ROUTE CHOICE ANALYSIS: DATA, MODELS, ALGORITHMS AND APPLICATIONS

THÈSE N[O] 4009 (2008)

PRÉSENTÉE LE 30 AVRIL 2008

À LA FACULTÉ DES SCIENCES DE BASE
LABORATOIRE TRANSPORT ET MOBILITÉ
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Emma FREJINGER

M.Sc. in Industrial Engineering and Management, Linköping University, Suède
et de nationalité suédoise

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. M. Bierlaire, directeur de thèse
Prof. M. Ben-Akiva, rapporteur
Prof. P. H. Bovy, rapporteur
Prof. T. Liebling, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2008

Till mamma och pappa

&

À Étienne

Acknowledgments

Many people have directly or indirectly helped me with the research presented in this thesis. This is an attempt to thank them although I know, even before starting, that these words are not representative of the importance of their contributions.

First of all, I cannot enough express my gratitude to my thesis supervisor Michel Bierlaire. His original ideas, perfectionism, enthusiasm and the confidence he has shown in me, have not only directly contributed to the work presented in this thesis, but have also meant a lot to me personally. Moreover, he has introduced me to a network of interesting people spread all over the world. I could not imagine a better mentor and I am forever grateful to him.

During my first two years as research assistant at EPFL, I worked in the Operations Research Group (ROSO) directed by Thomas Liebling. I would like to thank him for the responsibilities he has given me (undergraduate and postgraduate courses), the tricky questions during my presentations and especially the nice atmosphere that he creates in his group. I would also like to thank the other members of ROSO and in particular, Frank Crittin, Rodrigue Oeuvray, Marco Ramaioli and Michela Thiémard for the great moments we have spent together. A special thanks to Gautier Stauffer who hosted me in Boston and who has an inspiring personality.

I would also like to thank all the members of TRANSP-OR who create a very nice work environment. In particular, I would like to thank Carolina Osorio, with whom I am sharing office, for always having a smile on her face, Marianne Ruegg for always being helpful and Michaël Thémans for his friendship and the work we have done together.

I am very grateful to Moshe Ben-Akiva who has contributed to this thesis in several ways. First of all, I thank him for giving me the opportunity to experience the stimulating environment in his lab at MIT. Moreover, during my stay in Boston, he introduced me to Song Gao and we started a project on adaptive route choice modeling which is presented in Chapter 6 of this thesis. He has also contributed with important feedback on all the other chapters in this thesis, in particular, the sampling approach (Chapter 5). I would also like to thank the members of the MIT ITS lab for interesting discussions, especially, Maya Abou Zeid and Charisma Choudhury with whom I have spent many nice moments.

It has been a pleasure collaborating with Song Gao and I thank her for the time we have spent discussing adaptive route choice.

Piet Bovy has given interesting comments on this research, in particular on the sampling approach (Chapter 5) and the analysis of the Path Size Logit model (Chapter 3). I thank him for his feedback and I have also very much appreciated collaborating with him.

I express my gratitude to Mogens Fosgerau who has the gift of giving original comments and finding points of views that nobody else does. We have especially discussed issues related to choice set generation (Chapter 5).

The modeling approach presented in Chapter 4 originated from a research project on mobility pricing sponsored by the Swiss Federal Department of the Environment, Transport, Energy and Communications and the Swiss Federal Roads Authority. This was a collaboration with the Institute for Transport Planning and Systems (Swiss Federal Institute of Technology, Zurich) and of the Institute of Economics (University of Lugano), directed by Kay Axhausen and Rico Maggi respectively. I would like to thank them and their collaborators for valuable comments. The dataset was collected by the Swiss Railways. I am also grateful to Jelena Stojanovic who verified the coherence of the data.

Three years of this research have been financed by the Swiss National Science Foundation, grant 200021-107777/1 (May 2005 to April 2008).

All the good feedback and the simulating work environment I have had during my research have been of great importance. However, the value of sometimes gaining distance to the research should not be neglected. I have spent great time doing ski touring, cross-country skiing, running or just having nice dinners with my friends from Sweden, Switzerland and elsewhere. Thank you!

I could not have finished this thesis without the support from my nearest and dearest. I am more than grateful to my family and I would especially like to thank my parents for always giving me their unconditional support for all the crazy and not so crazy things I have done. I hope that I can manage to become such a good parent.

Étienne, mon amour, has been there all the time during these years of research. No words could thank him enough for his support and constant encouragements. Finally, I am very grateful to our not yet born baby who has been cheering me up with acrobatic movements in my belly during the intense weeks of thesis writing.

Abstract

This thesis focuses on the route choice behavior of car drivers (uni-modal networks). More precisely, we are interested in identifying which route a given traveler would take to go from one location to another. For the analysis of this problem we use discrete choice models and disaggregate revealed preferences data. Route choice models play an important role in many transport applications, for example, intelligent transport systems, GPS navigation and transportation planning.

The route choice problem is particularly difficult to analyze because it involves the modeling of choice behavior in large transportation networks. Several issues need to be addressed in order to obtain an operational model. First, trip observations in their original format rarely correspond to link-by-link descriptions of chosen paths and they therefore need to be matched to the network representation used by the modeler. This involves data processing that can introduce bias and errors. Second, the actual alternatives considered by the travelers are unknown to the analyst. Since there is a large, possibly infinite, number of feasible paths in the network, individual specific choice sets of paths need to be defined. Third, alternatives are often highly correlated due to physical overlap between the paths (shared links). Models with flexible correlation structure are complex to specify and to estimate. Simple models are therefore often used in practice even tough the associated assumptions about correlation are violated. Fourth, most route choice models assume that the decision is performed pre-trip. Their application in a context where drivers receive real-time information about traffic conditions is questionable.

In this thesis we address each of the aforementioned issues. First, we propose a general modeling scheme that reconciles network-free data with a network based model so that the data processing related to map-matching is not anymore necessary. The framework allows the estimation of any existing route choice model based on original trip observations that are described as sequences of locations. We illustrate the approach with a real dataset of reported long distance trips in Switzerland.

Second, a new paradigm for choice set generation in particular and route choice modeling in general is presented. Instead of focusing on finding alternatives actually considered by travelers, we propose an approach where we focus on obtaining unbiased parameter estimates. We present a stochastic path generation algorithm based on an importance sampling approach and derive the corresponding sampling correction to be added to the path utilities in the route choice model. This new paradigm also has implications on the way to describe correlation among alternatives. We argue that the correlation should be based not only on the sampled alternatives but also on the general network topology. Estimation results based on synthetic data are presented which clearly show the strength of the approach.

Third, we propose an approach to capture correlation that allows the modeler to control the trade-off between the simplicity of the model and the level of realism. The key concept capturing correlation is called a subnetwork. The importance and originality of this approach lie in the possibility to capture the most important correlation without considerably increasing the model complexity. This makes it suitable for a wide spectrum of applications, namely involving large-scale networks. We illustrate the model with a GPS dataset collected in the Swedish city of Borlänge.

The final contribution of this thesis concerns adaptive route choice modeling in stochastic and time-dependent networks, as opposed to the static network setting assumed in existing models. Optimal adaptive routing problems have been studied in the literature but the estimation of such choice models based on disaggregate revealed preference data is a new area. We propose an estimator for a routing policy choice model and use synthetic data for illustration. Given the uncertainty related to travel times and traffic conditions in transportation networks, we believe that adaptive route choice modeling is an important direction for future research.

To summarize, this thesis addresses issues related to *data* processing (network-free data approach), *algorithms* for choice set generation (sampling of alternatives) and *models* (subnetwork approach and adaptive route choice model). Moreover, we use real *applications* (Borlänge GPS dataset and reported trips in Switzerland) to illustrate the models and algorithms.

Keywords: route choice analysis, discrete choice models, choice set generation, sampling of alternatives, adaptive route choice, disaggregate revealed preferences data, GPS data

Résumé

Dans cette thèse, nous nous concentrons sur le comportement des automobilistes dans leur choix d'itinéraire. Plus précisément, nous nous intéressons à identifier l'itinéraire qu'un voyageur prendrait pour aller d'un endroit à un autre. Les modèles de choix discret et des données désagrégées de préférences révélées sont utilisés pour l'analyse de ce problème.

Le problème de choix d'itinéraire est particulièrement difficile car il implique une analyse de comportement de choix dans des réseaux de transport de grande taille. Pour obtenir un modèle opérationnel il est nécessaire d'aborder certains problèmes. Premièrement, les observations de choix d'itinéraire correspondent rarement à des descriptions détaillées (arc par arc) du trajet choisi et il faut associer les observations au réseau utilisé par l'analyste. Ceci implique un traitement des données qui peut introduire des erreurs et des biais. Deuxièmement, l'analyste ne connaît pas les alternatives qui sont réellement considérées par les voyageurs. Des ensembles de choix doivent être générés car le nombre d'alternatives possibles dans le réseau est très grand, voir infini. Troisièmement, les alternatives peuvent partager des arcs et sont pour cette raison souvent corrélées. Les modèles qui permettent de définir une structure de corrélation sont complexes à spécifier et à estimer. En conséquence, des modèles simples sont souvent utilisés en pratique malgré le fait que les hypothèses associées ne sont pas vérifiées. Quatrièmement, la majorité des modèles de choix d'itinéraire supposent que le choix est fait avant d'entreprendre le trajet. L'application de ces modèles dans un contexte où les voyageurs reçoivent des informations en route doit être mit en question.

Cette thèse aborde chacun des problèmes susmentionnés. Premièrement, nous proposons un modèle général qui rend le traitement des données superflu en permettant l'utilisation directe des choix observés. Cette approche permet d'estimer n'importe quel modèle de choix d'itinéraire existant basé sur des descriptions de choix sous forme d'une suite de lieux. Nous appelons ce type de données "indépendantes du réseau". Un jeu de données réel, récolté au moyen d'un sondage et composé de choix d'itinéraire de longue distance en

Suisse, est utilisé pour illustrer l'approche.

Deuxièmement, un nouveau paradigme est proposé pour la génération d'ensembles de choix et la modélisation de choix d'itinéraire en général. Plutôt que de se focaliser sur la génération des alternatives réellement considérés par les voyageurs, nous proposons une approche dont le but est d'obtenir une estimation non biaisée des paramètres. Nous présentons un algorithme stochastique pour la génération d'alternatives basé sur une approche d'échantillonnage et développons la correction d'échantillonnage correspondante à ajouter dans les utilités du modèle de choix. Ce nouveau paradigme a aussi des implications sur la manière de modéliser la corrélation qui ne devrait pas seulement être basée sur les alternatives dans les ensembles de choix mais également sur la topologie du réseau. Des résultats d'estimations avec des données synthétiques sont présentés et montrent clairement les forces de cette approche.

Troisièmement, nous proposons une approche pour modéliser la corrélation qui permet à l'analyste de contrôler le compromis entre la simplicité du modèle et le niveau de réalisme. Le concept clé capturant la corrélation est appelé sous-réseau. La contribution principale et l'originalité de cette approche résident dans la possibilité de modéliser la corrélation la plus importante sans augmenter déraisonnablement la complexité du modèle.

La dernière contribution de cette thèse porte sur la modélisation de choix d'itinéraire adaptatif dans un réseau stochastique et dynamique, contrairement à l'hypothèse d'un réseau statique dans les modèles existants. Des problèmes d'itinéraire adaptatif optimal ont été étudiés dans la littérature mais l'estimation de tels modèles de choix basée sur des données désagrégées de préférences révélées est un nouveau domaine. Nous proposons un estimateur pour un modèle de choix de "routing policy" et utilisons des données synthétiques comme illustration. Étant donnée l'incertitude liée à l'état de la circulation dans les réseau de transport, nous pensons que la modélisation de choix d'itinéraire adaptatif est une direction importante pour la poursuite de la recherche.

En résumé, cette thèse aborde des problèmes liés au traitement des *données* (approche avec des données "indépendantes du réseau"), aux *algorithmes* pour la génération d'ensembles de choix (échantillonnage d'alternatives) et aux *modèles* (approche sous-réseau et modèle de choix d'itinéraire adaptatif). De plus, des *applications* réelles sont utilisées pour illustrer les modèles et les algorithmes.

Mots-clés : analyse de choix d'itinéraire, modèles de choix discret, génération d'ensembles de choix, échantillonnage d'alternatives, choix d'itinéraire adaptatif, données désagrégées de préférences révélées, données GPS

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traveling is an important part of many peoples everyday life. Numerous trips are made for going to work, pick up children, going shopping, attending social activities and so forth. Many of these trips are made by car which has led to a number of problems with, for example, congestion and pollution (both in terms of noise and emissions of chemical substances). In turn, these problems have a negative impact on the environment and on peoples' well-being. In order to decrease the negative impact of travel it is essential to first understand travel behavior.

Many different aspects of traveling are of interest for travel behavior analysis. General questions such as "*Why do people travel?*" and "*Where do they go?*" are of great importance for understanding what factors drive the demand and which areas are the most effected by this demand. Other questions are related to a given trip, "*When was the trip made?*", "*What transportation mode was used?*","*Which route was taken?*", and allow to identify, for instance, which infrastructures are used and how the transportation network is effected at different points in time. Of course, all aspects of travel behavior are interrelated which makes its analysis highly complex.

This thesis focuses on the analysis of route choice behavior. More precisely, we are interested in identifying which route a given traveler would take to go from one location to another in a transportation network.

Route choice models can assess travelers' perceptions of various route characteristics such as distance, travel time, cost, number of traffic lights and road types, and relate the results to the individuals' characteristics (e.g. gender, age, income and trip purpose).

Route choice models are also a powerful tool for predicting behavior under different scenarios. Consider for example a project for building a new tunnel. By using the tunnel travelers can save a certain amount of time compared to existing routes but they will be charged a fee for each passage. Route choice models can be used to analyze questions like "*Which is the probability that a 30 year old man with high income chooses the tunnel if the fee is 5 EUR with a potential travel time saving of 10 minutes?*" or "*What share of the travelers would take the tunnel if the travel time savings are approximately 30 minutes and the fee 10 EUR?*".

The aforementioned examples illustrate typical applications of route choice models. Another application requiring advanced route choice modeling is Dynamic Traffic Management Systems. Such systems aim at improving traffic conditions by controlling the supply of the network and by providing real-time information to travelers to help them make better route choice decisions. An important component is the prediction of future traffic conditions so that consistent and unbiased information can be given to the travelers. In order to provide such information it is important to know how travelers react to information and how they adapt their route choices in consequence.

After this general introduction to route choice analysis we give a more detailed overview of the modeling process in the following section.

1.1 Route Choice Modeling Overview

Consider a transport network composed of links and nodes. For a given origin-destination pair and a given transport mode, the route choice problem deals with identifying which route a given traveler would take.

A good overview of this problem can be found in Bovy and Stern (1990) which is the first book entirely dedicated to the topic. The route choice depends, on the one hand, on the attributes of the available routes, such as travel time, type of road, number of traffic lights etc. On the other hand, characteristics and preferences of the traveler also influence the choice. Some travelers like high speeds on freeways while others prefer small scenic roads, some avoid left turns, others traffic lights and so forth.

Several aspects of the route choice problem make it particularly complex. The fact that it deals with paths in networks introduces a combinatorial dimension, it also involves the modeling of choice behavior which is a complex task as such. Moreover, route choice analysis is mainly of interest in dense urban areas and the models therefore need to be operational for large networks.

The efficiency of shortest path algorithms has been a strong motivation for researchers to assume that travelers choose the shortest path, with respect to some link additive generalized cost function. This leads to a simple deterministic route choice model which has the advantage of being operational in large networks. Even though the cost function can be made individual specific, this model cannot directly capture uncertainty related to human behavior and is therefore not realistic. There are indeed several sources of uncertainty in a route choice context. First, travelers have imperfect knowledge of the network and may not be aware of the shortest path. Second, some path attributes, such as perceived travel time, that can be included in the generalized cost function are uncertain. Finally, all individual characteristics and preferences are most likely not available to the modeler and an exact "true" cost function cannot be defined.

The random utility model framework is particularly suitable, and also the most widely used approach, to model choice behavior and related uncertainty. Within this framework, we assume that a traveler n associates a utility U_{in} with each alternative i in his/her choice set $\mathcal{C}_n$. The utility is defined as $U_{in} = V_{in} + \varepsilon_{in}$ and has both a deterministic term V_{in} and a random term ε_{in} capturing uncertainty. The deterministic term can include attributes of the alternative as well as socio-economic characteristics of the traveler. It has in general a linear-in-parameters formulation $V_{in} = \beta x_{in}$ where β is a vector of unknown coefficients to be estimated and x_{in} a vector of attributes. Travelers are assumed to maximize utility and the probability that an alternative i is chosen by traveler n from $\mathcal{C}_n$ is therefore $P(i|\mathcal{C}_n) = P(U_{in} \geq U_{jn} \ \forall \ j \in \mathcal{C}_n) = P(U_{in} = \max_{j\in\mathcal{C}_n} U_{jn})$. Different assumptions on the random terms lead to different types of discrete choice models. For example, the Multinomial Logit model assumes that the random terms are independent and identically distributed Extreme Value, which results in the following probability formulation

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{\mu V_{jn}}},$$

where μ is the positive scale parameter of the Extreme Value distribution. The unknown parameters β can be identified with maximum likelihood estimation. This model was used by Dial (1971) who was one of the first (Burrell, 1968, also uses a random utility model) to address the stochastic dimension of the route choice problem in an operational model.

