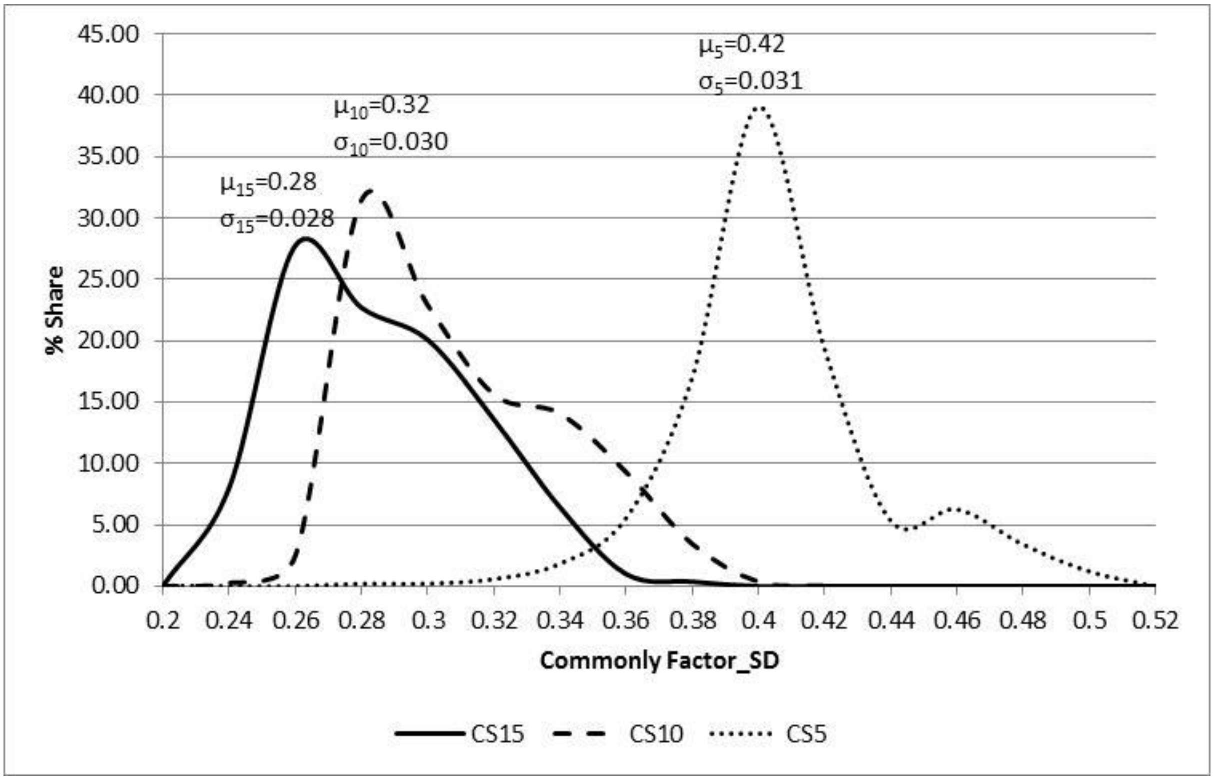


Figure 5-5. Frequency distribution of CF's standard deviation within a choice set

To prepare the samples for model estimations, the chosen route is manually added to the choice set if it is not already present and the extra alternative is eliminated to maintain the choice set size.

## 5.3 Summary

Out of the 3513 observations, for which the map-match algorithms generated routes, 2850 observations had trip- and traveler- characteristics determined. Further, an enhanced version of the Breadth First Search Link Elimination (BFS-LE) was used to generate choice sets for 2692 observations that were shorter than 25 miles. The enhancements to the BFS-LE were aimed to generate diverse and yet attractive routes in the choice sets. The generated choice sets were assessed for the choice set size, coverage and heterogeneity. Over 72% of the generated choice sets contained 14 alternatives or higher. Also, the alternatives in the choice sets provided reasonable coverage of the chosen routes and were found to be different from each other.

# CHAPTER 6
# ROUTE CHOICE MODELS

This chapter presents the empirical results of the route-choice models estimated using data from Chicago. The first section presents an overview of the modeling methodology. The second section presents the model results. The third section discusses the predictive assessments made using the estimated models. The last section presents an overall summary of the chapter.

## 6.1 Path Size Logit (PSL) Model

The path size logit (PSL) model is used in the analysis of route choices. As already discussed, this approach provides a closed-form structure for choice probability, has been shown to have good empirical performance.

This model introduces a similarity term called the Path Size (PS) Factor within the deterministic component of the utility to account for similarities among alternatives in a choice set. The path size factor indicates the fractions of the path that constitutes a “full” alternative (Prato, 2009) and is greater than 0 but less than or equal to 1. In the original formulation, path size for route i in choice set C is given by the following (Ben-Akiva and Bierlaire,1999):

$$\mathrm{PS_i} = \sum_{a\epsilon\Gamma_i} \frac{\mathrm{L_a}}{\mathrm{L_i}} \frac{1}{\sum_{l\epsilon C}\delta_{al}}$$

Where, $\Gamma_i$ is the set of links in path i, $\mathrm{L_a}$is the length of link a in path i, $\mathrm{L_i}$ is the length of the path i, $\delta_{al}$ is the link-path incidence dummy, it is equal to 1 if link a is on path l, and 0 otherwise. If a path is unique, i.e., does not share any link with any of the other alternatives, its PS value is 1.

The probability expression of choosing a route i from the choice set C is given by:

$$\mathrm{P(i)} = \frac{\exp(\mathrm{V_i} + \beta_{PS} * \ln \mathrm{PS_i})}{\sum_{j\epsilon C} \exp(\mathrm{V_j} + \beta_{PS} * \ln \mathrm{PS_j})}$$

Where, $\mathrm{V_i}$ and $\mathrm{V_j}$ are deterministic utilities for routes I and j in choice set C, $\beta_{PS}$ is the parameter for Path Size to be estimated.

In this study, for any given size of choice set, “base” and “full” models are estimated. The two models are estimated for three choice set sizes (5, 10, and 15 alternatives). The base model includes only the route attributes (such as distance, turns, intersections, and circuity) on in the utility function. The trip- and person- characteristics are introduced in the full model in addition to the route attributes. Note that the alternatives in the choice set are not labeled and therefore the trip- and person-attributes are introduced into the utility function by interacting them with the route attributes. Thus, in the full model, the sensitivity to various route attributes are allowed to vary based on the trip- and traveler- characteristics.

## 6.2 Estimation Results

The PSL models were estimated using NLOGIT software. As described in Chapter 5, a total of 1913 observations are selected to develop route choice models. Table 6-1 presents the descriptive of the observations. Table 6.1 presents the descriptive statistics of the explanatory variables of the observations. On average, the households consist of 3 members and 2 vehicles. Most of them are single family and owned the house.

Further, 20.3% households have annual income of less than $60k and 35.7% have annual income of more than $100k. 7.3% households refused to report their annual income. In total, 46.5% of the households were at the current home location for a long period (more than 10 years) and about 10% were new to the location.

Table 6-1. Estimation data descriptive

| Attribute | Share (%) | Attribute | Share (%) |
|---|---|---|---|
| **Household Characteristics** | | **Person Characteristics** | |
| *Household Size* | 2.84 (1.32)* | *Gender* | |
| *Household Vehicles* | 2.18 (0.98)* | Male | 44.1 |
| *Home Type* | | Female | 55.9 |
| 1-family detached | 80.7 | *Age (years)* | 48.25 (15.30)* |
| 1-family attached to other houses | 11.8 | < 16 | -- |
| Building with multiple apartments | 5.8 | 16 – 25 | 9.0 |
| Refused | 1.7 | 25 – 35 | 11.6 |
| *Home Ownership* | | 35 – 45 | 19.8 |
| Owned | 89.8 | 45 – 55 | 27.7 |
| Rented | 8.4 | >55 | 32.0 |
| Refused | 1.9 | *Ethnicity* | |
| *Household Income(per year)* | | White | 55.7 |
| Less than $20,000 | 1.3 | Black/ African American | 3.1 |
| $20,000 - $34,999 | 6.7 | Other | 1.5 |
| $35,000 - $49,999 | 7.6 | Refused | 39.6 |
| $50,000 - $59,999 | 4.7 | *Employment* | |
| $60,000 to $74,999 | 9.9 | Yes-Full Time | 51.7 |
| $75,000 to $99,999 | 26.9 | Yes-Part Time | 20.9 |
| More than $100,000 | 35.7 | Not | 26.8 |
| Refused | 7.3 | Refused | 0.6 |
| *Time at current home location* | | *Student* | |
| New ( < 2 years) | 9.7 | Yes-Full Time | 6.2 |
| Medium (2 – 10 years) | 43.8 | Yes-Part Time | 4.0 |
| Experienced ( > 10 years) | 46.5 | No | 89.8 |
| **Trip Characteristics** | | | |
| *Trip Type* | | | |
| From Home | 31.4 | | |
| To Home | 25.1 | | |
| NHB | 43.5 | | |
| *Travel Day* | | | |
| Weekday (Monday - Thursday) | 62.4 | | |
| Friday | 11.8 | | |
| Weekend | 25.8 | | |
| *Travel Time* | | | |
| Peak | 31.8 | | |
| Off-Peak | 68.2 | | |

Note: * presents mean (std. dev.)