Despite the fact that discrete choice models are appropriate for route choice analysis, a number of issues need to be addressed. In the following we discuss each of these issues by giving an overview of the modeling process (see Figure 1.1 for a schematic view). Before estimating a route choice model (bottom of the figure) three main steps involving various modeling assumptions need to be performed.

The model is to be estimated based on trips performed by travelers in a real network. Trip observations can be obtained by either asking travelers to describe chosen routes, or by passive monitoring using the Global Positioning System (GPS). In both cases, the data collection is difficult and the descriptions of the chosen routes are often ambiguous. The modeler has a representation of the real network at hand consisting of links and nodes as well as associated attributes (traffic lights, travel times etc.). In order to obtain link-by-link descriptions of chosen routes (denoted path observations in the figure), the first step in the modeling process is to match the trip observations to the network representation used by the modeler.

In a route choice context, the alternatives considered by each traveler are in general unknown to the analyst. It is therefore necessary to generate a choice set $\mathcal{C}_n$ for each path observation, which is far from a trivial task. The number of physically feasible alternatives for a given origin-destination pair is huge, actually unbounded if paths with loops are considered. Path generation algorithms are therefore used to define subsets of alternatives.

Even though the number of alternatives are limited by the choice set generation, it can still be considered large compared to other discrete choice applications. Moreover, the alternatives are highly correlated due to overlapping between the paths (shared links). Models with flexible correlation structure are complex to estimate especially for large number of alternatives. The third and final step before the model estimation involves an appropriate description, or rather approximation, of the correlation among alternatives.

It is worth mentioning that some literature use discrete choice models to analyze specific aspects of route choice behavior. To give two examples:

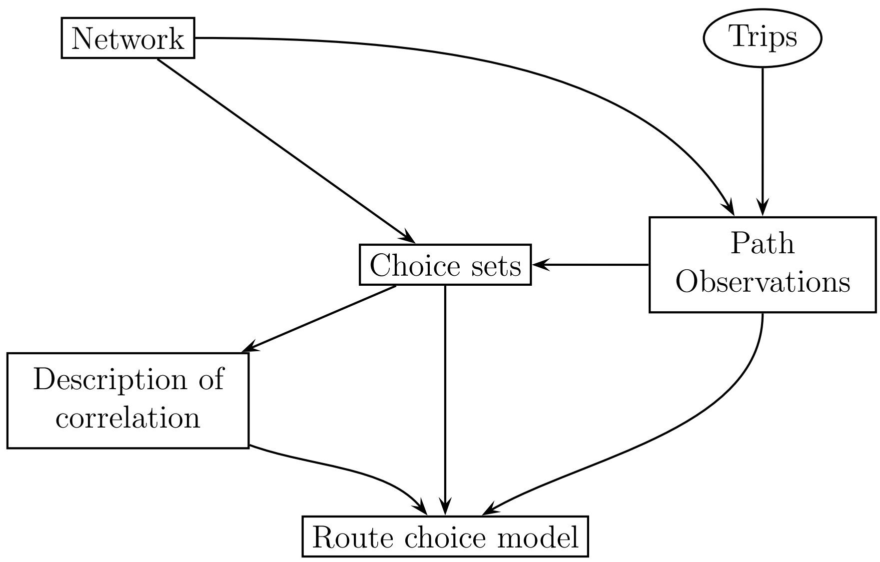

Figure 1.1: Route Choice Modeling Overview

de Palma and Picard (2005) study risk aversion in route choice decisions under travel time uncertainty using stated preferences data and Bogers et al. (to appear) model learning in day-to-day route choice behavior using data collected with a travel simulator.

Finally we note that other frameworks than random utility have been used in the literature for modeling route choice behavior. Several different models based on fuzzy logic have been proposed, see for example Lotan and Koutsopoulos (1993), Lotan (1997), Henn (2000) and Rilett and Park (2001). A review of work using artificial neural networks is given by Dougherty (1995) and Yamamoto et al. (2002) use decision trees for modeling the route choice between two alternatives. This list of literature is not exhaustive but gives some existing alternatives to random utility models.

1.2 Objectives and Scope of the Thesis

This thesis concerns route choice analysis for car trips (uni-modal networks) using discrete choice models and disaggregate revealed preferences data. The latter refers to observations of trips actually performed by travelers in a real network. Note that the same models can be used with stated preferences data (hypothetical trips) if the proposed alternatives correspond to paths in a network. Such a setting is however less complex since the number of alternatives can be limited by design.

The objectives of this thesis can be summarized under the following four keywords.

Data A considerable amount of data processing is in general required to match trip observations to the network used by the modeler. This is not only time consuming but may also introduce bias in the data used for estimation. One objective is to propose an approach that limits the data processing and allows to use original trip descriptions in the route choice model.

Algorithms Many choice set generation algorithms have been proposed in the literature, each generating different choice sets. Furthermore, it is known that estimation results vary depending on the definition of the choice sets. Since the actual alternatives considered by the travelers are unknown to the modeler, it is difficult to analyze which algorithm generates the most accurate choice sets. This research aims at designing route choice set generation algorithms which are both operational and realistic.

Models Path alternatives often share links and are therefore correlated. Discrete choice models with flexible correlation structure are complicated to specify and to estimate. Simple models are therefore used in practice even though the associated assumptions about correlation are violated. A goal of this thesis is to develop a model that captures the correlation among alternatives but that is still suitable for large scale route choice analysis.

Most route choice models assume that the decision is performed pre-trip. Their application in a context where drivers receive real-time information about traffic conditions is questionable. Another objective of this research is to propose route choice models applicable in stochastic and time-dependent networks.

Applications This research concerns fundamental methodological developments for route choice analysis but has the objective of being application oriented. When appropriate, the goal is to use real networks and datasets to illustrate proposed models and algorithms.

1.3 Contributions

In order to clearly show the contributions of this thesis we illustrate them in the previously discussed route choice modeling overview (see Figure 1.2). Brief descriptions of the main contributions are given below.

Network-free Data We propose a modeling approach that makes the explicit matching of trip observations to the network redundant. The dashed lines in Figure 1.2 correspond to the process that is now unnecessary. The original descriptions of trip observations can be directly used for choice set generation and in the route choice model.

Real data, in their original format, rarely correspond to path definitions. This is the case for GPS data as well as trips reported by interviewees. We advocate that the data manipulation, such as map matching, required by the underlying network model to obtain link-by-link descriptions of chosen routes introduces bias and errors and should be avoided. We propose a general modeling scheme that reconciles network-free data (original trip observations) with a network based model without such manipulations. The framework allows for several paths to correspond to a same observation. Fewer assumptions are therefore needed in case of ambiguous trip descriptions. We illustrate the framework with a dataset of reported long distance route choices in Switzerland.

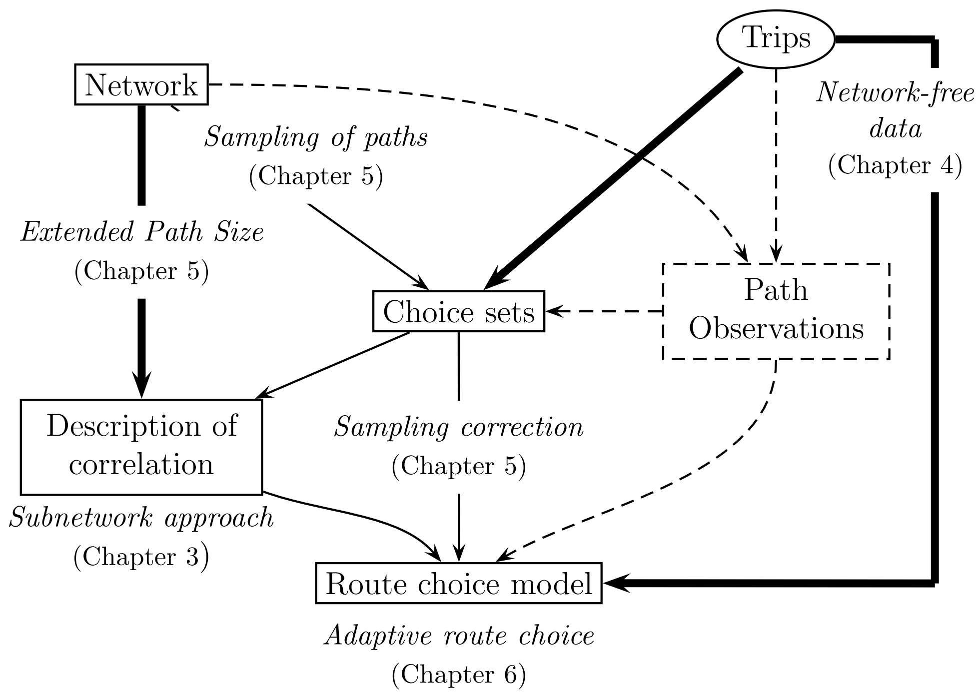

Figure 1.2: Contributions

Sampling of Paths A new paradigm for choice set generation is presented. Existing approaches assume that actual choice sets are found with path generation algorithms. However, none of them is actually able to completely reproduce even the chosen paths. We prefer to assume that the true choice sets contain all paths connecting each origin-destination pair. Although this is behaviorally questionable, we expect this assumption to avoid bias in the econometric model. In this context, we propose a stochastic path generation algorithm that corresponds to an importance sampling approach. The path utilities must then be corrected according to the used sampling protocol in order to obtain unbiased parameter estimates. We derive such a sampling correction for the proposed algorithm. Furthermore, we argue that the description of the correlation should be based on the true choice set and not only on sampled paths. We propose to do this within a Path Size Logit model using an *Extended Path Size* attribute which approximates the true correlation structure.

As shown in Figure 1.2, this novel view on choice set generation has several implications on the traditional route choice modeling process. First, the route choice model should be corrected according to the used path generation algorithm which has earlier been ignored. Second, the description of correlation should not only be based on sampled choice sets but also on the general topology of the network.

Subnetwork Approach We propose an approach for describing correlation that allows the modeler to control the trade-off between the simplicity of the model and the level of realism. Within this framework, the key concept capturing correlation is called a *subnetwork*. The importance and originality of this approach lie in the possibility to capture the most important correlation without considerably increasing the model complexity. This makes it suitable for a wide spectrum of applications, namely involving large-scale networks. We illustrate the model on a GPS dataset collected in the Swedish city of Borlänge.

This approach is an important contribution to the literature where there is a lack of models that can capture correlation in a realistic way and be estimated on data collected in large networks.

Adaptive Route Choice The final contribution of this thesis is based joint work with Moshe Ben-Akiva and Song Gao and concerns adaptive route choice models in stochastic and time-dependent networks. Most existing models assume that travelers make their complete path choice at the origin (static and deterministic network). Such an assumption ignores an important aspect of route choice behavior in real networks. Namely, travelers can adapt their route choices en-route for example in response to real-time information about traffic conditions. We estimate and analyze prediction results (based on synthetic data) of two types of adaptive route choice models: an adaptive path model where a sequence of (non-adaptive) path choice models are applied at intermediate nodes, and a routing policy choice model where alternatives correspond to routing policies (Gao, 2005) rather than paths.

There has been several algorithmic studies of optimal adaptive routing problems presented in the literature but the estimation of such choice models is a new area. This is therefore an important contribution to route choice behavior analysis and to the literature on evaluation of real-time information systems.

In Figure 1.2 this contribution is illustrated on the route choice model level because the routing policy choice model affects all the previous modeling steps. Ideally, even the trip observations are effected and should, in addition to descriptions of chosen routes, contain information about travelers' information access.

Applications Two different types of real data have been used in this thesis. A GPS dataset collected in Sweden and reported long distance trips from a survey in Switzerland. The Swiss network is to our knowledge the largest one used in the literature on route choice analysis based on revealed preferences data. Given the difficulties of making route choice models operational for large networks, illustrating the methodology on real data is a contribution as such.

To summarize, we have addressed several issues in route choice analysis related to *data* (network-free data approach), *models* (Subnetwork approach and adaptive route choice models) and *algorithms* (sampling of paths). Moreover, we have used real *applications* (Borlänge and Switzerland datasets).

1.4 Thesis Outline

This thesis is structured on four papers and the chapters are given in the their chronological order. The outline of the thesis is presented in the following and for each chapter we give the reference to the publication on which it is based.

- Chapter 2 reviews the literature. We focus on analyzing the state of the art rather than giving full technical details on all models and algorithms.

- Chapter 3 deals with correlation in route choice models. We make an in-depth analysis of the Path Size Logit model and present the Subnetwork approach. This chapter has been published as:

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

  Ranked 13 in the top 25 hottest articles of Transportation Research B for July-September 2007.

- Chapter 4 presents the framework for estimating existing route choice models with network-free data. This chapter has been published as:

  Bierlaire, M., and Frejinger, E. (to appear). Route choice modeling with network-free data, Transportation Research Part C: Emerging Technologies, Accepted for publication on July 23, 2007, doi:10.1016/j.trc.2007.07.007.

- Chapter 5 focuses on choice set generation and describes the importance sampling approach. This chapter has been published as:

  Frejinger, E. (2007). Random Sampling of Alternatives in a Route Choice Context. Proceedings of the European Transport Conference (ETC), Leiden, The Netherlands.

  Received the Neil Mansfield Award by the Association for European Transport for the best paper by a sole author aged 35 or under.

- Chapter 6 presents a joint work with Moshe Ben-Akiva and Song Gao that deals with adaptive route choice models in stochastic and time-dependent networks. This chapter has been published as:

  Gao, S., Frejinger, E., and Ben-Akiva, M. (2007). Adaptive Route Choice Models in Stochastic Time-Dependent Networks. Technical report TRANSP-OR 070730. Transport and Mobility Laboratory, ENAC, EPFL.

  Accepted for presentation at the 87th Annual Meeting of the Transportation Research Board, January 2008, and currently under review for possible publication in the Transportation Research Records.

- Chapter 7 provides conclusions and future research perspectives.

Chapter 2

Literature Review

In this chapter we present the state of the art in route choice analysis based on disaggregate data using discrete choice models. It is assumed that the reader is familiar with discrete choice modeling. We refer to Bierlaire (1998) for a concise and comprehensive introduction and to Ben-Akiva and Lerman (1985) which is an excellent textbook on the discrete choice analysis. Train (2003) is a recent textbook covering more advanced topics and focuses on models that require estimation by simulation.

In route choice analysis we are concerned with identifying which route a traveler would select in a transportation network (an introduction is given in Section 1.1). Existing models are based on a static network setting. That is, travelers are assumed to choose a path at the origin and follow it to the destination without considering changes in traffic conditions. In Sections 2.1 to 2.4, we present a literature review of data collection, choice set generation and models for route choice analysis in a static network setting. In the last section we change perspective and present literature related to adaptive route choice in stochastic and time-dependent networks. We introduce concepts that are later used in Chapter 6.

2.1 Route Choice Data

Rather few studies of route choice behavior in uni-modal networks using revealed preferences data are available in the literature. One of the reasons is that data are difficult to collect. Route choice models are in general based on link-by-link descriptions of observed routes. Such data can be collected by either asking travelers to describe chosen routes or by passive monitoring using the Global Positioning System (GPS). Each of these data collection methods have issues. In this section we review some route choice modeling applications and discuss how data related issues have been dealt with.

Mail, telephone and more recently web-based surveys are conventional methods for collecting trip data. Travelers are asked to describe chosen routes and give related information. Various collection methods are proposed in the literature, see for example Mahmassani et al. (1993) and Abdel-Aty et al. (1995). One of the first applications of route choice modeling is presented in Ben-Akiva et al. (1984). They use data collected in 1979 between Utrecht and Amersfoort in the Netherlands (1515 observations). Cars were stopped at the road side or license plates were recorded and surveys were either handed to the driver or mailed to the owner of the car. In order to simplify the choice context, the route choice is not directly modeled but rather the choice among a given set of "labels".

Ramming (2001) presents data collected by asking travelers to describe a chosen path with a set of route segments. This lead to a number of incomplete path descriptions. He uses the shortest path between two known points in order to obtain connected paths. The final sample size is 159 observations.

Prato (2004) uses data collected with a web-based survey for the city of Turin. Respondents were asked to indicate their route choice on an interactive map of the city center. 236 drivers reported 575 routes (one chosen route and alternatives to it). Incomplete trip descriptions are ignored which resulted in a sample of 276 observations.

Vrtic et al. (2006) present route choice data collected in Switzerland. They performed telephone interviews where intermediate locations of long distance trips were reported. This data (940 observations) are used to illustrate the modeling approach with network-free data described in Chapter 4.

In the past decade several studies presented in the literature (e.g. Murakami and Wagner, 1999, and Jan et al., 2000) compare data obtained with conventional survey methods with GPS data. There is a consensus that passive monitoring has several advantages over conventional surveys. For instance, multiple days of trip data can be collected automatically and are directly available in electronic format. However, GPS data also have issues (see Wolf et al., 1999, and Zito et al., 1995, for detailed discussions).

First, constraints of the technology, such as satellite clock errors, receiver noise errors and type of receiver limit the accuracy of the data.

Second, depending on the number of available satellites, atmospheric conditions, and local environment (high buildings, bridges, tunnels) the GPS receiver can compute an inaccurate position or fail to compute the position which introduces gaps in the data. Wolf et al. (1999) state that an accuracy level of 10 meters is required in order to map match GPS points in urban areas without ambiguity. In their tests using data collected in Atlanta, the best performing receiver achieves this level for 63% of the GPS points on average. Note that this rather poor performance is partly due to the effect of selective availability which is not an issue any more. However, Nielsen (2004) observes that 90% of the trips collected in the Copenhagen region had missing data.

A third issue is that the data are stored in one stream of GPS points and data processing is required in order to reconstruct the trips. Such data processing involves map matching, trip end identification and assumptions of missing data. Recently, Marchal et al. (2005) propose a map matching algorithm for large choice sets. They evaluate the performance in terms of computation time and underline the difficulty of evaluating accuracy since the actual chosen routes are unknown (see Quddus et al., 2003, for an overview of map matching algorithms). Du and Aultman-Hall (2007) discuss trip end identification algorithms. They manually identified trip ends in a GPS data stream and evaluate the performance of the algorithms. They find that the best algorithm correctly identified 94% of the trip ends.

Finally, we note that the data processing is highly dependent on the accuracy of the geographical information system data base that is used.

Despite of the aforementioned issues, GPS data have been used for route choice analysis. Nielsen (2004) studies route choice behavior and responses to road pricing schemes based on a large GPS dataset (100 thousand observations) collected in Copenhagen. The author underlines the problems associated with missing data and technical problems. The used map matching approach is described in Nielsen and Joergensen (2004).

In Chapter 3 we estimate route choice models based on a GPS dataset collected in the Swedish city of Borlänge (see Schönfelder et al., 2002, Axhausen et al., 2003, and Frejinger, 2004, for details on the data). The data processing was performed by the Atlanta based company GeoStats. Due to data accuracy issues, observed routes could only be reconstructed for a subset (24 vehicles, 2978 observations) of the complete sample of 186 vehicles.

2.2 Choice Set Generation

We start by giving an overview of the choice set modeling process in Figure 2.1. In a real network a very large number of paths (actually infinitely many if the network contains loops) connect an origin s_o and destination s_d. This set, referred to as the universal choice set $\mathcal{U}$, cannot be explicitly generated. In order to estimate a route choice model a subset of paths needs to be defined and path generation algorithms are used for this purpose. There exist deterministic and stochastic approaches for generating paths. The former refers to algorithms always generating the same set $\mathcal{M}$ of paths for a

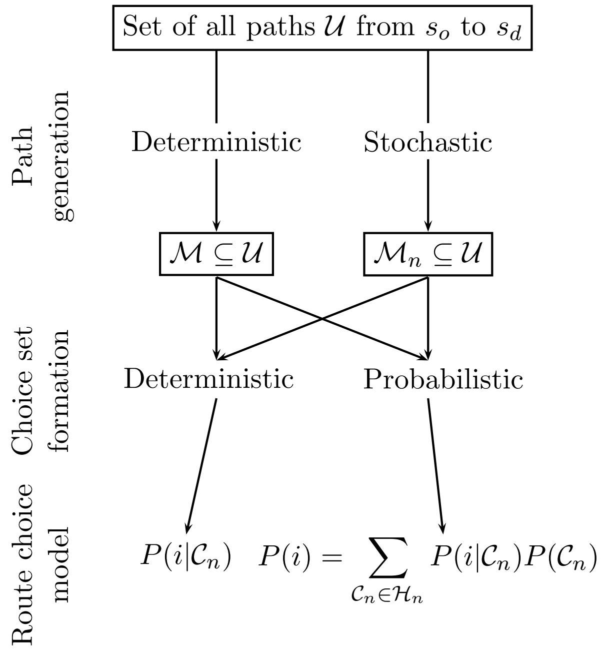

Figure 2.1: Choice Set Generation Overview

given origin-destination pair, whereas an individual (or observation) specific subset $\mathcal{M}_n$ is generated with stochastic approaches. A choice set $\mathcal{C}_n$ for individual n can be defined based on $\mathcal{M}$ (or $\mathcal{M}_n$) in either a deterministic way by including all feasible paths, $\mathcal{C}_n = \mathcal{M}$ (or $\mathcal{C}_n = \mathcal{M}_n$), or by using a probabilistic model $P(\mathcal{C}_n)$ where all non-empty subsets $\mathcal{H}_n$ of $\mathcal{M}$ (or $\mathcal{M}_n$) are considered. $P(i|\mathcal{C}_n)$ is the probability of route i given $\mathcal{C}_n$. Defining choice sets in a probabilistic way is complex due to the size of $\mathcal{H}_n$ and has never been used in a real size application. See Manski (1977), Swait and Ben-Akiva (1987), Ben-Akiva and Boccara (1995) and Morikawa (1996) for more details on probabilistic choice set models. Cascetta and Papola (2001) (Cascetta et al., 2002) propose to simplify the complex probabilistic choice set models by viewing the choice set as a fuzzy set in a implicit availability/perception of alternatives model.

In the following two sections we give brief overviews of existing deterministic and stochastic path generation algorithms focusing on the approaches used in subsequent chapters. For recent and more detailed overviews the reader is referred to Fiorenzo-Catalano (2007) and Bovy (2007b).

2.2.1 Deterministic Approaches

The majority of existing path generation algorithms are deterministic approaches and most of them are based on some form of repeated shortest path search. This type of approach is computationally appealing thanks to the efficiency of shortest path algorithms.

Azevedo et al. (1993) propose the *link elimination* approach that consists in computing the shortest path with respect to some generalized cost function and add it to the choice set. Each link or some links belonging to the shortest path are then removed and a new shortest path in the modified network is computed and introduced in the choice set.

Instead of eliminating links, de la Barra et al. (1993) propose to increase the generalized cost on links in the shortest path and then compute a shortest path for the new cost structure. On the one hand, this *link penalty* approach allows for essential links (e.g. bridges) to be used and a connected network is guaranteed. On the other hand, a same path can be generated repeatedly depending on how the cost structure is updated. Ramming (2001) concludes that the computational time is prohibitively large and disregards it for further consideration in his work.

The above mentioned algorithms may generate paths which are very similar to each other. This is the motivation to use a *constrained k-shortest paths* approach which is another variant of repeated shortest path search. Recent work is presented by Van der Zijpp and Fiorenzo-Catalano (2005) who propose an algorithm to efficiently identify k-shortest paths with respect to detour and overlap constraints.

Ben-Akiva et al. (1984) propose a *labeling* approach that includes in the choice set paths meeting specific criteria such as fastest, shortest or most scenic paths. Shortest paths are therefore repeatedly computed based on different generalized cost functions.

Instead of performing repeated shortest path searches, a constrained enumeration approach referred to as *branch-and-bound* has recently been proposed. Friedrich et al. (2001) present an algorithm for public transport networks, Hoogendoorn-Lanser (2005) for multi-modal networks and Prato and Bekhor (2006) for route networks. These algorithms build a tree where each branch correspond to a path and generate all paths satisfying some constraints. Prato and Bekhor (2006) use directional, temporal, detour, similarity and movement constraints. This type of approach does not benefit from the efficiency of shortest path algorithms and it is therefore crucial to define the constraints such that the size of the tree is limited. This is particularly important for route networks since the number of links in generated paths may be large and this defines the depth of the tree. As stated by Prato and

Bekhor (2006) the speed of the algorithm depends linearly on the width of the tree (number of paths) but exponentially on its depth.

2.2.2 Stochastic Approaches

Most of the deterministic approaches can be made stochastic by using random generalized cost for the shortest path computations. In the following we present two algorithms which are stochastic in their original version. This type of approaches is of particular interest in Chapter 5 where we view path generation as random sampling of alternatives.

Ramming (2001) proposes a *simulation* method that produces alternative paths by drawing link costs from different probability distributions. The shortest path according to the randomly distributed generalized cost is calculated and introduced in the choice set.

Recently, Bovy and Fiorenzo-Catalano (2006) proposed the so-called *doubly stochastic* choice set generation approach. It is similar to the simulation method but the generalized cost functions are specified like utilities and both the parameters and the attributes are stochastic. They also propose to use a filtering process such that, among the generated paths, only those satisfying some constraints are kept in the choice set.

2.2.3 Evaluation of Generated Choice Sets

The evaluation of generated choice sets is difficult since the actual choice sets in general are unknown to the modeler. The following measures, proposed by Ramming (2001) (see also Bekhor et al., 2006), are often used:

- computational time,
- number of routes in the choice set,
- number of links in the choice set and
- coverage of the observed routes (called prediction success rate by Bovy and Fiorenzo-Catalano, 2006, and Bovy, 2007b).

In addition, Prato and Bekhor (2006) and Bekhor and Prato (2006) use estimation results and "reduced choice sets". The latter consists in estimating models based on subsets of the generated choice sets in order to evaluate the influence of number of paths in the choice set.

Fiorenzo-Catalano (2007) focuses on choice set generation for prediction and provides definitions of *reasonable* routes (note that this definition is different from the one proposed by Dial, 1971) and of *adequate* choice sets.

These definitions are based on different criteria similar to the ones used in constrained enumeration algorithms and filtering processes. For example, detour and acyclic criteria at route level and overlap, size, comparability criteria at choice set level.

It is important to underline the limitations of the quality measures proposed in the literature. The coverage indicates if the observed routes have been generated. If there is one observation per origin-destination pair, which is often the case, the corresponding choice sets can contain only one path and still have perfect coverage. The measure is however useful to have an indication on the reasonableness of the generated paths.

Is is not not straightforward to statistically compare estimation results (model fit measures and parameter estimates) for different choice sets. If two estimations of the same model using the same dataset are made but where the definition of the choice sets differ, the results cannot be compared with classic tests such as the likelihood ratio test. Indeed the choice contexts are different and hence so are the null log likelihood values.

The number of routes and links in the choice sets, overlap measures etc. are useful for interpreting the estimation results. However, there exist no reference on "ideal" values for these measures. For this reason it is also difficult to define threshold values for constraints used in filtering processes and constrained enumeration algorithms. The choice of such threshold values must be based on the modeler's intuition and knowledge of the problem taking computation time and memory constraints into account.

For in-depth comparisons of the algorithms described in the previous sections with respect to these measures we refer the reader to Ramming (2001), Hoogendoorn-Lanser (2005), Bovy and Fiorenzo-Catalano (2006), Prato and Bekhor (2006), Bekhor and Prato (2006), Fiorenzo-Catalano (2007) and Bovy (2007b).

Finally we note that Van Nes et al. (2006) discuss generated versus observed choice sets (alternatives reported by travelers) using multi-modal route choice data (Hoogendoorn-Lanser, 2005). On average only 2.8 non-chosen routes are reported. Clearly, using only reported routes does not provide enough variability for model estimation. The authors underline the difficulty of collecting data on alternatives considered by the travelers and they recommend to use generated choice sets for both estimation and prediction.

2.3 Route Choice Models

In this section we give an overview of route choice models in general and those used in subsequent chapters in particular. We focus on models which have been used in real applications and discuss advantages and drawbacks of each model as opposed to giving all technical details of the underlying discrete choice models. Other reviews of route choice models are available in the literature. Ben-Akiva and Bierlaire (2003) give a concise description of discrete choice methods related to route choice applications and Prashker and Bekhor (2004) focus on route choice models used in the stochastic user equilibrium problem.

2.3.1 Multinomial Logit

The Multinomial Logit (MNL) model is simple but restricted the assumption that the error terms are identically and independently distributed (i.i.d.) which does not hold in the context of route choice due to overlapping paths. Due to its simplicity it is one of the most commonly used models in practice. Efforts have therefore been made to overcome this restriction by making a deterministic correction of the utility for overlapping paths. Cascetta et al. (1996) were the first to propose such a deterministic correction. They include an attribute, called Commonality Factor (CF), in the deterministic part of the utility obtaining the C-Logit model. The CF value of one path is proportional to the overlap with other paths in the choice set. Cascetta et al. (1996) present three different formulations of the CF attribute. They do however not provide any guidance about which of the formulations to use.

The lack of theoretical guidance for the C-Logit model was the motivation for Ben-Akiva and Ramming (1998) and Ben-Akiva and Bierlaire (1999a) to propose the Path Size Logit (PSL) model. The idea is similar to the C-Logit model. A correction of the utility for overlapping paths is obtained by adding an attribute to the deterministic part of the utility. In this case, the Path Size (PS) attribute. The original PS formulation is derived from discrete choice theory for aggregate alternatives (Ben-Akiva and Lerman, 1985). The utility associated with path i by individual n is $U_{in} = V_{in} + \beta_{\mathrm{PS}} \ln \mathrm{PS}_{in} + \varepsilon_{in}$ where V_{in} is the deterministic part of the utility and ε_{in} is the random part. The PS attribute is defined as

$$\mathrm{PS}_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum\limits_{j \in \mathcal{C}_n} \delta_{aj}}, \tag{2.1}$$

where Γ_i is the set of all links of path i, L_a is the length of link a and L_i is the length of path i. δ_{aj} equals 1 if link a is on path j and 0 otherwise. $\sum_{j \in \mathcal{C}_n} \delta_{aj}$ is therefore the number of paths in choice set $\mathcal{C}_n$ sharing link a.

Ben-Akiva and Bierlaire (1999b) present another version of this formulation including the length of the shortest path in the choice set, $L^*_{\mathcal{C}_n}$,

$$\mathrm{PS}_{in} = \sum_{a\in\Gamma_i} \frac{L_a}{L_i} \frac{1}{\displaystyle\sum_{j\in\mathcal{C}_n} \frac{L^*_{\mathcal{C}_n}}{L_j}\delta_{aj}}. \tag{2.2}$$

Ramming (2001) introduces a third PS formulation, called Generalized PS

$$\mathrm{PS}_{in} = \sum_{a\in\Gamma_i} \frac{L_a}{L_i} \frac{1}{\displaystyle\sum_{j\in\mathcal{C}_n} \left(\frac{L_i}{L_j}\right)^{\varphi}\delta_{aj}}, \tag{2.3}$$

where φ is a parameter greater or equal to zero. Note that when $\varphi = 0$ the formulation corresponds to the original PS formulation (2.1). Ramming (2001) proposes this formulation in order to decrease the impact of unrealistically long paths in the choice set. In the original PS formulation (2.1) the contribution of a link is decreased by the number of paths that share the link. If there are very long paths that no traveler is likely to choose sharing a link, then these long paths have a negative impact on the utility of shorter, more reasonable paths.

Ramming (2001) compares estimation results of the C-Logit and PSL models with the different formulations but does not provide a theoretical comparison. He finds that the PSL model with the Generalized PS formulation (2.3) outperforms the C-Logit model.

Hoogendoorn-Lanser et al. (2005) (see also Hoogendoorn-Lanser, 2005) study how to define overlap in multi-modal networks. Based on the conclusions of Ramming (2001), they do not further analyze the C-Logit model but focus on the PSL model. They investigate if the β_{PS} parameter should be estimated or be set to one, and conclude that it should be estimated since the PS attribute can capture behavioral perceptions regarding overlapping paths. Moreover, they compare different PS formulations in terms of model fit measures and find that the generalized formulation (2.3) with $\varphi = 14$ shows best results. They also observe best model fit when overlap is expressed in terms of number of legs[1] compared to time or distance.

Motivated by the derivation presented in Frejinger and Bierlaire (2007) (detailed in Chapter 3), Bovy (2007a) propose an alternative derivation which results in the so-called Path Size Correction (PSC) factor

$$\mathrm{PSC}_{in} = \sum_{a\in\Gamma_i} \frac{L_a}{L_i} \ln\left(\frac{1}{\sum_{j\in\mathcal{C}_n}\delta_{aj}}\right). \tag{2.4}$$

[1]A leg is a part of a route between two nodes in which a single mode or service type is used.

The utility is then specified as $U_{in} = V_{in} + \beta_{\text{PSC}}\text{PSC}_{in} + \varepsilon_{in}$. The difference between this and the original formulation is the way the logarithm enters the equation. The PSC is theoretically more appealing than the original PS but estimation results presented by Bovy et al. (2008) do not indicate significant differences between the two formulations. We discuss the theoretical differences further when deriving the original formulation in Section 3.1.

Given the shortcomings of the MNL model, more complex models have been proposed in the literature to explicitly capture path overlap within the error structure. However, rather few of these models have been applied to real size networks and large choice sets.

2.3.2 Multinomial Probit

The error terms are distributed Normal in a Multinomial Probit (MNP) model (see for example Burrell, 1968, and Daganzo, 1977) which permits an arbitrary covariance structure specification. It is well adapted for application and simulation when utilities are link additive. However, it does not have a closed form, and its evaluation requires a great deal of computing time. Consequently, it is rarely adequate for real applications.

Yai et al. (1997) propose a MNP model with structured covariance matrix in the context of route choice in the Tokyo rail network. This considerably limits the number of covariance parameters to be estimated. They still use maximum three alternatives in their application.

An efficient estimation method for MNP is proposed by Bolduc (1999) where he estimates a model with 9 alternatives. However, the choice set sizes in real route choice applications are often considerably larger.

2.3.3 Multivariate Extreme Value

The Multivariate Extreme Value (MEV), also called Generalized Extreme Value (GEV), is a family of models proposed by McFadden (1978) and includes for example the MNL and Nested Logit models. Contrary to the MNL model, the MEV model allows for some correlation and has the advantage of having a closed form.