About 60% of the survey respondents were male and on average the age of the individuals lies at the higher end (48.25 years). Specifically, about 60% are older than 45 years in age. About 56% are white Americans and 39.6% did not report their ethnicity. Over 72% are workers and about 90% are not students. In terms of trips, 56.5% are home-based trips and among them more than 31% originated at home. Most of the trips (62.4%) were performed during weekdays with about 12% on Fridays.

Out of the 1913 observations, 80% (1530) were selected to estimate PSL models and remaining 20% (383) were set aside for the validation of the models. Three datasets with the same 1530 observations but different choice set sizes of 15, 10, and 5 alternatives were constructed for the estimations. The results for the “base” models are discussed first followed by a discussion of the “full” models.

The empirical results for the base models (only route attributes are introduced into the utility functions) are presented in Table 6-2. The results are presented for the three choice-set sizes (15, 10, and 5 alternatives). The following discussion applies to the estimation results all three models unless stated otherwise explicitly.

Table 6-2. Base model estimations

| Variable | **CS 15** | | **CS10** | | **CS5** | |
|---|---|---|---|---|---|---|
| | Est. | T-stat | Est. | T-stat | Est. | T-stat |
| Time | -0.081 | -2.85 | -0.046* | -1.37 | -0.066 | -1.64 |
| Shortest time path | 0.899 | 8.68 | 0.519 | 4.92 | -0.060* | -0.60 |
| Longest leg time | 0.059 | 2.32 | 0.051 | 1.84 | 0.051 | 1.60 |
| Left turns per min | -8.257 | -27.77 | -8.110 | -25.98 | -7.626 | -21.86 |
| Right turns per min | -7.791 | -26.92 | -7.691 | -25.25 | -7.272 | -20.79 |
| Intersections per min | -0.483 | -6.04 | -0.484 | -5.67 | -0.538 | -5.65 |
| Proportion of time on Local Roads | 0.017 | 7.33 | 0.018 | 7.29 | 0.021 | 7.31 |
| Circuity | -2.851 | -11.70 | -2.528 | -9.98 | -1.802 | -7.19 |
| Ln (Path Size) | 0.869 | 11.84 | 0.800 | 8.86 | 1.192 | 8.16 |
| Number of Observations | 1530 | | 1530 | | 1530 | |
| Log-likelihood (at convergence) | -2000.70 | | -1698.40 | | -1172.42 | |
| Log-likelihood (equal shares) | -4143.32 | | -3522.96 | | -2462.44 | |
| R-Sqr[1] | 0.517 | | 0.518 | | 0.524 | |
| Adj. R-Sqr[2] | 0.515 | | 0.515 | | 0.520 | |

Notes:
* indicates that the estimate is insignificant at 95% confidence interval
1. R-Sqr = 1-LL(at convergence)/LL(equal shares)
2. Adj. R-Sqr = 1 – {LL(at convergence)-No. of parameters}/LL(equal shares)

As expected, the free-flow travel time was found to be negatively associated with the probability of choosing a route indicating that travelers do avoid the routes with higher travel time. However, the effect of travel time became insignificant at 95% confidence interval if the alternatives in the choice sets are reduced to 10. Further, the shortest time path between the origin-destination (OD) pair is also more likely to be chosen by the traveler (after controlling for the effect of travel time in general). The effect was estimated insignificant for the dataset with the least alternatives (CS5).

Longest (time units) stretch of roadway without intersections also enhances the attractiveness of a route. Higher number of left or right turns per minute of travel along a route was found to negatively affect the choice of the corresponding route. Similarly, intersections per minute of travel were negatively associated with the selection of a route. The time spent on local roads was positively associated with the possibility of choosing a route. However, the effect does not seem unreasonable considering that a significant fraction of the trips in the dataset were of shorter length and did not use the freeway. Further a high-resolution network was used in the choice-set generation resulting in significant volume of local streets in the choice set. Circuity was found to be a negative factor in choosing a route. Travelers prefer to travel on a route that deviate less from the straight-line path between an origin and destination.

Finally, the positive sign on the path size indicates that if a route is less similar to the alternatives, its chances of getting chosen will be high. This positive effect of the path-size variable was also reported in several past studies on route choice models for auto trips (for ex. Prato and Bekhor, 2006; Prato and Bekhor, 2007; Bierlaire, and Frejinger, 2008).

Overall the base models do indicate intuitive effects of several route-specific attributes on choice of route for any trip.

Next, the route attributes were interacted with the trip- and person- characteristics and the results of the “full models” are presented in Table 6-3. Once again the models are estimated for the three choice-set sizes. The discussion applies to the estimation results all three models unless stated otherwise explicitly.

Table 6-3. Full model estimations

| Variable | **CS 15** | | **CS10** | | **CS5** | |
|---|---|---|---|---|---|---|
| | Est. | T-stat | Est. | T-stat | Est. | T-stat |
| Time | -0.102 | -3.18 | -0.059 | -1.58 | -0.068 | -1.50 |
| Time * Peak | 0.104 | 2.20 | 0.076* | 1.49 | 0.034* | 0.61 |
| Shortest time path (STP) | 0.598 | 3.58 | 0.316 | 1.87 | -0.285 | -1.64 |
| STP * Experienced | 0.367 | 1.95 | 0.229* | 1.21 | 0.281* | 1.42 |
| Longest leg time (LLTM) | 0.072 | 2.72 | 0.059 | 2.01 | 0.061 | 1.84 |
| Left turns per min (LTRN) | -8.734 | -16.84 | -8.434 | -15.80 | -7.784 | -13.27 |
| LTRN * From Home | 1.374 | 1.92 | 0.623* | 0.82 | 0.648* | 0.78 |
| LTRN * To home | 2.005 | 2.79 | 1.969 | 2.67 | 1.574 | 1.88 |
| LTRN * New | -2.536 | -2.64 | -1.818 | -1.87 | -2.522 | -2.14 |
| LTRN * Medium | -2.273 | -2.93 | -1.957 | -2.42 | -1.721 | -1.95 |
| Right turns per min (RTRN) | -10.227 | -15.12 | -10.118 | -14.23 | -9.930 | -11.96 |
| RTRN * From Home | 1.404 | 2.06 | 1.352 | 1.85 | 1.595 | 1.91 |
| RTRN * To home | 1.178 | 1.63 | 1.093* | 1.44 | 1.563 | 1.76 |
| RTRN * Experienced | 2.030 | 3.08 | 2.018 | 2.90 | 2.167 | 2.66 |
| Intersections per min (INTR) | -1.223 | -4.50 | -1.298 | -4.46 | -1.217 | -3.67 |
| INTR * Male | 0.348 | 2.28 | 0.393 | 2.44 | 0.491 | 2.65 |
| INTR * Age | 0.008 | 1.70 | 0.008* | 1.51 | 0.005* | 0.90 |
| INTR * Weekend | 0.294 | 1.70 | 0.253* | 1.38 | 0.041* | 0.18 |
| INTR * Non-home based | 0.262 | 1.73 | 0.399 | 2.50 | 0.380 | 2.06 |
| Proportion of time on Local Roads (LRTP) | 0.016 | 3.97 | 0.017 | 4.12 | 0.019 | 3.96 |
| LRTP * Male | -0.008 | -1.80 | -0.008 | -1.79 | -0.008 | -1.59 |
| LRTP * From Home | 0.011 | 2.11 | 0.014 | 2.42 | 0.015 | 2.36 |
| LRTP * To Home | 0.010 | 1.80 | 0.008* | 1.34 | 0.011 | 1.59 |
| Circuity (CRCT) | -6.663 | -11.79 | -6.152 | -10.74 | -5.384 | -8.22 |
| CRCT * Male | 1.940 | 4.84 | 2.117 | 5.17 | 1.471 | 3.24 |
| CRCT * Experienced | 3.310 | 6.24 | 2.923 | 5.56 | 3.110 | 5.03 |
| Ln (Path Size) | 0.884 | 11.77 | 0.822 | 8.87 | 1.280 | 8.47 |
| Number of Observations | 1530 | | 1530 | | 1530 | |
| Log-likelihood (at convergence) | -1935.21 | | -1637.41 | | -1126.24 | |
| Log-likelihood (equal shares) | -4143.32 | | -3522.96 | | -2462.44 | |
| R-Sqr[1] | 0.532 | | 0.535 | | 0.543 | |
| Adj. R-Sqr[2] | 0.526 | | 0.528 | | 0.532 | |

Notes:
* indicates that the estimate is insignificant at 95% confidence interval
1. R-Sqr = 1-LL(at convergence)/LL(equal share)
2. Adj. R-Sqr = 1 – {LL(at convergence)-No. of parameters}/LL(equal share)

The travel time was negatively associated with the probability of choosing a route. Further, travelers indicated less sensitivity to travel time during peak hours than off-peak. The effect is probably capturing the influence of congestion during peak period. As the study uses free-flow travel time, congestion on shortest travel time route may have forced travelers to take a route with higher free-flow travel time. For CS5, this effect was estimated insignificant at 95% confidence interval.