Vovsha and Bekhor (1998) propose the Link-Nested Logit (LNL) model, which is a Cross-Nested Logit (CNL) formulation where each link of the network corresponds to a nest, and each path to an alternative. This model allows for a rich correlation structure but due to a high number of nests the nesting parameters cannot be estimated. Vovsha and Bekhor (1998) therefore propose to use the network topology (lengths of links and paths) to approximate the nesting parameters. Ramming (2001) estimated the LNL model on route choice data collected in Boston. He concludes that the PSL model with the generalized formulation (2.3) outperforms the LNL model. It is possible that the poor performance of the model is due to the approximation of the nesting parameters. Indeed, the imposed correlation structure may not reflect the actual structure.

Abbé et al. (2007) analyze the CNL model and derive the exact correlation structure. The nesting parameters can be computed by solving a system of equations involving numerical integration. This has not yet been tested for a real size route choice application and is an interesting topic for future research.

Two other MEV route choice models have been proposed in the literature. The Paired Combinatorial Logit model, developed by Chu (1989), has been adapted to the route choice problem by Prashker and Bekhor (1998) and Gliebe et al. (1999). Similar to the LNL model, they propose two different ways to approximate the nesting parameters based on the network topology. Recently, the Link-Based Path-Multilevel Logit model has specifically been developed for the route choice problem by Marzano and Papola (2004). In the same way as the MNL based approach proposed by Dial (1971), it allows for stochastic network loading with "implicit path enumeration" (all efficient paths, i.e. paths that do not backtrack). Both model types are only illustrated on toy networks where no estimation is performed.

2.3.4 Error Component

An Error Component (EC) model is a Normal Mixture of MNL (MMNL) model and was introduced namely by Bolduc and Ben-Akiva (1991) and is designed to be a compromise between the MNL and MNP models. The utilities have Normal as well as Extreme Value distributed error terms. A flexible correlation structure can therefore be defined while keeping the form of a MNL model. The estimation is simpler than for MNP but simulated maximum likelihood estimation is required.

The EC model can be combined with a factor analytic specification where some structure is explicitly specified in the model to decrease its complexity (Ben-Akiva and Bolduc, 1996). Bekhor et al. (2002) (see also Ramming, 2001) estimate an EC model based on route choice data collected in Boston. The utility vector $\mathbf{U}_n$ ($J_n \times 1$, where J_n is the number of paths in $\mathcal{C}_n$) is defined by

$$\mathbf{U}_n = \mathbf{V}_n + \varepsilon_n = \mathbf{V}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n, \tag{2.5}$$

where $\mathbf{V}_n$ ($J_n \times 1$) is the vector of deterministic utilities, $\mathbf{F}_n$ ($J_n \times M_n$) is the link-path incidence matrix (M_n is the number of links in $\mathcal{C}_n$), $\mathbf{T}$ ($M_n \times M_n$) is the link factors variance matrix, and ζ_n ($M_n \times 1$) is the vector of i.i.d. Normal variables with zero mean and unit variance. Bekhor et al. (2002) assume that link-specific factors are i.i.d. Normal and that variance is proportional to link length so that $\mathbf{T} = \sigma \operatorname{diag}\left(\sqrt{l_1}, \sqrt{l_2}, \ldots, \sqrt{l_{M_n}}\right)$ where σ is the only parameter to be estimated. The covariance matrix can then be defined as follows

$$\mathbf{F}_n \mathbf{T}\mathbf{T}^T \mathbf{F}_n^T = \sigma^2 \begin{bmatrix} L_1 & L_{12} & \ldots & L_{1J_n} \\ L_{12} & L_2 & \ldots & L_{2J_n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{1J_n} & L_{2J_n} & \ldots & L_{J_n} \end{bmatrix}$$

where L_{ij} is the length by which path i overlaps with path j.

MMNL models have been used in several studies on real size networks with stated preferences data. The size of the choice set is then limited. Han (2001) (see also Han et al., 2001) use a MMNL model to investigate taste heterogeneity across drivers and the possible correlation between repeated choices. Paag et al. (2002) and Nielsen et al. (2002) use a MMNL model with both random coefficient and error component structure to estimate route choice models for the harbor tunnel project in Copenhagen.

2.4 Sampling of Alternatives

The MNL model can be consistently estimated on a subset of alternatives (McFadden, 1978) using classical conditional maximum likelihood estimation. The probability that an individual n chooses an alternative i is then conditional on the choice set $\mathcal{C}_n$ *defined by the modeler*. This conditional probability is

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in} + \ln q(\mathcal{C}_n|i)}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn} + \ln q(\mathcal{C}_n|j)}} \tag{2.6}$$

and includes an alternative specific term, $\ln q(\mathcal{C}_n|j)$, correcting for sampling bias (μ is a scale parameter). This correction term is based on the probability of sampling $\mathcal{C}_n$ given that j is the chosen alternative, $q(\mathcal{C}_n|j)$. See for example Ben-Akiva and Lerman (1985) for a more detailed discussion on sampling of alternatives. Bierlaire et al. (to appear) have recently shown that MEV models can also be consistently estimated and propose a new estimator.

If all alternatives have equal selection probabilities, the estimation on the subset is done in the same way as the estimation on the full set of alternatives. Namely, $q(\mathcal{C}_n|i)$ is equal to $q(\mathcal{C}_n|j)\ \forall\ j \in \mathcal{C}_n$ and the corrections for sampling bias cancel out in Equation (2.6). A simple random sampling protocol is however not efficient if the full set of alternatives is very large. Indeed, the sample should include attractive alternatives since comparing a chosen alternative to a set of highly unattractive alternatives would not provide much information on the choice. In order to ensure that attractive alternatives are included, the sample would need to be prohibitively large.

When using a sampling protocol selecting attractive alternatives with higher probability than unattractive alternatives (importance sampling), the correction terms in Equation (2.6) do not cancel out. Note however that if a full set of alternative specific constants are estimated, all parameter estimates except the constants would be unbiased even if the correction is not included in the utilities (Manski and Lerman, 1977).

Importance sampling of alternatives has been used in the literature. For example, Ben-Akiva and Watanatada (1981) use samples of destinations for prediction and Train et al. (1987) sample alternatives for the estimation of local telephone service choice models. A sampling of alternatives approach has however never been used in a route choice modeling context, to the best of our knowledge.

2.5 Adaptive Route Choice

Traffic conditions in transportation networks are inherently uncertain due to disturbances such as traffic lights, incidents, vehicle breakdowns, work zones, bad weather conditions, special events and so forth. Travelers receive information about traffic conditions during trips and can adapt their route choices accordingly. A static network setting is assumed in the route choice models presented in Section 2.3 and the dynamic route choice process is therefore neglected. These existing models can in principle be applied successively in a stochastic and time-dependent network to model adaptive route choice behavior. DynaMIT (Ben-Akiva et al., 2001) and DYNASMART (Mahmassani, 2001) are examples of dynamic traffic assignment models that use this kind of approach. Calibration of DynaMIT's route choice model based on aggregate data is reported in Balakrishna (2006) and Balakrishna et al. (2007). We are however unaware of any estimation of sequential route choice models based on disaggregate data.

In a stochastic and time-dependent network setting it is assumed that the link travel times are random variables with time dependent distributions. Since the travel time is random, a traveler entering a link at a given time might exit the link at various times, which in turn can result in different travel time distributions on the downstream links. A path is a purely topological concept and is therefore not appropriate in this setting. Instead, adaptive routing concepts have been proposed, referred to as routing policy (Gao, 2005), hyperpath, strategy (see Marcotte and Nguyen, 1998, for an overview) or online path with recourse (Polychronopoulos and Tsitsiklis, 1996). The literature includes a number of algorithmic studies of optimal adaptive routing problems but the estimation of such choice models is a new area. Ukkusuri and Patil (2006) apply sequential Logit loading of hyperpath flows in an equilibrium traffic assignment model where they assume that travelers learn realized travel times on outgoing links. The estimation problem is however not addressed.

Finally we note that there have been a large number of studies evaluating the potential benefits of providing pre-trip and en-route real-time information to travelers. A recent literature review can be found in Abdel-Aty and Abdalla (2006). We are however unaware of any studies estimating route choice models based on real trip observations, instead, interactive simulation, synthetic or stated preference data are mainly used.

In the following we focus on the concept of routing policy proposed by Gao (2005) (see also Gao and Chabini, 2006) which is used in Chapter 6.

2.5.1 Routing Policy

For this concept, we assume a stochastic and time-dependent network defined as $G = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{P})$ where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ the set of links and $\mathcal{T}$ the set of time periods. The travel time on each link ℓ at each time period t is a random variable $\widetilde{T}_{\ell t}$ with a given probability mass function (PMF). $\mathcal{P}$ is the probabilistic description of the link travel times. The most general definition of $\mathcal{P}$ is a joint probability distribution of all link travel time random variables: $\mathcal{P} = \{\varrho_1, \varrho_2, \ldots, \varrho_r, \ldots, \varrho_R\}$ where ϱ_r is a matrix $(|\mathcal{T}| \times |\mathcal{E}|)$ and R the number of support points. Each support point r has a probability $P(r)$ and $\sum_{r=1}^{R} P(r) = 1$. An element of ϱ_r is denoted $T_{\ell t}^r$ and represents the realized travel time of link ℓ at time t for support point r.

At each node in the network, a traveler can decide which is the next node depending on the current state $\{v, t, I\}$ where v is the current node, t the current time and I the current information. The latter refers to a set of realized link travel times. A routing policy is defined as a mapping from states to decisions (next nodes in the network) and it manifests itself as a path for each support point of the network. A routing policy can therefore be viewed as a collection of paths, each with a certain probability. For more details we refer the reader to Gao (2005) and Gao and Chabini (2006).

Gao (2005) propose a Policy Size Logit model which is similar to the Path Size Logit model (Ben-Akiva and Ramming, 1998, and Ben-Akiva and Bierlaire, 1999b). It is designed to model the probability of a routing policy γ

given an individual specific choice set of routing policies $\mathcal{G}_n$, $P(\gamma|\mathcal{G}_n)$, using a MNL model. The utilities are corrected with the Policy Size (PoS) attribute $U_{\gamma n} = V_{\gamma n} + \beta_{\text{PoS}} \ln \text{PoS}_{\gamma n} + \varepsilon_{\gamma n},\ \forall\ \gamma \in \mathcal{G}_n$, that is defined as

$$\text{PoS}_{\gamma n} = \sum_{r=1}^{R} \left(\sum_{a \in \mathcal{I}_\gamma^r} \left(\frac{T_a^r}{T_\gamma^r} \right) \frac{1}{M_{an}^r} \right) P(r) \tag{2.7}$$

where

- $\mathcal{I}_\gamma^r$ is the set of links of the realized path of routing policy γ for r,
- T_a^r is the realized travel time of link a for r,
- T_γ^r is the realized travel time of routing policy γ for r and
- M_{an}^r the number of routing policies in $\mathcal{G}_n$ using link a for r.

Note that all variables are time-dependent, but the time subscript is omitted to make the notation light. The PoS attribute may be viewed as an "expected" PS. For each support point a routing policy manifests itself as a path. The PS attribute is computed for each support point and the expectation is taken over all possible support points.

Chapter 3

Modeling Correlation in Route Choice Models

When using random utility models for a route choice problem, a critical issue is the significant correlation among alternatives, as discussed in Section 2.3. There are basically two types of models proposed in the literature to address it: (i) a deterministic correction of the path utilities in a Multinomial Logit model (such as the Path Size Logit or the C-Logit models) and (ii) an explicit modeling of the correlation through assumptions about the error terms and the use of advanced discrete choice models such as the Cross-Nested Logit or the Error Component models. The first is simple, easy to handle and often used in practice. Unfortunately, it does not correctly capture the correlation structure, as we discuss in details in the following section. The second is more consistent with the modeling objectives, but complicated to specify and estimate.

The modeling framework presented in this chapter allows the analyst to control the trade-off between the simplicity of the model and the level of realism. Within this framework, the key concept capturing the correlation structure is called a *subnetwork*. A subnetwork is a simplification of the road network only containing easy identifiable and behaviorally relevant parts. In practice, the subnetwork can easily be defined based on the road network hierarchy. The importance and the originality of our approach lie in the possibility to capture the most important correlation without considerably increasing the model complexity. This makes it suitable for a wide spectrum of applications, namely involving realistic large-scale networks.

As an illustration, we present estimation results of a factor analytic specification of a mixture of Multinomial Logit model, where the correlation among paths is captured by error components. The estimation is based on a GPS dataset collected in the Swedish city of Borlänge. The results show a significant increase in model fit and forecasting performance for the Error Component model compared to a Path Size Logit model. Moreover, the correlation parameters are significant.

In the following section we analyze the Path Size Logit model. In Section 3.2 we introduce the new modeling approach based on the concept of subnetworks. Finally, we present estimation and prediction results for real data of Error Component models based on subnetworks and compare the results with Multinomial Logit and Path Size Logit models.

3.1 Deterministic Correction for Correlation

Several formulations of the Path Size (PS) attribute have been proposed in the literature. Here, we show that the original PS formulation (2.1), or the Path Size Correction (PSC) factor (2.4), should be used for correcting utilities of overlapping paths. These are the formulations that both show intuitive results and have a theoretical motivation. We start by showing how and under which assumptions the original PS formulation can be derived from the theory on aggregation of alternatives (Ben-Akiva and Lerman, 1985).

A nested structure is assumed where each nest corresponds to an aggregate alternative grouping elemental alternatives. In a route choice context the elemental alternatives correspond to the paths and the aggregate alternatives to the links. For the derivation of the original PS formulation we are interested in the choice of elemental alternative (route choice) as well as the size of the aggregate alternatives, where the size of an aggregate alternative, a link, equals the number of paths using the link.

We denote by $\mathcal{C}_n$ the set of paths considered by individual n, and we define subsets, $\mathcal{C}_{an} \subseteq \mathcal{C}_n,\ a = 1, \ldots, M_n$, where $\mathcal{C}_{an}$ is the set of paths using link a, and M_n is the number of links in $\mathcal{C}_n$. The utility U_{in} individual n associates with path i is $U_{in} = V_{in} + \varepsilon_{in}$ where V_{in} represents the deterministic part of the utility and ε_{in} the random part. The link utility U_{an} is defined by $U_{an} = \max_{j \in \mathcal{C}_{an}}(V_{jn} + \varepsilon_{jn}),\ a = 1, \ldots, M_n$. U_{an} can also be expressed as the sum of its expectation V_{an} and its random term ε_{an}, that is, $U_{an} = V_{an} + \varepsilon_{an}$ where $V_{an} = E[\max_{j \in \mathcal{C}_{an}}(V_{jn} + \varepsilon_{jn})]$. The average deterministic utility of paths using link a is defined by $\overline{V}_{an} = (1/M_{an}) \sum_{j \in \mathcal{C}_{an}} V_{jn}$ where M_{an} is the number of paths in $\mathcal{C}_n$ using link a (the size of link a). That is, $M_{an} = \sum_{j \in \mathcal{C}_n} \delta_{aj}$, where δ_{aj} is the link-path incidence variable that equals one if link a is on path j and zero otherwise.

According to the theory, if we assume that the size of $\mathcal{C}_{an}$ is large for all links, that the path utilities using a link have equal means and the random terms ε_{in} are independent and identically distributed (i.i.d.), then the utility individual n associates with link a is defined by

$$U_{an} = \overline{V}_{an} + \frac{1}{\mu} \ln M_{an} + \varepsilon_{an},$$

where μ is a positive scale parameter (see Ben-Akiva and Lerman, 1985, p. 256, Eq. (9.7)).

The original PS formulation, correcting the path utility U_{in}, is based on the definition of the link utility U_{an}. Accordingly, the positive correction for the size of an aggregate alternative, results in a negative correction of the utility of an elemental alternative. Moreover, there is no correction of an elemental alternative which belongs to a nest with size one. The size correction for an elemental alternative can therefore be defined as $(1/\mu)\ln(1/M_{an})$. The contribution of a link a is thus

$$\frac{1}{\mu} \ln \frac{1}{\sum\limits_{j \in \mathcal{C}_n} \delta_{aj}}$$

where δ_{aj} is the link-path incidence variable. Furthermore, we assume that the size of a path is proportional to the length of its links. If L_a denotes the length of link a and L_i the length of path i, we have derived the original PS formulation

$$\mathrm{PS}_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum\limits_{j \in \mathcal{C}_n} \delta_{aj}}.$$

Including a PS correction in the utility U_{in} gives

$$U_{in} = V_{in} + \beta_{\mathrm{PS}} \ln \mathrm{PS}_{in} + \varepsilon_{in}, \quad i \in \mathcal{C}_n,$$

where $\beta_{\mathrm{PS}} = 1/\mu$. β_{PS} should therefore be estimated and be strictly positive in order to be consistent with the theory.

In the derivation of the PSC factor Bovy (2007a) defines the contribution of each link in path j as

$$\frac{L_a}{L_j} \frac{1}{\mu_a} \ln \frac{1}{M_{an}}$$

which seems more appropriate from a theoretical perspective than weighting with (L_a/L_j) inside the logarithm as is done in the original PS. The utility is then defined as

$$U_{in} = V_{in} + \beta_{\mathrm{PSC}} \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \ln \frac{1}{\sum_{j \in \mathcal{C}_n} \delta_{aj}} + \varepsilon_{in}$$

where it assumed that $\mu_a = \mu\ \forall a \in \Gamma_i$ and $\beta_{\text{PSC}} = 1/\mu$.

Ben-Akiva and Bierlaire (1999b) do not include a β_{PS} in their utility specification. Ramming (2001) argues that according to discrete choice theory, β_{PS} should be fixed to one. However, his β_{PS} estimate is significantly different from both zero and one. Hoogendoorn-Lanser et al. (2005) suggest that the PS attribute can have a behavioral interpretation and therefore argues that β_{PS} should be estimated. They also get better empirical results when estimating β_{PS}. When deriving the original PS formulation, we show that $\beta_{\text{PS}} = 1/\mu$ where μ is a positive scale parameter.

Both Ramming (2001) and Hoogendoorn-Lanser et al. (2005) conclude that the PS attribute only corrects the utility for a part of the correlation. In the derivation of the PS attribute, the error terms of paths using a same link are assumed to be i.i.d. The cross-nested structure and the correlation due to paths using more than one link is therefore neglected. This explains the PS attribute's limited capacity of capturing correlation.

Ben-Akiva and Bierlaire (1999b) present an alternative PS formulation (2.2) including the length of the shortest path in the choice set $L^*_{\mathcal{C}_n}$. The correlation of the utility $\ln \text{PS}_{in}$ can be written as follows:

$$\ln \text{PS}_{in} = -\ln L_i - \ln L^*_{\mathcal{C}_n} + \ln \sum_{a\in\Gamma_i} \frac{L_a}{\displaystyle\sum_{j\in\mathcal{C}_n} \frac{1}{L_j}\delta_{aj}}.$$

Note that, including $L^*_{\mathcal{C}_n}$ adds a constant $\ln L^*_{\mathcal{C}_n}$ to all path utilities in the choice set which does not change their relative utility. The length component in the denominator, $1/L_j$, does however play a role.

The Generalized PS formulation (2.3) is introduced by Ramming (2001) in order to decrease the influence of unrealistically long paths on the utility of shorter paths in the choice set. The formulation is however difficult to interpret for $\varphi > 0$. (Note that $\varphi = 0$ corresponds to the original PS formulation.)

In order to analyze the influence of the φ parameter, we write $\ln \text{PS}_{in}$ as follows:

$$\ln \text{PS}_{in} = -(\varphi + 1)\ln L_i + \ln \sum_{a\in\Gamma_i} L_a \frac{1}{\displaystyle\sum_{j\in\mathcal{C}_n} \left(\frac{1}{L_j}\right)^{\varphi} \delta_{aj}}. \tag{3.1}$$

Independently of the value of the φ parameter, this formulation yields a zero correction when path i has no overlap with any other path in the choice set. When $\varphi \to +\infty$, if we assume that $L_i > 1 \quad \forall\ i \in \mathcal{C}_n$, the limits of the two

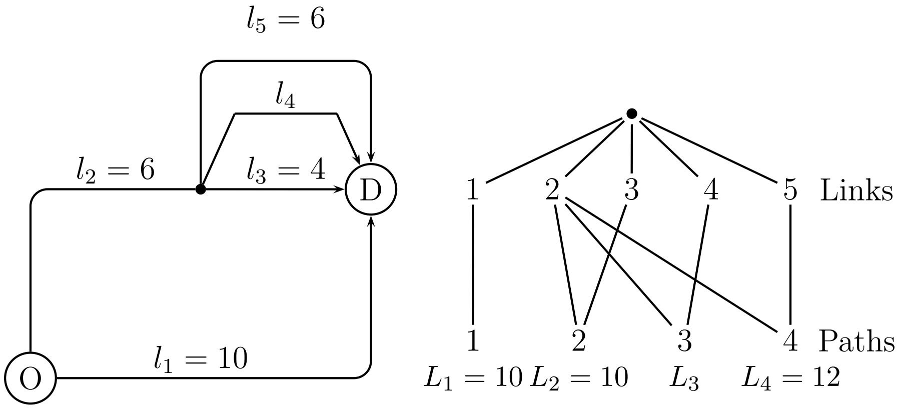

Figure 3.1: Example for Deterministic Correction Formulation

terms in Equation (3.1) are

$$\lim_{\varphi\to+\infty} -(\varphi+1)\ln L_i = -\infty \qquad \lim_{\varphi\to+\infty} \ln \sum_{a\in\Gamma_i} L_a \frac{1}{\displaystyle\sum_{j\in\mathcal{C}_n} \left(\frac{1}{L_j}\right)^{\varphi} \delta_{aj}} = +\infty.$$

This result can be explained by the fact that the sum in the denominator of formulation (2.3) is composed of terms $(L_i/L_j)^{\varphi}$ where (L_i/L_j) can be greater or equal to one, or less than one depending on the lengths L_i and L_j. Since Ramming (2001) considered an example with only two correlated alternatives this effect was not illustrated in his thesis. Here we consider instead an example with three correlated alternatives (shown in Figure 3.1) where the length of path 3, L_3, varies with the length of link 4, l_4.

In Figure 3.2 we compare the values of the original PS formulation (2.1), $\varphi = 0$ (thin lines), with the generalized formulation (2.3) using a high value of φ (thick lines) as a function of l_4. Only the PS values for the correlated alternatives are shown.

The original PS formulation penalizes path 2 the most and path 4 the least. This is intuitive since the correlated part (link 2) has a higher proportion of the total length for path 2 than path 4. Moreover, path 3 is penalized proportionally to the length of link 4, as expected.

For high values of φ the results are problematic. Firstly (and most importantly), we observe a high sensitivity of the PS value with respect to small modifications in the length of the paths. This happens when two paths, including the shortest one, have almost the same length. This is illustrated by values of l_4 close to 4.0 in Figure 3.2. Clearly, this may have a disastrous impact on the probabilities in the presence of tiny errors in the route length

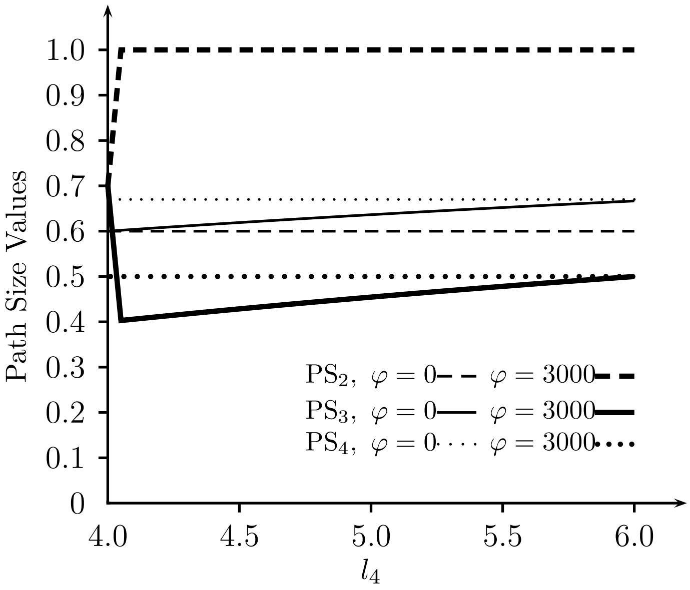

Figure 3.2: PS Values ($\varphi = 0$ and $\varphi = 3000$) for Correlated Alternatives in Example 3.1 as a Function of l_4

data. Secondly, when l_4 increases, we obtain an extreme situation where the shortest path (path 2) is not penalized at all.

We now consider a choice set where two correlated alternatives have almost the same length and one of those alternatives is the shortest path, that is $L_1 = 10.0, L_2 = 10.0, L_3 = 10.1$ and $L_4 = 12$. This case is common in practice. In Figure 3.3 we show the PS values for this case as φ varies. First of all, note that the ordering of the paths changes. Path 4 is more penalized than path 3 for $\varphi < 170$ and then the order is inverted. Second, even though path 3 is only 1% longer than path 2, its PS value decreases as φ increases.

We conclude that the generalized formulation may produce counter intuitive results and the original PS formulation (or the recently proposed PSC factor) should therefore be preferred, with the additional motivation that it has a theoretical foundation. However, as pointed out earlier, the PS attribute can only capture part of the correlation. It is preferable to use a model that accounts explicitly for correlation within the error structure, but without considerably increasing the complexity. For this purpose, we propose to use subnetworks which are discussed in the next section.

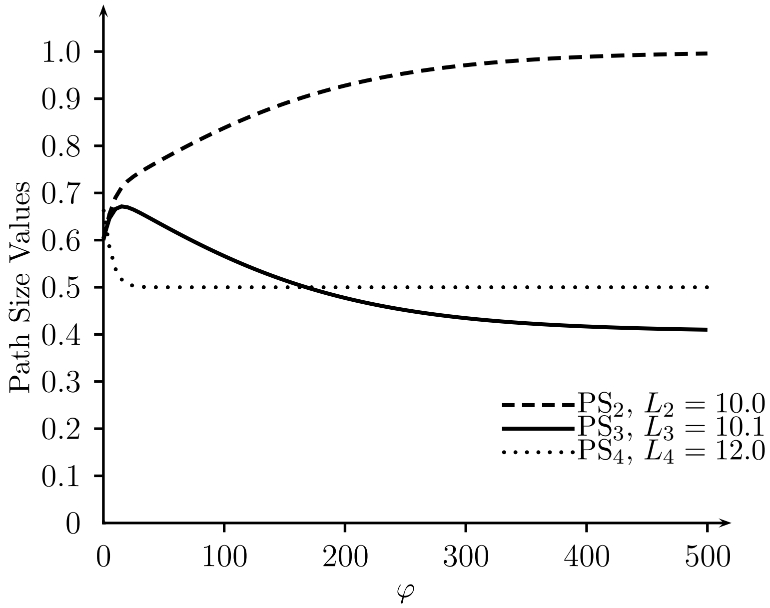

Figure 3.3: PS Values for Correlated Alternatives as a Function of φ

3.2 Subnetworks

We are proposing a modeling approach which is designed to be both behaviorally realistic and convenient for the analyst. We define a *subnetwork component* as a set of links corresponding to a part of the network which can be easily labeled and is behaviorally meaningful in actual route descriptions (Champs-Elysées in Paris, Fifth Avenue in New York, Mass Pike in Boston, etc.). The analyst defines subnetwork components either by arbitrarily selecting, for example, motorways and main roads in the network hierarchy, or by conducting simple interviews to identify the most frequently used names when people describe itineraries. Note that the actual relevance of a given subnetwork component can be tested after model estimation, so that various hypotheses can be tried.

The model is designed such that paths sharing a subnetwork component are correlated. This allows for a great deal of modeling flexibility, including the possibility to capture perceptual correlation among paths that are not physically overlapping. For instance, two paths going through the city center may share unobserved attributes, even if they do not share any link.

We propose to explicitly capture this correlation within a factor analytic specification of an Error Component (EC) model. The model specification is combined with a PS attribute that accounts for the topological correlation on the complete network. The EC model specification is an extension of the model presented by Bekhor et al. (2002). We define the utility as

$$\mathbf{U}_n = \mathbf{V}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n \tag{3.2}$$

where $\mathbf{F}_{n\ (J_n \times Q)}$ is the factor loadings matrix (J_n is the number of paths in choice set $\mathcal{C}_n$ and Q is the number of subnetwork components), $\mathbf{T}_{(Q\times Q)} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_Q)$ (σ_q is the covariance parameter associated with subnetwork component q, to be estimated), $\zeta_{n\ (Q\times 1)}$ is a vector of i.i.d. N(0,1) variates, and $\nu_{n(J_n\times 1)}$ is a vector of i.i.d. Extreme Value distributed variates. An element $(f_n)_{iq}$ of $\mathbf{F}_n$ equals $\sqrt{l_{niq}}$ where l_{niq} is the length by which path i in choice set $\mathcal{C}_n$ overlaps with subnetwork component q.

We illustrate the model specification with a small example presented in Figure 3.4. We consider one origin-destination pair, three paths and a subnetwork composed of two subnetwork components (S_a and S_b). Path 1 uses both subnetwork components whereas path 2 only uses S_a and path 3 only S_b. Path 1 is assumed to be correlated with both path 2 and path 3 even though path 1 and path 2 do not physically overlap. The path utilities for this example are consequently

$$\begin{aligned}
U_1 &= V_1 + \sqrt{l_{1a}}\sigma_a\zeta_a + \sqrt{l_{1b}}\sigma_b\zeta_b + \nu_1 \\
U_2 &= V_2 + \sqrt{l_{2a}}\sigma_a\zeta_a + \nu_2 \\
U_3 &= V_3 + \sqrt{l_{3b}}\sigma_b\zeta_b + \nu_3,
\end{aligned}$$

where ζ_a and ζ_b are distributed N(0,1), l_{iq} is the length path i shares with subnetwork component q. σ_a and σ_b are the covariance parameters to be estimated. The variance-covariance matrix of ζ for this example is

$$\mathbf{FTT}^T\mathbf{F}^T = \begin{bmatrix} l_{1a}\sigma_a^2 + l_{1b}\sigma_b^2 & \sqrt{l_{1a}}\sqrt{l_{2a}}\sigma_a^2 & \sqrt{l_{1b}}\sqrt{l_{3b}}\sigma_b^2 \\ \sqrt{l_{1a}}\sqrt{l_{2a}}\sigma_a^2 & l_{2a}\sigma_a^2 & 0 \\ \sqrt{l_{3b}}\sqrt{l_{1b}}\sigma_b^2 & 0 & l_{3b}\sigma_b^2 \end{bmatrix}.$$

3.2.1 Empirical Results

The estimation results presented in this section are based on a GPS dataset collected during a traffic safety study in the Swedish city of Borlänge. Nearly 200 vehicles were equipped with a GPS device and the vehicles were monitored within a radius of about 25 km around the city center. Since the dataset was not originally collected for route choice analysis, an extensive amount

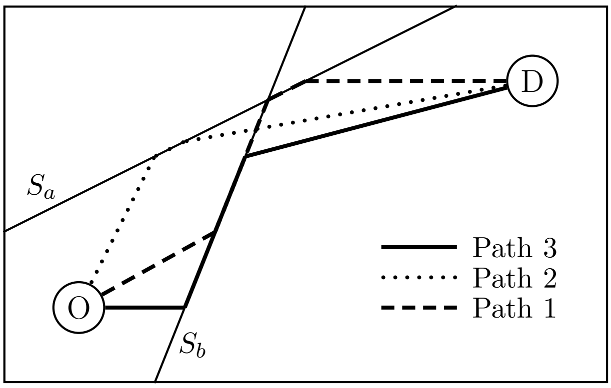

Figure 3.4: Example of a Subnetwork

of data processing has been performed in order to obtain coherent route observations. The data processing was mainly performed by the company GeoStats in Atlanta. Data of 24 vehicles and a total of 16 035 observations are available for route choice analysis (see Axhausen et al., 2003, Schönfelder and Samaga, 2003, and Schönfelder et al., 2002, for more details on the Borlänge GPS data set). For the model estimations we consider a total of 2 978 observations corresponding to 2 244 observed simple routes of 24 vehicles and 2 179 origin-destination pairs. Note that we make a distinction between observations and observed routes since a same route can have been observed several times.

Only individuals who had access to their own vehicle were recruited for the survey. Moreover, we do not have access to the characteristics of the drivers. We therefore assume that each vehicle correspond to one single individual.

Borlänge is situated in the middle of Sweden and has about 47 000 inhabitants. The road network contains 3 077 nodes and 7 459 unidirectional links. We have defined a subnetwork based on the main roads traversing the city center. Two of the Swedish national roads ("riksväg") traverse Borlänge. The subnetwork is composed of these national roads (referred to as R.50 and R.70) and we have defined two subnetwork components for each national road (north and south directions). In addition, we have defined one subnetwork component for the road segment in the city center where R.50 and R.70 overlap (called R.C.). The Borlänge road network and the subnetwork are shown in Figure 3.5. In Table 3.1 we report for each subnetwork component its length and the number of observations that use the component. Table 3.1 also reports the weighted number of observations N_q, defined by $N_q = \sum_{o \in \mathcal{O}} (l_{oq}/L_q)$, where l_{oq} is the common length between the

	R.50 S	R.50 N	R.70 S	R.70 N	R.C.
Component length [m]	5255	4966	11362	7028	1733
Nb. of Observations	173	153	261	366	209
Weighted Nb. of Observations (N_q)	36	88	65	73	116

Table 3.1: Statistics on Observations of Subnetwork Components

route corresponding to observation o and subnetwork component q, L_q is the length of q, and $\mathcal{O}$ is the set of all observations.

A link elimination approach (Azevedo et al., 1993) has been used for the choice set generation. This algorithm computes the shortest path and adds it to the choice set. One link at a time is then removed from the original shortest path, and a new shortest path in the modified network is computed and added to the choice set, if it is not already present.

The main drawback of the link elimination approach is that it generates similar routes. When one link is removed, there exists often a short deviation using roads next to the removed link. In order to address this drawback we have used two generalized cost functions for the shortest path computation. In addition to estimated travel time, we have also used link length divided by the number of lanes. For each origin-destination pair, the link elimination algorithm is therefore applied to two shortest paths.

The observed routes that were not found by the choice set generation algorithm were added afterward. The algorithm found all the observed routes for 80% of the origin-destinations pairs. However, for 20% of the origin-destination pairs, none of the observed routes were identified. Typically, this is the case when the observed routes make long detours compared to the shortest path, for example, in order to avoid the city center. These results are consistent with the findings of Ramming (2001) who at best found 84% of the observed routes by combining all the choice set generation algorithms that he had tested. The number of paths in the choice sets varies between 2 and 43 where a majority of the choice sets (93%) include less than 15 paths.