If a route is the shortest time path, a higher preference was given to it than the other alternatives. The travelers who have been living at the current home location for more than 10 years (experienced travelers) preferred the shortest time path even more than other travelers (people who have lived in their current residence for less than 10 years). With experience, travelers have an enhanced knowledge of the network, and therefore, they are more likely to be aware of the shortest time paths than the other travelers.

Left turns per minute of travel were negatively associated with the probability of choosing a route. Further, home-based trips were less sensitive to left turns than non-home-based trips. Among home-based trips, travelers going to home showed more tolerance to left turns than travelers departing on a trip from home. For smaller number of alternatives (10 and 5), the effect of left turns on the trips originating at home was insignificant at 95% confidence interval.

The number of years spent at the current home location was also found to be a factor in willingness to go through the left turns during a trip. The experienced travelers were more tolerant to the left turns and travelers who were relatively inexperienced (less than 2 years at their current home location) showed the most sensitivity to left turns. Perhaps, less time at the current home locations reflects in their limited knowledge of the roadway network and intersection performance of the area. The estimates of the CS10 indicated that the travelers who were living for 2 to 10 years were the most sensitive to the left turns and the ones who had been living for longer durations (more than 10 years) were the least sensitive.

Number of right turns per min of travel was also estimated as a negative factor in choosing a route. Similar to left turns, the route choice decisions for home-based trips were affected less by right turns than non-home-based trips. The travelers were willing to take right turns while leaving from home than when there are going back home. This effect can be explained in conjunction with the effect of left turns on home-based trips. As left turns going to a place becomes right turns coming back using the same path, the results might be reflective of travelers' preference to re-trace on-ward paths when going back home. Experienced travelers were relatively less concerned about making right turns during a trip.

Number of intersections per minute of travel on a route reduced the attractiveness of a route. While choosing a route, men were less sensitive to intersections than women. While making a decision for route choice, younger travelers were more concerned about the number of intersections than the older travelers. The effect is expected as young people are generally like to drive on a route with fewer interruptions. For CS10 and CS5, the effect was estimated insignificant at 95% confidence interval. Further, travelers were more willing to go through the intersections on weekends than weekdays possibly reflecting the difference in traffic volumes at intersections across the days of the week. This effect of travel day was insignificant at 95% confidence interval for datasets with fewer alternatives (10 and 5) in a choice set.

Non-home-based trips were less sensitive to intersections than the home-based trips. These trips are more likely to be made in less-familiar areas (not in the vicinity of home) and, hence, the travelers may not be aware of alternate options.

The proportion of the total travel time spent on the local roads was positively associated with the probability of choosing a route. The effect is a result of high proportion of local roads in the datasets. Men showed less preference to a route with high local street time than women. Time on local roads was given more preference for the home-based trips than the non-home-based trips. Among home-based trips, the trips originating at home chose routes with more time on local roads than the trips that were going to home.

Circuity was a negative factor in choosing a route. Male were less sensitive to circuity than female. Similarly, travelers who were living at the home location for longer time (more than 10 years) were less sensitive to circuity than other travelers. Once again, this could be indicating the knowledge of congested routes in the roadway network. The people with more time in the area are more likely to be aware of the congested routes, therefore, choose to take more circuitous route rather than spending time in traffic.

## 6.3 Predictive Assessment

To assess the predictive quality, the models (base and full) developed using 15 alternatives (the CS15 models) were applied to the validation sample with 383 observations. The prediction performances of the models were compared against the performance of the deterministic method (shortest time path). In each case the extent of the distance overlap of the predicted route with the chosen route was compared.

In the case of the deterministic model, there is exactly one predicted route (the shortest time path). Therefore the determination of the overall of the predicted path and the observed path is straightforward. However, the prediction from the path-size-logit model is in the form of the probability of choosing a route in the choice set. Therefore, we focus on comparing the expected overlap by considering the overlap of each option in the choice set with the chosen route and the probability assigned to each option.

The expected overlap, $\mathrm{E(O)_i}$, for a trip *i* is the probability weighted average of the alternatives' overlap with the chosen route and is calculated as following:

$$\mathrm{E(O)_i} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{P_j} * \mathrm{O_j}$$

Where, N is the total number of alternatives in the choice set, Pj is the probability of choosing route j, and Oj is the physical overlap of route j with the chosen route.

If there are more alternatives in the choice set, then it is more likely to have an option that is closer (greater overlap) to the chosen route. At the same time the probability assigned to each option is reduced (since the probabilities should sum to 1 across the options). If there are fewer options in the choice set used for prediction, the probability assigned to each option is generally higher but the chance of having an option with a greater overlap is lesser.

In the context of the above discussion, we examine the predictive performance when the model (estimated using 15 alternatives) is applied to three different choice-set sizes (4, 9, and 14 alternatives). The last alternative in any choice set assembled for model estimation was the chosen alternative which was added in to the other options generated by the BFS-LE algorithm (if the choice set generated did not automatically include the chosen option). Therefore the alternatives used for predictions represent the first 4, 9, and 14 alternatives that would be generated by the BFS-LE algorithm.

### 6.3.1 Comparison of Predicted Overlaps

The deterministic overlap was simply calculated as the distance overlap between the shortest time path and the chosen path. Table 6-4 shows the statistic of the predicted overlaps.

Table 6-4. Overlaps statistic (full sample)

| | **Base model** | | | **Full model** | | | **Shortest time** |
|---|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| Avg | 50.46 | 51.43 | 52.25 | 50.56 | 51.57 | 52.40 | 58.48 |
| Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max | 96.54 | 96.12 | 97.49 | 96.53 | 96.11 | 97.47 | 100.00 |
| Std Dev | 27.63 | 28.33 | 29.36 | 27.70 | 28.40 | 29.42 | 35.52 |

On average, the deterministic overlaps were higher than the expected overlaps. However, with fewer alternatives, higher expected overlaps were generated. The lower expected overlaps could have been the result of the overlaps' dilution with the route probabilities. This was also reflected in the maximum overlap for the probabilistic routes as it was always less than 100. To get more insight into the results, frequency distributions of the predicted overlaps were performed, shown in Table 6-5.

Table 6-5. Cumulative share of expected overlaps (full sample)

| | **Base model** | | | **Full model** | | | **Shortest time** |
|---|---|---|---|---|---|---|---|
| Overlap | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| 10 | 9.40 | 10.70 | 10.97 | 10.44 | 10.44 | 11.49 | 14.10 |
| 20 | 17.23 | 17.23 | 19.06 | 17.23 | 17.23 | 19.06 | 21.15 |
| 30 | 27.68 | 28.20 | 27.68 | 27.94 | 27.94 | 27.68 | 28.46 |
| 40 | 38.90 | 38.38 | 37.60 | 38.38 | 38.38 | 37.86 | 34.99 |
| 50 | 48.30 | 45.69 | 46.74 | 45.69 | 45.69 | 46.74 | 42.56 |
| 60 | 58.22 | 55.09 | 55.35 | 55.09 | 55.09 | 55.09 | 47.78 |
| 70 | 68.41 | 66.32 | 64.23 | 66.58 | 66.58 | 63.45 | 53.52 |
| 80 | 80.94 | 79.37 | 76.76 | 78.85 | 78.85 | 76.24 | 61.88 |
| 90 | 93.21 | 93.73 | 89.56 | 93.73 | 93.73 | 89.56 | 70.50 |
| 100 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

About 30% of the shortest paths were very close (greater than 90% overlap) to the chosen routes. Moreover, about 71% of those were the chosen routes. This may have influenced the average overlaps for the deterministic paths. Therefore, the validation sample was segregated into two samples. First sample consisted of the 30%

observations for which the shortest path was very close to the chosen route. The remaining 70% observations formed the second sample.

For the first sample, the statistic and frequency distributions of the path overlaps were calculated and are presented in Table 6-6 and Table 6-7 respectively.