Model Specifications

We compare MNL and PSL models with five different specifications of an EC model based on the subnetwork defined previously. Two EC models are specified with a simplified correlation structure where the covariance parameters are assumed to be equal (models EC_1 and EC'_1). Two other EC models are specified with one covariance parameter per subnetwork component (models EC_2 and EC'_2).

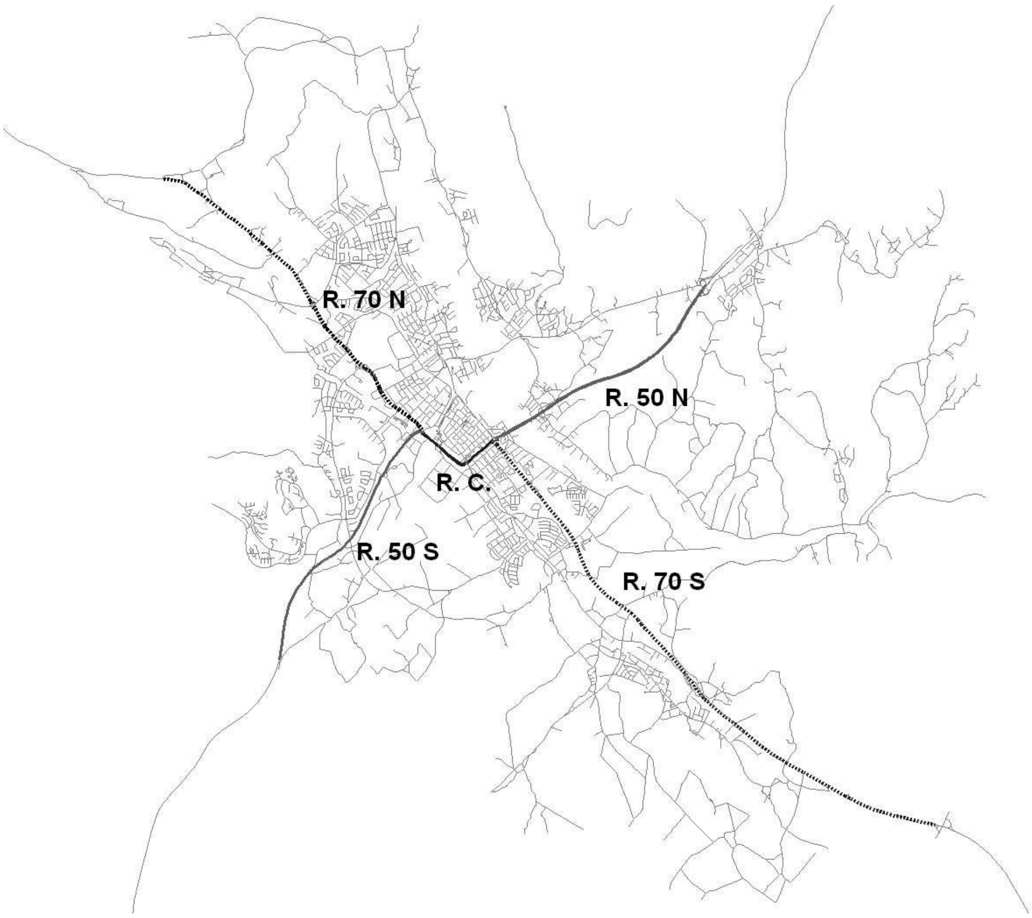

Figure 3.5: Borlänge Road Network and Subnetwork Definition

The PSL, EC_1 and EC_2 models are specified with the same linear in parameters formulation of the deterministic part of the utility function. The deterministic part V_i of the utility for alternative i is

$$\begin{aligned} V_i = \; & \beta_{\mathrm{PS}} \ln(\mathrm{PS}_i) + \beta_{\mathrm{EstimatedTime}} \mathrm{EstimatedTime}_i + \\ & \beta_{\mathrm{NbSpeedBumps}} \mathrm{NbSpeedBumps}_i + \beta_{\mathrm{NbLeftTurns}} \mathrm{NbLeftTurns}_i + \\ & \beta_{\mathrm{AvgLinkLength}} \mathrm{AvgLinkLength}_i . \end{aligned}$$

For comparison we also estimate models without a PS attribute (MNL, EC'_1 and EC'_2) which otherwise have the same deterministic utilities as the other models. In addition to attributes such as estimated travel time, number of speed bumps and number of left turns in uncontrolled crossings, we have included average link length which is intended to capture an attraction for routes with few crossings. The estimated travel time is computed for each link in the network based on its length and an average speed. We have used one average speed for each speed limit that corresponds to the observed average speed. Statistics on all attributes included in the model specifications are given in Table 3.2.

PS formulations based on length and estimated travel time show similar results in our case. We have therefore preferred the definition based on length, which is directly observed. A high correlation among the routes is expected since a link elimination approach has been used for generating the choice sets. In Figure 3.6 we show the PS values for all routes and all choice sets. The generated routes are shown with black bars and the observed routes with gray bars. A majority of the routes have a high overlap (low PS values). Only 5% of the routes have no overlap (PS value that equals 1). Note that almost 50% of the routes that have no overlap are observed routes. This can be explained by the poor performance of the choice set generation algorithm discussed in the previous section. Namely, for 20% of the origin-destination pairs, none of the observed routes were found by the algorithm. These observed routes are therefore expected to have a low overlap with the other routes in the choice set.

We deal with heteroscedasticity by specifying different scale parameters for different individuals. After systematic testing of various specifications, eight individuals have one scale parameter each which are estimated significantly different from one. For the remaining individuals the scale parameter is fixed to one.

Even though the number of individuals is small, we provide a model, EC_3, where we take into account that we have panel data. Otherwise the specification is the same as for EC_2. Our objective is purely illustrative,

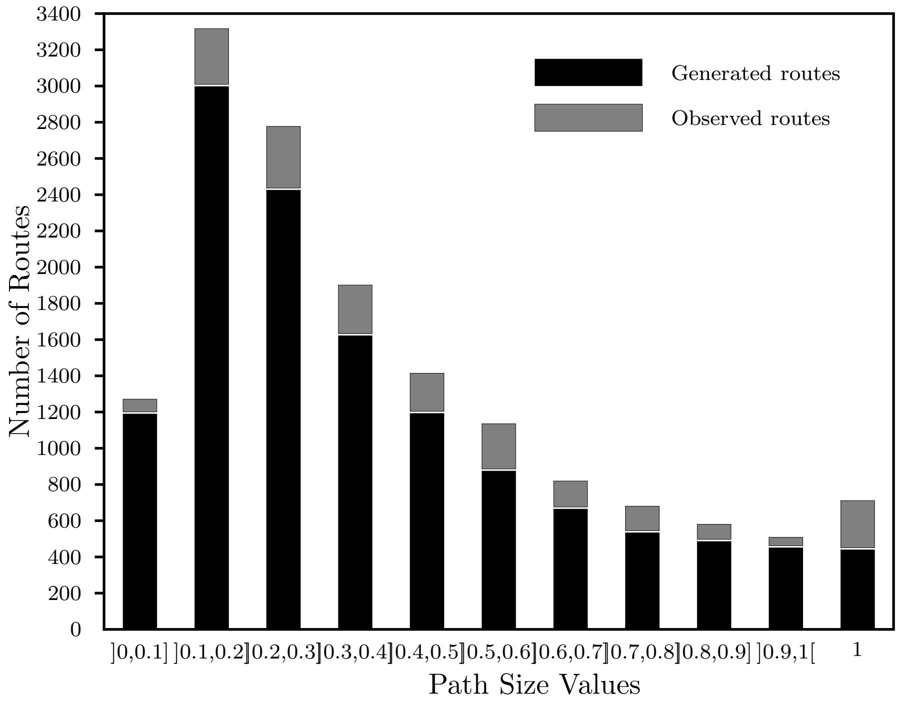

Figure 3.6: Number of Routes for PS Values

and this model would clearly need data from more than 24 individuals to be meaningful. We assume that the perception of correlated alternatives on the subnetwork is individual specific and that the taste is constant over choice situations. The error components in the correlation structure are therefore specified to be invariant across the observations of a given individual.

Model Estimation

The parameter estimates are given in Table 3.3. We have provided a scaled parameter estimate in order to facilitate the comparison of different models.

Attribute	Min	Average	Max
Estimated Travel Time [min]	0.5	4.2	37.5
Number of Left Turns	0	3.2	27
Average Link Length [m]	11	198.7	2947
Number of Speed Bumps	0	0.3	5
ln(PS)	-3.7	-0.9	0

Table 3.2: Statistics on Attributes

The scaling is based on the estimated travel time parameter in the MNL model. The magnitude of the scaled estimate for this parameter is consequently the same for all the models.

We start by comparing the models PSL, EC_1 and EC_2. The parameter estimates shown in Table 3.3 related to average link length, estimated travel time, number of left turns and number of speed bumps are all significantly different from zero. Moreover, the parameter values as well as the robust t-test statistics are stable when comparing the different models.

The PS parameter estimate, β_{PS}, is negative and significantly different from zero and from minus one in models PSL, EC_1 and EC_2. As discussed in Section 3.1 a negative value of β_{PS} is not consistent with choice theory since it corresponds to a scale parameter and consequently should be positive. The negative estimate suggests that the PS attribute captures an attractiveness for overlapping paths. An increase in magnitude and significance of the scaled β_{PS} estimates can be noted when comparing EC_1 with PSL and EC_2 with EC_1. More precisely, when the correlation structure on the subnetwork is explicitly captured by the error terms, the value of β_{PS} increases in magnitude and significance. Based on these results, we draw the conclusion that the PS attribute has an ambiguous interpretation. On the one hand, it negatively corrects the utility for the independence assumption on the random terms. On the other hand, it has a behavioral interpretation. Namely, it captures an attractiveness for overlapping paths, for example, because they provide the possibility of route switching (this has also been suggested by Hoogendoorn-Lanser et al., 2005, in the context of multi-modal route choice). Another possible explanation for the negative β_{PS} estimate is based on the choice set definition. A majority of the observed paths have a high overlap with other paths in the choice set (see Figure 3.6). Hence, the utility is increased for overlapping paths.

The estimates of σ_{R50S} (see Table 3.3) are not significantly different from zero. This can be explained by the limited number of observations using this subnetwork component. As shown in Table 3.1, 173 observations use R.50 S but since the number of weighted observations is only 36, the length by which they overlap with the subnetwork component is relatively short. The other covariance parameter estimates are all significant (σ_{R50N} in EC'_2 at 10% significance level).

The estimation results for the models without a PS attribute (MNL, EC'_1 and EC'_2) are comparable to the results for the other models. As expected, the covariance parameter estimates are slightly different but all significant (except σ_{R50S}). This is also the case for the parameters associated with the explanatory variables.

Based on the log likelihood values reported in Table 3.4, and the likelihood ratio tests shown in Table 3.5, the MNL model can, as expected, be rejected in favor of the PSL model. The EC_1 and EC_2 models are significantly better than the PSL model, and EC_2 has better model fit than EC_1. The hypothesis of equal covariance parameters for all subnetwork components can therefore be rejected although not as strongly as the PSL model. Moreover, we note that the two EC models including a PS attribute are significantly better than those that do not.

Considering the significant improvement in model fit for the EC models compared to the PSL model, as well as the significant covariance parameter estimates, we conclude that the specification based on the subnetwork captures an important correlation structure.

Finally, we compare EC_2 with EC_3 where EC_3 explores the panel data structure of the observations. Referring to the scaled parameter estimates in Table 3.3 for average link length, estimated travel time, number of left turns and number of speed bumps, the value of the estimates are stable. On the contrary, the value β_{PS} decreases in magnitude, breaking a trend where it has been increasing in magnitude for the models EC_1 and EC_2 compared to the PSL model. It is possible that the EC_3 model better captures individuals' perception of overlapping paths than EC_1 and EC_2. The behavioral aspect that the PS attribute captures in models EC_1 and EC_2 is therefore captured within the error structure of EC_3. This would explain the decreased magnitude of the β_{PS} value.

All the covariance parameter estimates, except for σ_{R50S}, are significant in the EC_3 model. The assumption that the perception of correlated alternatives on the subnetwork is individual specific and that the taste is constant over choice situations seems to correspond to the observations.

Due to the small number of individuals there is a systematic loss in significance for all parameters in EC_3 compared to EC_2. In spite of this, there is a noticeable increase in model fit (see Table 3.4) compared to EC_2.

Forecasting Results

Route choice models are often used to predict individual behavior. It is therefore important to compare models, not only in terms of model fit, but also regarding the performance of predicting choice probabilities.

In order to test the prediction power of the different models, we estimate models based on subsets of the data saving the rest to validate the predicted probabilities. The data selection is based on origin-destination (OD) pairs. More precisely, 80% of the OD pairs are randomly selected and all observations associated with these OD pairs are included in the estimation data set. This test is particularly challenging since the models predict choice proba-

Parameters	**MNL**	**PSL**	**EC_1**	**EC'_1**	**EC_2**	**EC'_2**	**EC_3**
Path Size		**-0.28**	**-0.49**		**-0.53**		**-0.32**
Scaled estimate		*-0.33*	*-0.53*		*-0.56*		*-0.36*
Rob. std		0.07	0.09		0.09		0.19
Rob. t-test 0		-4.05	-5.61		-5.91		-1.65
Avg Link Length	**3.85**	**4.15**	**4.98**	**4.45**	**5.06**	**4.50**	**4.75**
Scaled estimate	*3.85*	*4.85*	*5.35*	*3.86*	*5.39*	*3.82*	*5.29*
Rob. std	0.53	0.55	0.60	0.55	0.61	0.56	1.21
Rob. t-test 0	7.32	7.58	8.32	8.05	8.28	8.00	3.92
Estimated Time	**-0.46**	**-0.40**	**-0.43**	**-0.54**	**-0.44**	**-0.55**	**-0.42**
Scaled estimate	*-0.46*	*-0.46*	*-0.46*	*-0.46*	*-0.46*	*-0.46*	*-0.46*
Rob. std	0.05	0.05	0.06	0.06	0.06	0.06	0.10
Rob. t-test 0	-9.61	-7.85	-7.47	-8.74	-7.51	-8.76	-4.37
Nb. Left turns	**-0.31**	**-0.32**	**-0.33**	**-0.31**	**-0.33**	**-0.31**	**-0.33**
Scaled estimate	*-0.31*	*-0.37*	*-0.35*	*-0.27*	*-0.35*	*-0.27*	*-0.36*
Rob. std	0.02	0.02	0.02	0.02	0.02	0.02	0.04
Rob. t-test 0	-15.72	-15.73	-15.62	-15.49	-15.59	-15.39	-9.16
Nb. Speed Bumps	**-0.19**	**-0.23**	**-0.22**	**-0.15**	**-0.23**	**-0.15**	**-0.22**
Scaled estimate	*-0.19*	*-0.27*	*-0.24*	*-0.13*	*-0.24*	*-0.13*	*-0.24*
Rob. std	0.06	0.07	0.07	0.06	0.07	0.06	0.19
Rob. t-test 0	-3.04	-3.52	-3.14	-2.50	-3.14	-2.44	-1.11
σ			**1.44**	**1.09**			
Scaled estimate			*1.55*	*0.94*			
Rob. std			0.19	0.18			
Rob. t-test 0			7.57	6.02			
$\sigma_{\mathbf{R50N}}$					**1.07**	**0.70**	**1.78**
Scaled estimate					*1.14*	*0.60*	*1.99*
Rob. std					0.32	0.42	0.67
Rob. t-test 0					3.28	1.68	2.66
$\sigma_{\mathbf{R50S}}$					**0.27**	**0.18**	**0.69**
Scaled estimate					*0.28*	*0.15*	*0.77*
Rob. std					0.69	0.89	0.60
Rob. t-test 0					0.39	0.20	1.16
$\sigma_{\mathbf{R70N}}$					**2.04**	**1.53**	**0.65**
Scaled estimate					*2.17*	*1.30*	*0.73*
Rob. std					0.39	0.39	0.26
Rob. t-test 0					5.16	3.91	2.55
$\sigma_{\mathbf{R70S}}$					**1.52**	**1.19**	**0.83**
Scaled estimate					*1.62*	*1.01*	*0.93*
Rob. std					0.22	0.20	0.20
Rob. t-test 0					7.08	5.80	4.07
$\sigma_{\mathbf{RC}}$					**2.02**	**1.48**	**1.19**
Scaled estimate					*2.14*	*1.25*	*1.33*
Rob. std					0.66	0.75	0.32
Rob. t-test 0					3.05	1.96	3.75

Table 3.3: Estimation Results

Model	Nb. σ Estimates	Nb. Estimated Parameters	Final L-L	Adjusted Rho-Square
MNL	-	12	-4186.07	0.152
PSL	-	13	-4174.72	0.154
EC_1	1	14	-4142.40	0.161
EC'_1	1	13	-4165.59	0.156
EC_2	5	18	-4136.92	0.161
EC'_2	5	17	-4162.74	0.156
EC_3	5	18	-4109.73	0.166
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 2978 observations Null log likelihood: -4951.11 BIOGEME (biogeme.epfl.ch) has been used for all model estimations (Bierlaire, 2003, Bierlaire, 2007).				