Table 6-6. Overlaps statistic (30% observations)

| | **Base model** | | | **Full model** | | | **Shortest time** |
|---|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| Avg | 79.14 | 81.79 | 82.68 | 79.29 | 81.99 | 82.94 | 98.70 |
| Min | 15.81 | 20.22 | 41.33 | 16.55 | 21.29 | 43.22 | 90.49 |
| Max | 96.54 | 96.13 | 97.49 | 96.54 | 96.12 | 97.48 | 100 |
| Std Dev | 13.80 | 11.24 | 12.08 | 13.87 | 11.20 | 11.96 | 2.58 |

Table 6-7. Cumulative overlap with the chosen route (30% observations)

| | **Base model** | | | **Full model** | | | **Shortest time** |
|---|---|---|---|---|---|---|---|
| Overlap | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| <10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 20 | 0.88 | 0.00 | 0.00 | 0.88 | 0.00 | 0.00 | 0.00 |
| 30 | 0.88 | 0.88 | 0.00 | 0.88 | 0.88 | 0.00 | 0.00 |
| 40 | 0.88 | 0.88 | 0.00 | 0.88 | 0.88 | 0.00 | 0.00 |
| 50 | 3.54 | 0.88 | 1.77 | 3.54 | 0.88 | 1.77 | 0.00 |
| 60 | 11.50 | 3.54 | 5.31 | 11.50 | 3.54 | 5.31 | 0.00 |
| 70 | 19.47 | 14.16 | 15.04 | 19.47 | 15.04 | 14.16 | 0.00 |
| 80 | 39.82 | 33.63 | 33.63 | 37.17 | 32.74 | 32.74 | 0.00 |
| 90 | 76.99 | 78.76 | 66.37 | 76.11 | 78.76 | 66.37 | 0.00 |
| 100 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

For the observations where the shortest time path was chosen to make the trip, probabilistic methods predicted routes with at-least 50% overlap with the chosen route. Moreover, for about 80% observations the predicted routes were with an overlap of equal or higher than 70%. This suggests that although probabilistic methods predicted lower overlaps than the deterministic method, the attenuation was low as overlaps were at the higher end.

Further, the statistic and the frequency distributions of the path overlaps for the second sample are shown in Table 6-8, and Table 6-9. In general, the expected overlap was comparable to the overlap of the shortest path with the chosen route.

Table 6-8. Overlaps statistic (70% observations)

| | **Base model** | | | **Full model** | | | **Shortest time** |
|---|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| Avg | 38.46 | 38.74 | 39.52 | 38.54 | 38.84 | 39.62 | 41.66 |
| Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max | 83.17 | 84.56 | 90.77 | 83.16 | 84.52 | 90.54 | 89.72 |
| Std Dev | 22.70 | 23.21 | 24.75 | 22.77 | 23.29 | 24.78 | 28.74 |

Table 6-9. Cumulative overlap with the chosen route (70% observations)

| | Base model | | | Full model | | | Shortest time |
|---|---|---|---|---|---|---|---|
| Overlap | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 | CS1 |
| 10 | 13.33 | 15.19 | 15.56 | 13.70 | 14.81 | 16.30 | 20.00 |
| 20 | 24.07 | 24.44 | 27.04 | 24.07 | 24.44 | 27.04 | 30.00 |
| 30 | 38.89 | 39.63 | 39.26 | 39.63 | 39.26 | 39.26 | 40.37 |
| 40 | 54.81 | 54.07 | 53.33 | 55.19 | 54.07 | 53.70 | 49.63 |
| 50 | 67.04 | 64.44 | 65.56 | 66.67 | 64.44 | 65.56 | 60.37 |
| 60 | 77.78 | 76.67 | 76.30 | 77.78 | 76.67 | 75.93 | 67.78 |
| 70 | 88.89 | 88.15 | 84.81 | 89.26 | 88.15 | 84.07 | 75.93 |
| 80 | 98.15 | 98.52 | 94.81 | 98.15 | 98.15 | 94.44 | 87.78 |
| 90 | 100.00 | 100.00 | 99.26 | 100.00 | 100.00 | 99.26 | 100.00 |
| 100 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

### 6.3.2 Outperforming the Shortest Path

To examine the extent to which the probabilistic path outperforms the shortest time path, the probability of predicting a path with an equal or better overlap than the shortest path is calculated. Therefore, the probability of outperforming is calculated as following:

$$P_i^{\mathrm{Outperform}} = \sum_{j=1}^{N} P_j * \delta_j$$

Where, $P_j$ is the predicted probability of route j, and $\delta_j$ is the overlap performance index, which is equal to 1 if overlap between route j and the chosen route is equal to or greater than the overlap between the shortest time path and the chosen route, and 0 otherwise.

Table 6-10 shows the statistic of the probabilities of outperforming the shortest time path.

Table 6-10. Statistic of the outperforming probabilities (full sample)

| | Base model | | | Full model | | |
|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| Avg | 0.370 | 0.407 | 0.499 | 0.372 | 0.410 | 0.504 |
| Min | 0.025 | 0.042 | 0.093 | 0.021 | 0.034 | 0.076 |
| Max | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Std Dev | 0.349 | 0.357 | 0.336 | 0.348 | 0.356 | 0.333 |

On average, there is a probability of 50% (CS4) that the predicted path will have an overlap of equal to or better than the shortest time path. The frequency distributions of the outperforming probabilities are presented in Table 6-11.

Table 6-11. Cumulative share of the outperforming probabilities (full sample)

| | Base model | | | Full model | | |
|---|---|---|---|---|---|---|
| Probability | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| 0.1 | 35.77 | 37.34 | 0.26 | 35.25 | 35.51 | 0.52 |
| 0.2 | 48.30 | 48.04 | 26.37 | 47.52 | 46.74 | 19.58 |
| 0.3 | 55.09 | 53.52 | 47.26 | 55.09 | 53.52 | 47.26 |
| 0.4 | 61.88 | 57.70 | 48.83 | 61.36 | 57.70 | 48.04 |
| 0.5 | 69.19 | 63.19 | 61.10 | 69.19 | 62.92 | 60.31 |
| 0.6 | 72.85 | 69.45 | 63.71 | 72.58 | 69.45 | 63.71 |
| 0.7 | 75.46 | 73.89 | 65.80 | 75.20 | 73.89 | 65.80 |
| 0.8 | 80.68 | 77.02 | 74.67 | 80.42 | 77.02 | 74.67 |
| 0.9 | 85.12 | 83.03 | 75.98 | 85.12 | 82.51 | 75.98 |
| 1 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

For over 30% of the observations, there was at-least 50% probability of outperforming the shortest path. The probability of outperforming increased with the reduction in the choice set size, as for 4 alternatives about 40% observations showed a probability of 50% or higher.

For the observations for which the shortest path was very close (over 90% overlap) to the chosen route, the average probability of outperforming the shortest path was very low, Table 6-12. Similar results were indicated in the frequency distribution (Table 6-13) of the outperforming probabilities, as most of the observations showed the probability of less than 30%.

Table 6-12. Statistic of the outperforming probabilities (30% observations)

| | Base model | | | Full model | | |
|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| Avg | 0.055 | 0.086 | 0.204 | 0.058 | 0.090 | 0.211 |
| Min | 0.025 | 0.042 | 0.093 | 0.021 | 0.034 | 0.076 |
| Max | 0.135 | 0.223 | 0.515 | 0.140 | 0.223 | 0.511 |
| Std Dev | 0.022 | 0.031 | 0.071 | 0.022 | 0.031 | 0.071 |

Table 6-13. Cumulative share of the outperforming probabilities (30% observations)

| | Base model | | | Full model | | |
|---|---|---|---|---|---|---|
| Probability | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| 0.1 | 92.04 | 92.04 | 0.88 | 92.04 | 87.61 | 1.77 |
| 0.2 | 100.00 | 98.23 | 59.29 | 100.00 | 97.35 | 44.25 |
| 0.3 | 100.00 | 100.00 | 94.69 | 100.00 | 100.00 | 94.69 |
| 0.4 | 100.00 | 100.00 | 94.69 | 100.00 | 100.00 | 94.69 |
| 0.5 | 100.00 | 100.00 | 99.12 | 100.00 | 100.00 | 99.12 |
| 0.6 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 0.7 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.8 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 0.9 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 1 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

For the remaining sample (70% observations), on average (Table 6-14), the probability of outperforming was higher than 50%. Moreover, for about 65% observations, the chances of predicting a path with equal or better overlap than the shortest path was more than 50% (Table 6-15).

Table 6-14. Statistic of the outperforming probabilities (70% observations)

| | **Base model** | | | **Full model** | | |
|---|---|---|---|---|---|---|
| | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| Avg | 0.501 | 0.542 | 0.623 | 0.503 | 0.544 | 0.626 |
| Min | 0.044 | 0.068 | 0.135 | 0.040 | 0.066 | 0.123 |
| Max | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Std Dev | 0.337 | 0.346 | 0.326 | 0.336 | 0.344 | 0.323 |

Table 6-15. Cumulative share of the outperforming probabilities (70% observations)

| | **Base model** | | | **Full model** | | |
|---|---|---|---|---|---|---|
| Probability | CS14 | CS9 | CS4 | CS14 | CS9 | CS4 |
| 0.1 | 12.22 | 14.44 | 0.00 | 11.48 | 13.70 | 0.00 |
| 0.2 | 26.67 | 27.04 | 12.59 | 25.56 | 25.56 | 9.26 |
| 0.3 | 36.30 | 34.07 | 27.41 | 36.30 | 34.07 | 27.41 |
| 0.4 | 45.93 | 40.00 | 29.63 | 45.19 | 40.00 | 28.52 |
| 0.5 | 56.30 | 47.78 | 45.19 | 56.30 | 47.41 | 44.07 |
| 0.6 | 61.48 | 56.67 | 48.52 | 61.11 | 56.67 | 48.52 |
| 0.7 | 65.19 | 62.96 | 51.48 | 64.81 | 62.96 | 51.48 |
| 0.8 | 72.59 | 67.41 | 64.07 | 72.22 | 67.41 | 64.07 |
| 0.9 | 78.89 | 75.93 | 65.93 | 78.89 | 75.19 | 65.93 |
| 1 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

## 6.4 Summary

The PSL model was used to develop route choice models for three different choice sets sizes (CS15, CS10, and CS5). In addition to the trip and route attributes, the probability of choosing a route was examined in terms of the traveler's characteristics. In general expected effected were indicated by the estimation results. Specifically, free-flow travel time, left turns, right turns, intersections, and circuity are found negatively associated with the attractiveness of a route. If a route is the shortest time path, it was more attractive to the traveler. Also, if a route is very different from the other alternatives, it is more likely to be chosen. Additionally, travelers indicated less sensitivity to the travel time during peak period, thus suggesting a congestion effect. Trips going to home were the least sensitive to the travel time and right turns than the other trips. While determining a route, males cared less about the intersections, proportion of time on local roads and circuity than females. The sensitivity to the number of intersections in a route decreased with age. Compared to home-based trips, non-home-based trips were less sensitive to intersections and time on local roads. The effects were more or less similar across different choice set sizes, except that some effects became insignificant.

In terms of the predictive quality, when the shortest time path was very close to the chosen route, the probabilistic methods produced routes with lower overlaps. However, the overlaps were still reasonably high. For the other cases, the probabilistic methods predicted better overlaps than the deterministic method. On average, there was a probability of 50% that the predicted route will outperform the shortest time path.

# CHAPTER 7
# SUMMARY AND CONCLUSIONS

The study combines data from GPS-based travel surveys and Geographic Information Systems (GIS)-based roadway network databases to develop models for route choice. The GPS data comes from the Chicago Regional Household Travel Inventory (CRHTI) and a high-resolution roadway network was obtained from ArcGIS Data and Maps. Three modules are performed during the study: map matching, choice set generation, and route choice modeling. The summary of the three modules is presented below, followed by some insight on extending the research.

## 7.1 Map Matching

The GPS traces must first be mapped to the roadway network to identify the chosen route in terms of the network link actually traversed. Two broad classes of map-matching methods are available. The GPS-weighted shortest path (GWSP) is a simpler approach, but assumes the preference for a “shortest” path in determining the route-choice. The Multi Path (MP) algorithm is free from such assumptions, but is considerably elaborate in terms of methodology.

This study contributes to the map-matching literature by presenting enhanced implementations of both classes of algorithms and comparing their operational performance using data from a large-scale GPS survey. The implementation enhancements were aimed to achieve full automation of the method (no need for manual interventions such as visual inspections) and to improve computational performance. The methods were implemented in ArcGIS using VBA with ArcObjects.

The algorithms were first validated on locally collected GPS data on known routes. While both algorithms performed well, it is observed that the GWSP was more sensitive to missing links in the network than the MP approach. On applying the algorithms to large dataset comprising almost 4000 trips, both algorithms generated complete trips for a vast majority of the cases (again the MP produced mode complete trips than the GWSP).

On comparing the 3513 trips for which both algorithms produced results, we find a substantial similarity between the routes generated in terms of both common links and the extent of distance overlap. This holds irrespective of trip distance and time-of-day of travel. Broadly, this result suggests that the simpler GWSP algorithm might be appropriate to generate chosen routes from GPS traces if a good (complete) roadway network database is available.

On comparing the predicted “chosen” routes to the shortest-distance and shortest-path routes between the same locations, we find that the extent of overlap between the “chosen” and “shortest” paths in terms of the common links is small. The disparity between these routes generally increases with distance but is generally the same across the peak- and off-peak periods. Finally, the “chosen” paths follow the shortest-(free-flow)-time paths more than they follow the shortest distance paths, especially for longer trips. This clearly reflects a preference for using higher-speed facilities for longer trips.

Even though the chosen paths are quite deviant from the shortest paths in terms of the actual links traversed, the overall distance/time along the chosen path is fairly close to the distance/time along the shortest paths. This is possibly because of the presence of alternate (possibly parallel) links/paths in the network that are very comparable in terms of distances and times.

This entire study focused on GPS traces obtained from special-purpose logging devises used in travel surveys. With increasing volumes of GPS trace data available from other sources (tracking of cell phone, blue tooth devices), corresponding methodological innovations are needed to build route-determination algorithms.

## 7.2 Choice Set Generation

Once the chosen route has been determined, the next step is to determine the set of alternate paths available for the same trip. The shortest path based path generation algorithms are recommended for generating alternatives in a high-resolution roadway network. The Breadth First Search Link Elimination (BFS-LE) method has shown to generate attractive alternatives within a reasonable computation time. The study's contribution to the choice set generation literature includes the enhanced implementation of the BFS-LE algorithm. The enhancements were aimed to generate diverse and yet attractive routes in the choice sets. The method was implemented in ArcGIS using VBA with ArcObjects.

Out of the 3513 trips with observed routes, 2850 trips that had person information available were selected for generating choice sets. However, with the available resources and time, the choice sets were generated for 2692 trips. For over 72% of the observations, 14 or more alternative routes are generated within stipulated run times and among them over 51% of the observations replicated at-least 90% of the chosen route in the choice set. With this, 1913 observations that have at-least 14 alternatives are selected for model estimations and three estimation datasets with different choice set sizes (15, 10, and 5) were constructed. The chosen route is manually added to the choice set if not already present.

## 7.3 Route Choice Modeling

The Path Size Logit (PSL) model with original Path Size formulation was adopted to develop route choice models for each of the datasets. The study contributes to the route choice modeling literature empirically by expressing the utility functions in terms of route attributes (time, longest leg time, distance, number of intersections, left turns, right turns, time by facility type, and circuity), trip characteristics (home-based/ non-home-based, weekday/weekend, and peak/off-peak) and traveler's demographics (gender, age, employment, and household income). The models were estimated using NLOGIT software.

The estimation results indicated expected effects. Specifically, free-flow travel time, left turns, right turns, intersections, and circuity were found negatively associated with the attractiveness of a route. Travelers indicated a preference to the route with higher proportion of the travel time on local streets. The effect can be attributed to the large share of local roads in the dataset. The effect of the path size attribute was consistent with the other studies in the literature.

Further, a congestion effect was observed during the peak period as travelers chose to travel on a route with higher free-flow travel time. Sensitivity to the travel time was estimated the least for the trips going back to home. The result indicated that the travelers are more concerned about the travel time when they travel to a place other than home.

While the current study used route attributes that could be quickly generated using GIS methods, future efforts should also aim to incorporate additional qualitative aspects of the routes. Further, focus should also be placed on the conduciveness of the routes for alternate modes such as transit and pedestrians.

Men showed a higher tolerance to the number of intersections and the circuity of the route than women. Additionally, compared to women, they indicated a lower preference to the routes with higher travel time proportions on local roads. Sensitivity to number of intersections decreased with traveler's age. Younger travelers gave more preference to the routes with fewer intersections than the older travelers. Furthermore, the intersections were less of a negative factor in choosing routes for home-based trips than non-home-based trips.

Across different choice set sizes, the effects were more or less similar except that some effects became insignificant in estimations with fewer alternatives.

In terms of the predictive quality, when the shortest time path was very close to the chosen route, the probabilistic methods produced routes with lower overlaps. However, the overlaps were still reasonably high. For the other cases, the probabilistic methods predicted better overlaps than the deterministic method. Further, on average, there was a probability of 50% that the predicted route will outperform the shortest time path.