Table 3.4: Model Fit Measures

Model 1	Model 2	Test	Threshold (5%)
MNL	PSL	22.70	3.84
PSL	EC_1	64.64	3.84
PSL	EC_2	75.60	11.07
EC_1	EC_2	10.96	9.49
EC'_1	EC_1	46.38	3.84
EC'_2	EC_2	51.64	3.84

Table 3.5: Likelihood Ratio Test

	Estimation		Forecast	
	Sample Size	Null L-L	Sample Size	Null L-L
Data 1	2386	-3944.28	592	-1006.84
Data 2	2359	-3905.29	619	-1045.82
Data 3	2409	-3982.19	569	-968.918
Data 4	2328	-3952.23	650	-998.886
Data 5	2385	-4013.04	593	-938.078

Table 3.6: Information on Datasets used for Forecasting

bilities for OD pairs whose choice sets have not been used for estimating the models.

Five datasets have been generated. The size and null log likelihood for each of them are reported in Table 3.6. Since, in general, there is only one observation per OD pair, all the datasets have more or less the same size.

The same models as in the previous section are estimated (except EC_3). The model fit measures and the likelihood ratio tests are reported in Table 3.7. The general interpretation of the estimation results for all datasets remains the same as for the models estimated on the complete dataset. The conclusions regarding the model fit measures are also the same here as for the estimations based on the full dataset, with the only exception that the EC'_2 model is not significantly better than the EC'_1 model for dataset five.

We compare the performance of the different models using the log likelihood of the predicted probabilities computed on the data not used for estimation. The log likelihood values for all models and datasets are reported in Figure 3.7, which shows that the EC models are superior to the MNL and the PSL models. The EC'_1 and EC'_2 models are performing better than MNL and PSL for all datasets. The EC_1 and EC_2 models outperform all other models for all datasets except the fourth one. The prediction performance of the PSL and MNL models are very similar, while the fit of estimated data is better for the PSL model.

The results for the fourth dataset show that the MNL model performs better than the PSL, EC_1 and EC_2 models. The EC'_1 and EC'_2 are however good for forecasting in spite of the fact that they have worse model fit than the PSL model and the other two EC models. In this particular case, even though the models including a PS attribute have in general better model fit, they have poor prediction performance. This illustrates the general statement that the best model fit does not necessarily identify the best model, and comparing prediction performances is important.

Nb. of estimated parameters		MNL 12	PSL 13	EC_1 14	EC'_1 13	EC_2 18	EC'_2 17
Data 1	Final L-L	-3279.23	-3270.68	-3248.00	-3265.14	-3243.94	-3262.44
	$\bar{\rho}^2$	0.166	0.167	0.173	0.169	0.173	0.169
Data 2	Final L-L	-3313.27	-3305.81	-3279.57	-3298.17	-3273.73	-3294.58
	$\bar{\rho}^2$	0.149	0.150	0.157	0.152	0.157	0.152
Data 3	Final L-L	-3361.90	-3353.01	-3336.46	-3350.86	-3331.16	-3347.37
	$\bar{\rho}^2$	0.153	0.155	0.159	0.155	0.159	0.155
Data 4	Final L-L	-3358.66	-3338.48	-3311.16	-3342.87	-3306.56	-3340.68
	$\bar{\rho}^2$	0.147	0.152	0.159	0.151	0.159	0.150
Data 5	Final L-L	-3408.82	-3397.29	-3373.37	-3394.89	-3370.63	-3393.43
	$\bar{\rho}^2$	0.148	0.150	0.156	0.151	0.156	0.150

Table 3.7: Model Fit Measures (Forecasting Models)

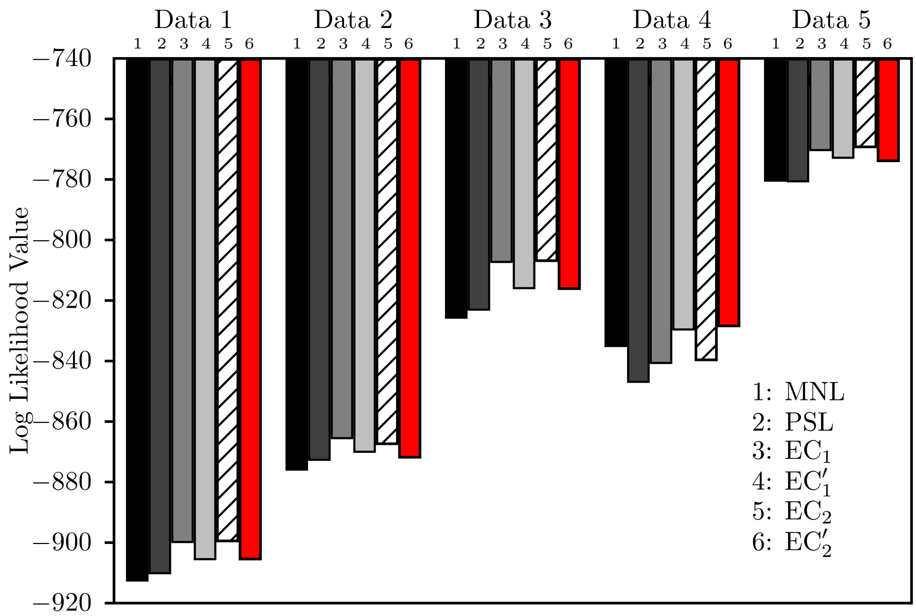

Figure 3.7: Log likelihood Values for Predicted Probabilities

3.3 Conclusions and Future Work

In this chapter we justify the use of the original Path Size formulation (or the recently proposed Path Size Correction factor) among the deterministic corrections of the i.i.d. assumption on the random terms in a MNL model. These are the formulations that both have a theoretical support and show intuitive results for the correction of the independence assumption on the random terms. Moreover, we have presented estimation results that suggest a behavioral interpretation of the Path Size attribute. Namely, overlap can be attractive for travelers since it provides the possibility of switching between different routes.

We propose a novel modeling approach based on subnetworks designed to enhance the performance of simple models with a limited increase in complexity. Estimation results show that this approach is significantly better than a simple Path Size Logit model. A subnetwork is a set of subnetwork components. Alternatives are assumed to be correlated if they use the same subnetwork component. This correlation is captured within a factor analytic specification of an Error Component model. The estimation results are promising and the estimates of the covariance parameters suggest that the specification captures an important correlation structure. Moreover, prediction tests are presented that clearly show the superiority of the Error Component model compared to Multinomial Logit and Path Size Logit models.

We believe that this model will open new perspectives for route choice modeling. It is a flexible approach where the trade-off between complexity and behavioral realism can be controlled by the analyst with the definition of the subnetwork.

More analysis is required in order to assess the sensitivity of the results with regard to the definition of the subnetwork. Both with respect to the choice and definition of components and with respect to different network topologies and datasets. In future work it would also be interesting to perform more thorough prediction tests using a higher number of samples.

Chapter 4

Route Choice Modeling with Network-free Data

The concept of path, which is the core of a route choice model, is usually too abstract for a reliable data collection process. Real data, in their original format, rarely correspond to path definitions. A typical example is GPS data, which are more and more available (see Section 2.1 for a literature review). As GPS devices do not explicitly use the transportation network, the coordinates of data points cannot be directly used and data manipulation is required in order to reconstruct paths. In the literature, such data manipulation involves map matching, trip end identification and assumptions on missing data.

Another context is when respondents are asked to describe a path that they have followed during a given trip. They are in general able to identify a sequence of locations that they have traversed, but have difficulties describing a full path in detail. For instance, Ramming (2001) estimated route choice models based on data collected in Boston. The respondents described chosen routes by naming street segments. In case of incomplete or ambiguous descriptions, the routes were reconstructed by taking the shortest path between known street segments.

We advocate that the data manipulation required by the underlying network model introduces biases and errors, and should be avoided. We propose a general modeling scheme that reconciles network-free data (such as GPS data or partially reported itineraries) with a network based model without such manipulations. The concept that bridges the gap between the data and the model is called *Domain of Data Relevance* (DDR) and corresponds to a physical area in the network where a given piece of data is relevant. The framework allows for several paths to correspond to a same observation. Fewer assumptions are therefore needed in the case of ambiguous data.

Note that some approaches have been proposed in the literature where the link between the concept of path and the data has been loosened, either in order to simplify the choice context, or because the observed choices are based on underlying latent choices. Ben-Akiva et al. (1984) construct latent alternatives in order to simplify the choice set definition in a route choice model. Instead of modeling the choice of routes where there are many feasible alternatives, they model the choice of labels, such as, fastest route, most scenic route, shortest route etc. The exact route choices are observed and used to estimate the model. Choudhury (2007) presents a general methodology for modeling choice behavior that is based on choices of plans. These underlying choices may not be observed. Both the choice of plan and observed choices are explicitly modeled in a multi-dimensional approach. She applies the methodology to freeway lane changing and merging from an on-ramp (see also Ben-Akiva et al., 2006a, and Ben-Akiva et al., 2006b).

In the following section we introduce the concept of DDR and illustrate the framework on simple examples for two different types of data (GPS data and reported trips). Moreover, we present estimation results (Section 4.4) of Path Size Logit and Subnetwork models based on a dataset of reported trips collected in Switzerland. The network is to our knowledge the largest one used in the literature for route choice analysis based on revealed preferences data.

4.1 Domain of Data Relevance

The common reference of our modeling scheme is a finite two-dimensional region with an appropriate coordinate system, typically longitude, latitude[1]. In general, it is simply the region of interest such as a city, or a country.

We define an *observation* as a sequence of individual pieces of data related to an itinerary, such as a sequence of GPS points or reported locations. For a given piece of data, the *Domain of Data Relevance* is defined as the physical area where the piece of data is relevant. Its exact definition depends on the context. For example, consider a GPS reporting coordinates (x, y). Due to the intrinsic technological limitations of the device, we can identify a 95% confidence interval around a point (x, y). This would be the DDR of this piece of data. An example of GPS data is shown in Figure 4.1 where the GPS points are represented by small circles and their corresponding DDR with dashed lines. The size of the DDR areas vary depending on the accuracy (e.g. quality of satellite signals) of each piece of data.

[1]Using a three-dimensional reference is possible and relatively straightforward. However, it would bring an unnecessary level of complexity to our model.

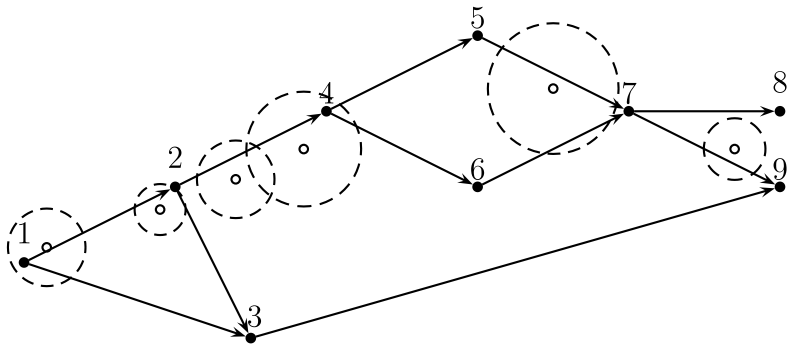

Figure 4.1: Example of GPS Data

In the context of reported paths, notions such as “downtown”, “next to the Eiffel Tower” or “intersection of Massachusetts Avenue and Newbury Street” can easily be associated with a DDR. The size of the DDR is inversely proportional to the fuzziness of the concept. It may be unambiguous (such as the area corresponding to “downtown”), or ambiguous and left to the modeler’s judgment (such as “next to the Eiffel Tower”). An example is shown in Figure 4.2 where the reported locations are “home”, “intersection Main St and Cross St”, “city center” and “mall”. The home and intersection correspond to exact locations in the network and the areas of the associated DDRs (dashed lines) are therefore small; they contain only one node. The two other reported locations are more fuzzy and the areas of the associated DDRs are therefore larger, in this case the DDRs contain two nodes.

In summary, the DDR is a modeling element whose exact definition is left to the analyst and depends on the data collection process and the network

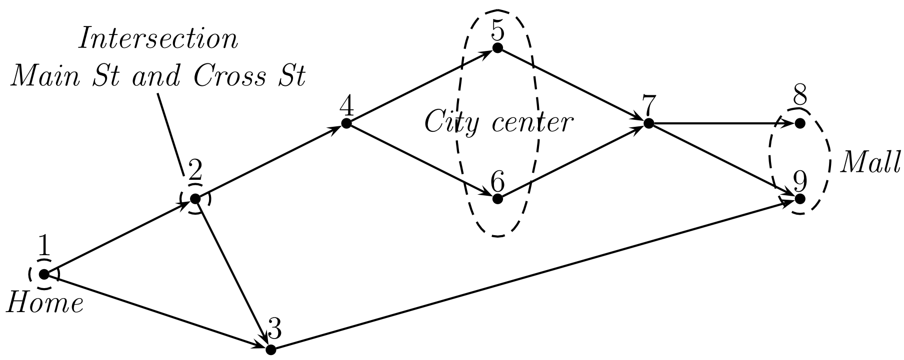

Figure 4.2: Example of a Reported Trip

topology. We now formally relate the DDR of each piece of data with the various network elements (that is, nodes and links). We define an indicator function $\delta(d, e)$ which is 1 if network element e is related with the DDR of data d, and 0 otherwise. In general, the definition of this indicator function is straightforward. If e is a node representing an intersection, it is easy to verify if it lies in the area of the DDR or not. If e is a node representing the centroid of a zone, we simply check if the zone intersects with the DDR area. Similarly, if e is a link representing a road segment, we identify if it crosses the DDR area. A node can also be associated with a DDR if it is the source or the sink node of a link crossing the DDR.

In practice, we generate for each piece of data a list of relevant network elements, which bridges the gap between the network-free data and the network model.

4.2 Model Specification

We aim at estimating the unknown parameters β of a route choice model $P(p|\mathcal{C}_n(s); \beta)$ where $\mathcal{C}_n(s)$ is the set of paths linking origin-destination (OD) pair s considered by traveler n and p is a path in $\mathcal{C}_n(s)$[2].

Let $\mathcal{S}$ be the set of all OD pairs in the network. For a given observation i of traveler n, that is a sequence of pieces of data $(d_1, d_2, \ldots, d_k)$, we first identify the set $\mathcal{S}_i$ of relevant OD pairs, that is OD pairs s such that the observation's origin node is related to the DDR of first data and the destination node is related to the last, that is

$$\mathcal{S}_i = \{s \in \mathcal{S} \mid \delta(d_1, s_o)\delta(d_k, s_d) = 1\}.$$

At least one relevant OD pair must exist and the set $\mathcal{S}_i$ must therefore be non empty. If it is empty, the definitions of the DDRs must be revised.

We derive the probability $P_n(i|\mathcal{S}_i)$ of reproducing observation i of traveler n given $\mathcal{S}_i$. It can be decomposed in the following way

$$P_n(i|\mathcal{S}_i) = \sum_{s\in\mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p\in\mathcal{C}_n(s)} P_n(i|p) P_n(p|\mathcal{C}_n(s); \beta), \tag{4.1}$$

where

- $P_n(s|\mathcal{S}_i)$ is the probability that the actual OD pair is s given the set of relevant OD pairs $\mathcal{S}_i$,

[2] Several choice sets can correspond to a same observation. Throughout this chapter we therefore explicitly note which OD pair a choice set corresponds to.

- $P_n(i|p)$ is the measurement equation, giving the probability of observing i if the actual path is p, and
- $P_n(p|\mathcal{C}_n(s); \beta)$ is the route choice model giving the probability that individual n selects path p within choice set $\mathcal{C}_n(s)$. The model depends on unknown parameters β which must be estimated. See Section 2.3 for a review of route choice models.

Since several paths can correspond to the same observation, the measurement equation plays a key role in this framework. It takes a value greater than zero if observation i corresponds to path p that is composed by links $(\ell_1, \ldots, \ell_P)$. This is the case if

- there is at least one link in the path related to each DDR, that is, for any $m = 1, \ldots, k$, there exists q, $1 \leq q \leq P$, such that $\delta(d_m, \ell_q) = 1$,
- the sequence of reported locations is consistent with the order of the links in the path, that is, for any $m_1 \leq m_2$, if $\delta(d_{m_1}, \ell_{q_1}) = 1$ and $\delta(d_{m_2}, \ell_{q_2}) = 1$, then $q_1 \leq q_2$.

We illustrate the measurement equation using the two data collection processes mentioned above.

In the context of reported trips a simple measurement equation can be defined since either the path goes through all reported locations or not. The measurement equation therefore takes the value 1 if this is the case and 0 otherwise.

For GPS collected data a more complex model may be necessary. For example, the probability that the observation i is generated by the real path p may be defined as a function of the distance between i and p. This distance can be computed since, unlike reported trips, each piece of data d is a coordinate in the network. We define a function $\Delta(d, \ell)$ which maps the euclidean distance from d to the closest point on link ℓ. The distance between a piece of data d and a path p is $D(d, p) = \min_{\ell \in A_{pd}} \Delta(d, \ell)$ where A_{pd} is the set of links that are part of path p and are located within the DDR of data d, $A_{pd} = \{\ell \in \ell_1, \ldots, \ell_P \mid \delta(d, \ell) = 1\}$. The global distance $D(i, p)$ between the observation i and the path p can be evaluated in several ways. For example, the sum of $D(d, p)$ for each piece of data in i or the average distance. A distributional assumption on $D(i, p)$ then defines the measurement equation $P(i|p)$. The evaluation of $D(i, p)$ and its distribution depend on the specific context and should be defined on a case to case basis.