We envision this study as an important contribution towards the development of empirically rich route choice models. With increasing numbers of GPS surveys and benefits of using high-resolution roadway network, the availability of speedy automatic procedures to generate the chosen routes and alternatives is critical. Further, the examination of route choice behavior in terms of travelers' demographics provides more insight into the route choice decisions.

# APPENDIX A
# LITERATURE SUMMARIES

## Map Matching

| Study | Dataset | Methodology | Chosen route |
|---|---|---|---|
| *On-line Methods* | | | |
| Velaga et al. (2009) | Three pre-defined routes: two in urban areas and one in a suburban area | Online map matching<br>Weight based topological map matching<br>GPS points to links<br>Used heading, proximity, and two weights for turn restrictions at junctions and link connectivity<br>Two consistency checks | For urban areas, 96.8% and 95.3% and for suburban area 96.71% of the total links were correctly identified |
| Quddus et al. (2003) | A test vehicle with a GPS receiver was driven on a carefully chosen route in London | Online map-matching<br>Used network topology, vehicle heading and speed information<br>Matches GPS points to the links | An efficient method in particular for conditions such as junctions and intersections |
| *Off-line Methods* | | | |
| Chung and Shalaby (2005) | 60 multimode trips including 24 auto trips in Downtown, Toronto | Approach by Greenfeld (2002)<br>GPS points to links<br>Used network topology<br>Distance and azimuth between GPS points and network links | Visual inspection<br>78.5% correctly for all modes<br>86% for autos |
| Tsui and Shalaby (2006) | Transit system | Method by Chung and Shalaby (2005)<br>Additionally, introduced interactive link matching sub-system to improve efficiency | |
| Du (2005) | 674 trips on 18 known routes in Lexington, KY | Shortest path satisfying network topology | Visual inspection, 95% routes constructed entirely |
| Griffin et al. (2011) | In a 70 square mile geographical area surrounding Wichita Falls, Texas<br>200 routes | Used driving directions (DD) services from web service provides to get the initial route by providing way points, shortest path<br>Set of rules to remove troublesome points<br>Manual inspection required to identify troublesome routes | Visual inspection confirmed a 100% accuracy<br>Significant amount of time is spent to identify problematic routes. |
| Spissu et al. (2011) | 697 trips | A geometric map matching in ArcGIS<br>Manual intervention is required | Matched only 58% of the trips (393 routes) |

| Study | Dataset | Methodology | Chosen route |
|---|---|---|---|
| | | | Reasons: missing GPS points, missing links in the network etc. |
| Song et al. (2010) | 12 trips in Osaka city, Japan | A pipeline approach<br>Data filtering to obtain high quality trajectories<br>Two curve-to-curve based map-matching algorithms based on – Hausdroff distance and Frechet distance | Frechet distance performs better but with high run time |
| Marchal et al. (2005) | 84 paths collected in the Zurich area | Multiple hypothesis technique<br>GPS points to link<br>Used network topology<br>Multiple paths are stored and the best path with a lowest score is chosen | In most of the cases, no continuous routes because of irregular GPS streams |
| Schüssler and Axhausen (2009) | 3932 car stages with 2.4 million GPS points<br>High-resolution swiss network with 408,636 nodes and 882,120 links<br>In all, 250 OD pairs | Multiple Hypothesis Technique<br>Follow the sequence of GPS points<br>Matches GPS points to the links<br>Several route candidates are kept in memory<br>Maximum saved paths between 20 and 40 | Very low matched routes<br>Main reasons: Missing links in the network, Off-network travel, U-turns<br>No validation |
| Menghini et al. (2010) | 3387 bike stages and 2498 unique OD pairs | Algorithm by Schüssler and Axhausen (2009) | |
| Zhou and Golledge (2006) | A test run with a single GPS trace | Multiple hypothesis technique with rank aggregation<br>Matches GPS points to the links<br>Combination of Accumulated 2-norm distance and rotational variation metric to decide the ranking of the candidate paths | Performance is comparable for small candidate size and large network |

Choice Set Generation

| Study | Dataset | Methodology | Choice set |
|---|---|---|---|
| Schussler et al., 2012 | On-person GPS data<br>36,000 car trips by 2,434 persons | Breadth-first search link elimination<br>Compared with stochastic | Choice set size: 20-100 routes |
| Spissu et al., 2011 | GPS data from smart phones<br>All purpose<br>393 observed routes<br>Two leg survey of 1 week each | Min-cost algorithm through existing Cagliari model with cost function of time and distance | Observed route and last 10 min-cost paths from the simulation |
| Quattrone and Vieteatta, 2011 | Road side survey (280 chosen routes) of trucks compared with on-board GPS (52 routes) trips | k-shortest path<br>Combination of 5 criteria | 30 routes<br>Overlap of 75% chosen routes |
| Papinski and Scott, 2011b | GPS data<br>Home-based Work Trips<br>In all 237 trips | Shortest path using Potential path area (PPA) concept<br>k-shortest path algorithm | k-shortest path generates only 52 out of 237 chosen routes |
| Pillat et al., 2011 | GPS sensor smart phones<br>Home to work Analysis<br>300 participants<br>total 18,300 trips with 61 trips per person<br>Personal interview for known routes by clicking link by link on a digital map | Path enumeration<br>Parameters were estimated by using known routes from the questionnaire | Replicates 60% of the actual chosen routes<br>Maximum allowed commonly factor of 0.90 |
| Menghini et al., 2010 | GPS data<br>3387 bike stages | Breadth-first search algorithm | |
| Frejinger et al., 2009 | Synthetic Data<br>A network with 38 nodes and 64 links | Probabilistic Method with Random Walk Algorithm | |
| Bekhor and Prato, 2009 | From two previous studies, Ramming 2002 and Prato 2005<br>228 observations in Turin, Italy dataset and 181 observations in Boston dataset | Labeling, Link elimination, Link penalty, Simulation, and Branch and bound | |
| Prato and Bekhor, 2006<br>Prato and Bekhor, 2007 | No GPS data<br>Home to Work Analysis<br>Web-based survey of faculty and staff members<br>Chosen routes and possible alternatives<br>236 routes, 339 possible alternatives and 182 different ODs | Branch and Bound<br>Compared with other three:<br>Labeling (4 labels)<br>Link Elimination (10 iterations)<br>Link Penalty (15 penalties)<br>Simulation (25 and 35 draws) | For BB, median size 17 and maximum of 44 routes<br>Merged choice set with median size of 32 and maximum 55 routes |

| Study | Dataset | Methodology | Choice set |
|---|---|---|---|
| Bierlaire and Frejinger, 2005 | GPS data<br>Road network with 3077 nodes and 3843 links<br>1282 observed routes with 927 OD pairs | Simulation method<br>Truncated normal distribution with 20 draws<br>Mean and variance based on the observations<br>The observed route is inserted to the choice set | Average of 9.3 routes<br>Maximum 22 and min 2 |
| Ramming, 2002 | No GPS data<br>Home to Work Analysis<br>Questionnaire survey<br>188 observations | Labeling (16 labels)<br>Link Elimination (2-49 unique paths)<br>Link Penalty (3% for origins that are close to MIT, 5% for most distant ones and 4% for remaining)<br>Simulation (48 draws – Gaussian distribution with mean and standard deviation from the model of travel time perception) | Median of 30 routes<br>Maximum up-to 51 routes<br>160 routes met 80% overlap criteria |
| Ben-Akiva, 1984 | | Labeling method<br>10 labels<br>Time, distance, scenic, signals, capacity, hierarchical travel pattern, quality of pavement, commercial dev., highway, congestion | Time and distance reproduce 70% of the chosen routes<br>All labels together reproduce 90%<br>In the end, 6 labels<br>Time and distance are most effective<br>Signals fails to be a significant factor<br>Factors other than time and distance plays effective role |

Route choice Modeling

| Study | Data | Route choice model | Explanatory variables |
|---|---|---|---|
| Frejinger et al. (2009) | Synthetic dataset | Expanded PS<br>Compared with the original PSL model | Length, and Number of speed bumps |
| Schussler and Axhausen (2010) | GPS data collected in Zurich with an on-person GPS logger<br>1500 observations | C-logit, and PSL | Time-of-day dependent travel times on each road types (motorway(MW), extra-urban main(EUM), urban main(UM), and local road(LR)), travel time proportions on each road type, and road type specific path sizes. |
| Bekhor et al. (2006) | No GPS data<br>Questionnaire survey of faculty and staff at MIT, Boston<br>Home-to-work<br>159 observations with choice sets consisting more than 1 routes | Four models: MNL, PSL, and CNL (two models) | Distance, free-flow time, dummy variables for road sections (Mass. Pike, Tobin Bridge, and Sumner Tunnel), time spend on government numbered routes, delays for different income categories, and dummy for least distance and time paths |
| Bovy et al. (2008) | Two datasets from different regions, Turin (228 observations) and Boston (181 observations)<br>Forecasting probabilities: a simple hypothetical network of a single OD pair with 12 available routes | Proposed Path Size Component Logit (PSCL) model.<br>Compared with: MNL, PSL, and PSCL | Total length, travel time, % of travel time on major roads, dummy for the path with the maximum average speed, and % delay with respect to the free-flow time |
| Prato and Bekhor (2006) | No GPS data<br>Web-based survey in Turin, Italy<br>Home-to-work<br>236 chosen routes<br>339 possible alternatives<br>182 different ODs | Six models:<br>MNL, C-logit, PSL, GNL, CNL, and LNL | Level-of-service(distance, free-flow time, and travel time), landmark dummy variables (1 or 0), and behavioral variables (habit, spatial ability, and familiarity) |
| Bliemer and Bovy, 2008 | Forecasting probabilities: a simple hypothetical network of a single OD pair with 12 available routes | MNL, C-logit, PSL, PSCL, PCL, and CNL | |

| Study | Data | Route choice model | Explanatory variables |
|---|---|---|---|

| Study | Data | Route choice model | Explanatory variables |
|---|---|---|---|
| Bekhor et al. (2002) | No GPS data<br>Questionnaire survey of faculty and staff at MIT, Boston<br>Home-to-work<br>159 observations | Adaptation of the logit kernel (LK) model to a route choice situation.<br><br>Compared with: MNL, and PSL | Distance, free-flow time, dummy variables for road sections (Mass. Pike, Tobin Bridge, and Sumner Tunnel), time spend on government numbered routes, delays for different income categories |
| Prato and Bekhor (2007) | No GPS data<br>Web-based survey in Turin, Italy<br>Home-to-work<br>216 observations with at-least 5 alternatives except the chosen route | Six models: MNL, C-logit, PSL, GNL, CNL, and LK with a factor analytic | Level-of-service(distance, free-flow time, and travel time for experience and non-experienced drivers), landmark dummy variables (1 or 0), and behavioral variables (habit, spatial ability, and familiarity) |
| Frejinger and Bierlaire (2007) | Borlange GPS data<br>2978 observations<br>2244 unique observed routes<br>2179 OD pairs | Proposed an Error Component (EC) model using Subnetwork<br>Compared five different specifications of an EC model with: MNL, and PSL<br><br>Forecasting Models: 80% of the observations for the estimation, remaining 20% to validation.<br>In all, five datasets | Path size, estimated travel time, number of speed bumps, number of left turns, and avg. link length |
| Bierlaire and Frejinger (2008) | No real GPS data<br>Reported trip dataset collected in Switzerland<br>780 observations | Two models: PSL, and EC with Subnetwork | Free-flow travel time for each road types with linear time category specification, proportion of the travel time on each road type (freeway, cantonal/ nations, main, and small roads) |
| Bekhor and Prato (2009) | No GPS data<br>Turin and Boston datasets | Three models: MNL, PSL, and LK | Distance, travel time, % of time on major roads, dummy for the path with max. average speed, and % of delay w.r.t. free flow time |

# LIST OF REFERENCES

Azevedo, J.A., M.E.O. Santos Costa, J.J.E.R. Silvestre Mederra, and E.Q. Vieira Martins. An algorithm for the Ranking of Shortest Paths. *European Journal of Operation Research*, Vol. 69(1), 1993, pp. 97-106

Bekhor, S., and C.G. Prato. Methodological Transferability in Route Choice Modeling. *Transportation Research Part B*, Vol. 43, 2009, pp. 422-437.

Bekhor, S., and J.N. Prashker. GEV-based Destination Choice Models that Account for Unobserved Similarities Among Alternatives. *Transportation Research Part B*, Vol. 42, 2008, pp. 243-262.

Bekhor, S., and J.N. Prashker. Stochastic User Equilibrium Formulation for Generalized Nested Logit Model. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1752, 2001, pp. 84-90.

Bekhor, S., and C.G. Prato. Methodological Transferability in Route Choice modeling. *Transportation Research Part B*, Vol. 43(4), 2009, pp. 422-437.

Bekhor, S., M.E. Ben-Akiva, and S. Ramming. Evaluation of Choice Set Generation Algorithms for Route Choice Models. *Annals of Operation Research*, Vol. 144(1), 2006, pp. 235-247.

Bekhor, S., M.E. Ben-Akiva, and M.S. Ramming. Adaptation of Logit Kernel to Route Choice Situation, *Transportation Research Record: Journal of the Transportation Research Board*, no. 1805, 2002, pp. 78-85.

Ben-Akiva, M.E., M.J. Bergman, A.J. Daly, and R. Ramaswamy. Modeling Inter-Urban Route Choice Behaviour. In: Volmuller, J. Hamerslag, R. (Eds.), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*. VNU Science Press, Utrecht, The Netherlands, 1984, pp. 299-330.

Bierlaire, M., and E. Frejinger. Route Choice Modeling with Network-Free Data. Transportation Research Part C, Vol. 16, 2008, pp. 187-198.

Bierlaire, M., and E. Frejinger. Route Choice Models With Subpath Components. In Proceedings of the *5th Swiss Transport Research Conference*, Ascona, Switzerland, 2005.

Bliemer, M.C.J., and P.H.L. Bovy. Impact of Route Choice Set on Route Choice Probabilities. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2076, 2008, pp. 10-19.

Bovy, P.H.L. On Modeling Route Choice Sets in Transportation Networks: A Synthesis. *Transport Reviews*, Vol. 29(1), 2009, pp. 43-68.

Bovy, P.H.L., and S. Fieorenzo-Catalano. Stochastic Route Choice Set Generation: Behavioral And Probabilistic Foundations. *Transportmetrica*, Vol. 3(3), 2007, pp. 173-189.

Bovy, P.H.L., S. Bekhor, and C.G. Prato. The Factor of Revisited Path Size: Alternative Derivation. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2076, 2008, pp. 132-140.

Cascetta, E., and A. Papola. Random Utility Models With Implicit Availability Perception Of Choice Travel For The Simulation Of Travel Demand. *Transportation Research part C*, Vol. 9(4), 2001, pp. 249-263.

Chen, A., P. Kasikitwiwat, and Z. Ji. Solving the Overlapping Problem in Route Choice With Paired Combinatorial Logit Model. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1857, 2003, pp. 65-73.

Chung, E. and A. Shalaby. A Trip Construction Tool For GPS-Based Personal Travel Surveys. *Transportation Planning and Technology*, Vol. 28(5), 2005, pp. 381-401

De la Barra, T., B. Perez, and J. Anez. Multidimensional Path Search And Assignment. *In Proceedings of the 21st PTRC Summer Annual Meeting*, Manchester, England, 1993, pp. 307-319.

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

Dijkstra, E.W. A Note On Two Problems In Connection With Graphs. *Numerical Mathematics*, Vol. 1, 1959, pp. 269-271.

Du, J. Investigating Route Choice And Driving Behavior Using GPS-Collected Data. PhD Dissertation submitted at the University of Connecticut, 2005.

Eisner, J., S. Funke, A. Herbst, A. Spillner, and S. Storandt. Algorithms for Matching and Predicting Trajectories. *In Proc. of the 13th Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2011, pp. 84-95.

Frejinger, E. Random Sampling Of Alternatives In A Route Choice Context. *Proceedings of the European Transport Conference*, Leeuwenhorst, The Netherlands, 2007.

Frejinger, E., and M. Bierlaire. Capturing Correlation with Subnetworks in Route Choice Models. *Transportation Research Part B*, Vol. 41, 2007, pp. 363-378.

Frejinger, E., M. Bierlaire, and M. Ben-Akiva. Sampling of Alternatives for Route Choice Modeling. *Transportation Research Part B*, Vol. 43, 2009, pp. 984-994.

Friedrich, M., I. Hofsab, and S. Wekeck. Timetable-Based Transit Assignment Using Branch & Bound. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1752, 2001, pp. 100-107.

Greenfeld, J. S. Matching GPS Observations To Locations On A Digital Map. *Paper presented at the 81th Annual Meeting of the Transportation Research Board*, Washington, DC, January. (CD-ROM), 2002

Griffin, T., Y. Huang, and S. Seals. Routing-Based Map Matching For Extracting Routes From GPS Trajectories. *In proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*, Vol. 24, 2011, pp. 1-6.

Hood, J., E. Sall, and B. Charlton. A GPS-based Bicycle Route Choice Model for San Francisco, California. *Presented at the Third Innovations in Travel Modeling Conference*, Tempe, Arizona, 2010.

Hoogendoorn-Lanser, S. Modeling Travel Behavior In Multi-Modal Networks. Ph.D. Thesis, TRAIL Research School, Technical University of Delft, The Netherlands, 2005.

Koppelman, F.S., and C. Wen. The Paired Combinatorial Logit Model: Properties, Estimation and Application. *Transportation Research Part B*, Vol. 34, 2000, pp. 75-89.

Kuby, M., X. Zhongyi, and X. Xiaodong. A Minimax Method For Finding The K-Best Differentiated Paths. *Geographical Analysis*, Vol. 29(4), 1997, pp. 298-313.

Li, H., R. Guensler, and J. Ogle. Impact of Objective Route Attributes on the Choice of Primary Morning Commute Route. *Transportation Research Board Annual Meeting,* CD ROM, 2006.

Lou, Y., C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-Matching For Low-Sampling-Rate GPS Trajectories. *In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, 2009, pp. 352-361.

Manski, C. The Structure Of Random Utility Models. *Theory and Decisions*, Vol. 8(3), 1997, pp. 229-254.

Marchal, F., J.K. Hackney, and K.W. Axhausen. Efficient Map Matching Of Large Global Positioning System Data Sets: Tests On Speed Monitoring Experiment In Zürich, *Transportation Research Record: Journal of the Transportation Research Board*, No. 1935, 2005, pp. 93–100.

Menghini, G., N. Carrasco, N. Schussler, and K.W. Axhausen. Route Choice Of Cyclists In Zurich. *Transportation Research Part A: Policy and Practice*, Vol. 44(9), 2010, pp. 754-765.

Nerella, S., and C.R. Bhat. Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1894, 2004, pp. 11-19.

Nielsen, O.A. A Stochastic Transit Assignment Modeling Considering Differences In Passengers Utility Functions. *Transportation Research Part B: Methodological*, Vol. 34(5), 2000, pp. 377-402.

Nielsen, O.A., R.D. Frederisken. Optimization Of Timetable-Based, Stochastic Transit Assignment Models Based On MSA. *Annals of Operations Research*, Vol. 144(1),2006, pp. 263-285.

Nielsen, O. A., and R.M. Jørgensen. Map-Matching Algorithms for GPS Data—Methodology and Test on Data from the AKTA Roadpricing Experiment in Copenhagen. *Presented at 5th TRIannual Symposium on Transportation ANalysis* (TRISTAN V), Le Gosier, 2004.

NuStats. Chicago Regional Household Travel Inventory Draft Final Report. Prepared by Chicago Metropolitan Agency for Planning, 2008. Available at: http://www.cmap.illinois.gov/travel-tracker-survey

NuStats, GeoStats. Chicago Regional Household Travel Inventory Draft GPS Final Report. Prepared by Chicago Metropolitan Agency for Planning, 2008. Available at: http://www.cmap.illinois.gov/travel-tracker-survey

Papinski, D. and D.M. Scott. A GIS-Based Toolkit For Route Choice Analysis. *Journal of Transportation Geography*, Vol. 19, 2011a, pp. 434-442.

Papinski, D., and D.M. Scott. Modeling Home-to-Work Route Choice Decisions Using GPS Data: A Comparison of Two Approaches for Generating Choice Sets. *Presented at the Transport Research Board Annual Meeting* 2011b, Washington D.C.

Park, D., and L.R. Rilett. Identifying Multiple And Reasonable Paths In Transportation Networks: A Heuristic Approach*. Transportation Research Record: Journal of the Transportation Research Board*, no. 1607, 1997, pp. 31-37.

Pillat, J., E. Mandir, and M. Friedrich. Dynamic Choice set Generation Based on Global Positioning System Trajectories and Stated Preference Data. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2231, 2011, pp. 18-26.

Prashker, J.N., and S. Bekhor. Congestion, Stochastic, and Similarity Effects in Stochastic User-Equilibrium Models. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1733, 2000, pp. 80-87.

Prashker, J.N., and S. Bekhor. Investigation of Stochastic Network Loading Procedures. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1645, 1998, pp. 94-102.

Prato, C.G. Route Choice Modeling: Past, Present and Future Research Directions. *Journal of Choice Modeling*, Vol. 2 (1), 2009, pp. 65-100.

Prato, C.G., and S. Bekhor. Applying Branch-and-Bound Technique to Route Choice Set Generation. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1985, 2006, pp. 19-28.

Prato, C.G., and S. Bekhor. Modeling Route Choice Behavior: How Relevant Is the Composition of Choice Set? *Transportation Research Record: Journal of the Transportation Research Board*, no. 2003, 2007, pp. 64-73.

Pyo, J. S., D.H. Shin, and T.K. Sung. Development Of A Map Matching Method Using The Multiple Hypothesis Technique. *Proceedings of IEEE Intelligent Transportation Systems Conference*, 2001, pp. 23-27.

Quattrone, A., and A. Vietetta. Random and Fuzzy Utility models for Road Route Choice*. Transportation Research Part E*, Vol. 47, 2011, pp. 1126-1139.

Quddus, M., W. Ochieng, L. Zhao, and R. Noland. A General Map Matching Algorithm For Transport Telematics Applications. *GPS Solutions*, Vol. 7(3), 2003, pp. 157-167.

Quddus, M., W. Ochieng, L. Zhao, and R. Noland. Current Map-Matching Algorithm For Transport Applications: State Of The Art And Future Research Directions. *Transport Research Part C*, Vol. 15, 2007, pp. 312-328.

Ramming, S. Network Knowledge And Route Choice. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA, 2002

Richardson, A. Search Models And Choice Set Generation. *Transportation Research Part A: General*, Vol. 16(5-6), 1982, pp. 403-419.

Schuessler, N. Accounting For Similarities In Destination Choice Modeling. PhD Dissertation submitted to ETH ZURICH, 2009

Schüssler, N., and K.W. Axhausen. Map-matching of GPS Points on High-resolution Navigation Networks using Multiple Hypothesis Technique. Arbeitsberichte Verkehrs- und Raumplanung, 568, IVT, ETH Zürich, 2009.

Schussler, N., and K.W. Axhausen. Accounting for Route Overlap in Urban and Sub-urban Route Choice Decisions Derived from GPS Observations. *Paper Presented at Proceedings of 12th International Conference on Travel Behavior Research*, Jaipur, India, December 2009.

Schussler, N.R., Balmer, M., and Axhausen, K.W., 2012. Route Choice Sets For Very High-Resolution Data. *Working Paper for Transport and Spatial Planning*, 2012.

Scott, K., G. Pabon-Jimenez, and D. Bernstein. Finding Alternatives To The Best Path. *In Proceedings of the 76th Annual Meeting of the Transportation Research Board*, Washington, D.C., 1997.

Song, X., V. Raghavan, and D. Yoshida. Matching Of Vehicle GPS Traces With Urban Road Networks*. Current Science*, Vol. 98(12), 2010, pp. 1592 – 1598.

Spissu, E., I. Meloni, and B. Snjust. Behavioral Analysis of Choice of Daily Route with Data from Global Positioning System. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2230, 2011, pp. 96-103.

Syed, S., and M.E. Cannon. Fuzzy Logic-Based Map-Matching Algorithm For Vehicle Navigation System In Urban Canyons. *In proceedings of the Institute of Navigation (ION) national technical meeting*, California, USA (26–28 January), 2004.

Tsui, S. Y. A., and A. Shalaby. Enhanced System For Link And Mode Identification For GPS-Based Personal Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1972, 2006, pp. 38–45.

Van der Zij, N.J., and S. Fiorenzo-Catalano. Path Enumeration By Finding The Constrained K-Shortest Paths. *Transportation Research Part B: Methodological*, Vol. 39(6), 2005, pp. 545-563.

Velaga, N. R., M.A. Quddus, and A.L. Bristow. Developing An Enhanced Weight-Based Topological Map-Matching Algorithm For Intelligent Transport Systems. *Transportation Research Part C: Emerging Technologies*, Vol. 17 (6), 2009, pp. 672–683.

Vovsha, P., and S. Bekhor. Link-Nested Logit Model of Route Choice: Overcoming Route Overlapping Problem. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1654, 1998, pp. 133-142.

Wen, C., and F.S. Koppelman. The Generalized Nested Logit Model. *Transportation Research Part B*, Vol 35, 2001, pp. 627-641.

Yair, T., S. Iwakura, and S. Morichi. Multinomial Probit with Structured Covariance for Route Choice Behavior. *Transportation Research Part B*, Vol. 31 (3), 1997, pp. 195-207.

Yin, H., and W. Ouri. A Weight-Based Map Matching Method In Moving Objects Databases. *In Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM)*, Santorini Island, Greece, June 21-23, 2004.

Zhou, J., and R. Golledge. A Three-Step General Map Matching Method In The GIS Environment: Travel/Transportation Study Perspective. *International Journal of Geographical Information System*, Vol. 8(3), 2006, pp. 243–260