If there is at least one observation i for which $|\mathcal{S}_i| > 1$ then a model for $P_n(s|\mathcal{S}_i)$ needs to be defined. Different formulations are possible depending on the available information where the most simple one assigns equal probabilities to all OD pairs, that is

$$P_n(s|\mathcal{S}_i) = \frac{1}{|\mathcal{S}_i|} \; \forall s \in \mathcal{S}_i. \tag{4.2}$$

If additional information is available, a more sophisticated model can be specified. For instance, high probabilities can be assigned to OD pairs that include home and work locations.

As discussed in the previous section, the role of the DDR is to link the network-free data to the network. Special care must be taken to define the DDRs. Although no general instructions can be provided, as the exact definition is problem dependent, we emphasize two issues to consider in this process. First, the DDR of a data d cannot be empty. If $\delta(d, e) = 0 \; \forall e$, meaning that no network element correspond to this piece of data, the DDR is not properly defined and a new specification is necessary. A possible solution is to increase the size of the DDR so that at least one link crosses the DDR. Second, if the DDRs are too large and encompass a high number of relevant network elements, the model may not be identified. Moreover, the dimensionality of the sums involved in Equation (4.1) may increase exponentially. We advice to design the DDRs in order to reflect as close as possible the data uncertainty. For instance, in the case of GPS data, the error ellipse can be used. In the case of reported data the underlying zoning system is a natural choice.

Finally we note that the route choice model is only identifiable if at least one of the routes in $\mathcal{C}_n(s)$ corresponds to the observation and at least one of the routes in $\mathcal{C}_n(s)$ does not correspond to the observation.

Maximum likelihood estimation of model (4.1) can be performed with BIOGEME (Bierlaire, 2003, Bierlaire, 2007, `biogeme.epfl.ch`).

4.3 Illustrative Examples

We illustrate the modeling framework on the two examples used previously. We start with the reported trip shown in Figure 4.2. The exact origin node is known ("home" node) but there are two possible destination nodes (8 and 9 corresponding to "mall"). The set of relevant OD pairs for this observation i is therefore $\mathcal{S}_i = \{(1, 8), (1, 9)\}$ (referred to as s_1 and s_2). No additional information is available, so we assume that the OD pairs are equally probable, that is $P(s_1|\mathcal{S}_i) = P(s_2|\mathcal{S}_i) = 0.5$. There are two routes connecting the first OD pair, $\mathcal{C}(s_1) = \{(1, 2, 4, 5, 7, 8), (1, 2, 4, 6, 7, 8)\}$, that we denote p_1 and p_2 respectively. Note that we omit the notation for individual n since we only have one observation here. The observation corresponds to both routes and consequently $P(i|p_1) = P(i|p_2) = 1$. Four routes connect the second OD pair $\mathcal{C}(s_2) = \{(1,2,4,5,7,9), (1,2,4,6,7,9), (1,2,3,9), (1,3,9)\}$ (denoted $p_3, \ldots, p_6$, respectively) but the observation only corresponds to the first two, that is $P(i|p_3) = P(i|p_4) = 1$ and $P(i|p_5) = P(i|p_6) = 0$. For this example, Equation (4.1) is therefore defined as

$$\begin{aligned} P(i|\mathcal{S}_i) = \; & \frac{1}{2}\Big[P(p_1|\mathcal{C}(s_1);\beta) + P(p_2|\mathcal{C}(s_1);\beta)\Big] + \\ & \frac{1}{2}\Big[P(p_3|\mathcal{C}(s_2);\beta) + P(p_4|\mathcal{C}(s_2);\beta)\Big] \end{aligned}$$

where $P(p_g|\mathcal{C}(s_h);\beta)$ ($g = 1, \ldots, 4$ and $h = 1, 2$) is the network based route choice model to be estimated.

We now turn our attention to the example on GPS data shown in Figure 4.1. There is one relevant origin node but the DDR of the last piece of data does not contain any node. We therefore consider the sink node of the link that crosses this DDR. Hence, there is one relevant OD pair for this observation i, $\mathcal{S}_i = \{(1,9)\}$, that we denote s. Similar to the example on the reported trip, there are four routes in the choice set, $\mathcal{C}(s) = \{(1,2,4,5,7,9), (1,2,4,6,7,9), (1,2,3,9), (1,3,9)\}$, now denoted $p_1, \ldots, p_4$. The observation corresponds to the first two routes and therefore $P(i|p_3) = P(i|p_4) = 0$. $P(i|p_1)$ and $P(i|p_2)$ can be defined as a function of the distances between the observed locations and the path. In Figure 4.3 we show how the distance between the fourth piece of data and the paths could be computed. The figure shows links $(2,4)$, $(4,5)$ and $(4,6)$ that all cross the DDR of d_4 (see Figure 4.1). Since both p_1 and p_2 use link $(2,4)$ and $\Delta(d_4,(4,5)) = \Delta(d_4,(4,6)) > \Delta(d_4,(2,4))$ the distance between d_4 and the paths p_1 and p_2 is $\Delta(d_4,(2,4))$. For this example the model given by Equation (4.1) is

$$P(i|s) = P(i|p_1)P(p_1|\mathcal{C}(s);\beta) + P(i|p_2)P(p_2|\mathcal{C}(s);\beta).$$

4.4 Case Study

In this section we illustrate the modeling framework on a dataset collected in Switzerland. The data concern long distance route choice behavior and were collected via telephone interviews (Vrtic et al., 2006). The respondents were asked to describe their last long distance trip with the names of the origin and destination cities as well as maximum three intermediate cities or locations that they passed through. An example is shown in Figure 4.4 where

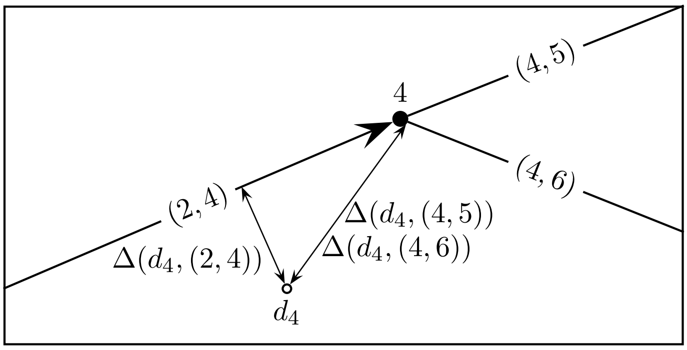

Figure 4.3: Example of GPS Data (Continued)

a traveler went from Belmont-sur-Lausanne to Vandoeuvres passing through Morges, Aubonne and Nyon. 940 reported trips are available for route choice analysis.

In this context, the DDR of each reported location is defined by the corresponding zip code. The size of the zip code areas vary (average 3.8 km^2) depending on their location. Zip codes in rural regions can cover a large area whereas zip code areas in urban regions can be small. When linking the network-free data with the network through the DDRs it is important to make sure that the precision level of the observations corresponds to the precision level of the network. Since the descriptions of the observations are approximate, we use a simplified transportation network (Swiss national model, Vrtic et al., 2005). This network covers all regions in Switzerland and contains 39411 unidirectional links and 14841 nodes (to be compared with the Swiss TeleAtlas network that contains approximately 1 million unidirectional links and half a million nodes). Very small roads are not considered which simplifies the network especially in urban areas. To our knowledge, this is the largest network used for estimation of route choice models based on revealed preferences data presented in the literature.

In order to estimate a route choice model we need to specify $P(s|\mathcal{S}_i)$ and choice sets $\mathcal{C}_n(s)\ \forall s \in \mathcal{S}$. The observations contain no information on relevant OD pairs. Due to the computationally demanding choice set generation as well as memory constraints, we do not consider all possible OD pairs for each observation but randomly choose two OD pairs (if more than one is available) and use the probability model given by Equation (4.2). For each OD pair we generate a choice set of 45 routes using a stochastic choice set generation approach (Bierlaire and Frejinger, 2007). After the choice set generation there are 780 observations available for model estimation. 160 observations are not considered because either all or none of the generated

Figure 4.4: Example of an Observation

routes correspond to the observation.

We estimate two different types of route choice models $P_n(p|\mathcal{C}_n(s);\beta)$: one Path Size Logit (PSL) model (Ben-Akiva and Ramming, 1998, and Ben-Akiva and Bierlaire, 1999b) and one Subnetwork model (Frejinger and Bierlaire, 2007, described in Chapter 3). With the latter, we explicitly model the correlation among paths on a subnetwork using an Error Component model. Here we create a subnetwork composed of all main freeways. We estimate one covariance parameter which is assumed proportional to the length by which the paths overlap with the subnetwork. The transportation network is shown in Figure 4.5 where the subnetwork is marked with bold lines.

Finally, we specify the deterministic utility functions using the attributes reported in Table 4.1. Namely, Path Size, free-flow travel time and road type attributes. Note that departure time is unknown, we therefore use free-flow travel time as an approximation of travel time. The type of road is defined according to an existing hierarchy of the links. We define four road types; freeway (FW), cantonal/national (CN), main and small roads. The cantonal/national roads connect different regions in Switzerland but have a lower capacity and speed limit than freeways. Main roads refer to fast local roads in urban or rural areas and small roads are the remaining ones.

Both models have the same linear-in-parameters specifications. More precisely, a piecewise linear specification of free-flow travel time (measured in hours) is used in order to capture travelers' sensitivity to changes in travel

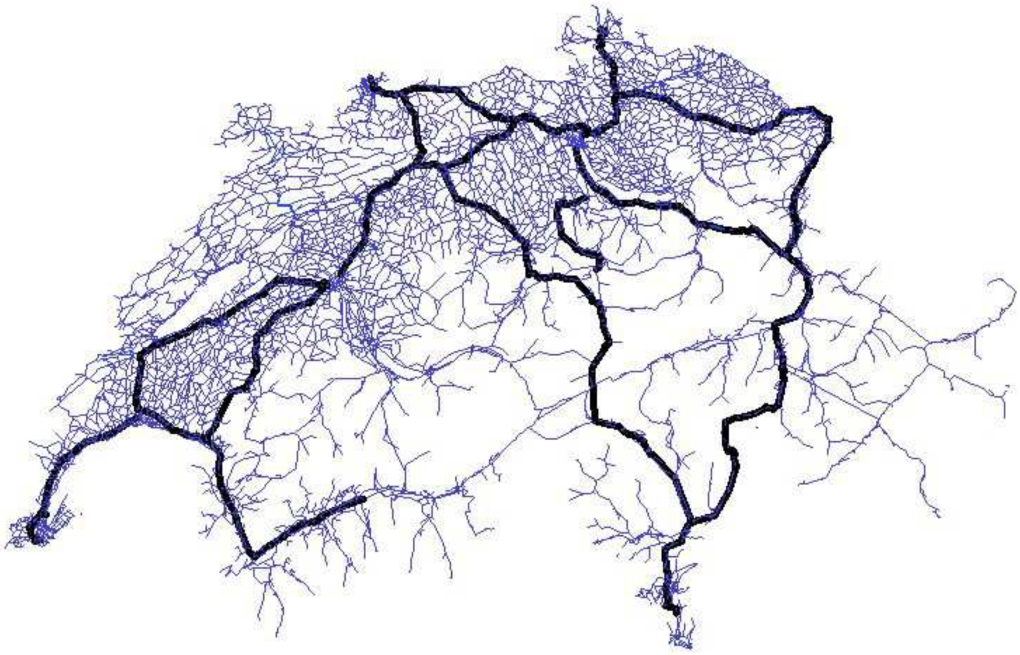

Figure 4.5: Swiss National Network

Attribute	Min	Average	Max
Path Size	0.02	0.17	0.96
ln(Path Size)	-3.74	-1.95	-0.04
Proportion of free-flow time on freeway	0.00	0.29	1.00
Proportion of free-flow time on CN	0.00	0.27	1.00
Proportion of free-flow time on main	0.00	0.23	1.00
Proportion of free-flow time on small	0.00	0.21	1.00
Free-flow travel time [minutes]	8	49.00	523

Table 4.1: Statistics on Routes Corresponding to Observations

time in different ranges of the variable. After systematic testing of different endpoints for the ranges we have defined a specific piecewise linear approximation of the contribution to utility of free-flow travel time for each of the four road types. The utility functions also include a Path Size attribute and the four variables representing the proportion of the total travel time on each type of road.

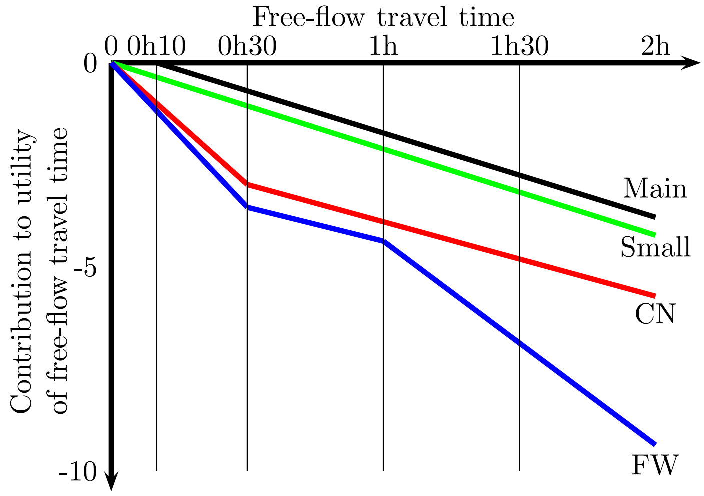

Figure 4.6: Piecewise Linear Specification - PSL Model

In Figure 4.6 we illustrate the piecewise linear specification of the free-flow travel time by graphically visualizing the estimates for the PSL model. The coefficient estimates for all the explanatory variables are reported in Table 4.2. The coefficients have their expected signs and are significantly different from zero. The estimates indicate that travelers are more sensitive to an increase in travel time on freeways than on other road types for travel times greater than one hour. This may be due to the free-flow travel time approximation underestimating the actual travel time on cantonal/national roads. For small and main roads there are few observations and the estimates can therefore not be considered reliable for this time interval. The advantage of a piecewise-linear approximation is that we can estimate with good precision coefficients for different intervals of the variable, provided that there are enough observations. In this case, more observations are needed for travel times greater than one hour on each road type in order to have reliable coefficient estimates.

We provide scaled coefficient estimates in order to facilitate the comparison of the two models. The scaling is based on the “freeway free-flow time 0-30 min” coefficient. The magnitude of the scaled estimate for this coefficient is hence the same for both models. The scaled estimates have comparable magnitudes for the two models. This is also the case for the robust standard errors and the t-test statistics are therefore similar. We conclude that the estimation results are stable for the different model structures.

The model fit measures and the coefficients related to the correlation structure are reported in Table 4.3. The Path Size coefficient estimates are positive which is consistent with theory (Frejinger and Bierlaire, 2007, and Chapter 3). Indeed, this results in a negative correction of the utility for overlapping paths. The covariance estimate is significantly different from zero which can be interpreted as there is a significant correlation among paths using freeways. Furthermore, the Subnetwork model has a significantly better model fit than the Path Size Logit model (the likelihood ratio test statistic is 6.756 to be compared with $\chi^2_{0.05,1} = 3.84$) which is also consistent with the findings in Frejinger and Bierlaire (2007), described in Chapter 3.

4.5 Conclusions and Future Work

Link-by-link descriptions of chosen routes are rarely directly available and data manipulation is necessary in order to obtain network compliant paths for the estimation of route choice models. We argue that data manipulation introduces biases and errors and should be avoided. We propose a general modeling framework that reconciles network-free data (for example partially reported trips and GPS data) with a network based model without such manipulations. The concept that bridges the gap between the data and the model is called *Domain of Data Relevance* and corresponds to a physical area in the network where a given piece of data is relevant. The DDR allows to avoid additional arbitrary assumptions. For instance, when there is ambiguity about the chosen path, the DDR allows to maintain the ambiguity in the model rather than assuming an "observed" path.

In this framework any existing route choice model can be estimated based on observations that are defined by sequences of individual pieces of data (estimation is available in BIOGEME). We illustrate the framework with simple examples for two different types of data: GPS data and reported trips. Moreover, we provide estimation results of Path Size Logit and Subnetwork models based on a real dataset of reported trips. The network is to our knowledge the largest one used in the literature for route choice analysis based on revealed preferences data.

We believe that this approach makes the route choice modeling results more accurate. Moreover, it makes the estimation of the models easier since the complex data manipulation can be limited to a minimum. We provide the methodology for estimating models based on GPS data. Since no GPS

Coefficient	**PSL**	**Subnetwork**
Freeway free-flow time 0-30 min	**-7.12**	**-7.45**
Scaled Estimate	*-7.12*	*-7.12*
(Rob. Std. Error) Rob. t-test	(0.877) -8.12	(0.984) -7.57
Freeway free-flow time 30min - 1 hour	**-1.69**	**-2.26**
Scaled Estimate	*-1.69*	*-2.16*
(Rob. Std. Error) Rob. t-test	(0.875) -1.93	(1.03) -2.19
Freeway free-flow time 1 hour +	**-4.98**	**-5.64**
Scaled Estimate	*-4.98*	*-5.39*
(Rob. Std. Error) Rob. t-test	(0.772) -6.45	(1.00) -5.61
CN free-flow time 0-30 min	**-6.03**	**-6.25**
Scaled Estimate	*-6.03*	*-5.97*
(Rob. Std. Error) Rob. t-test	(0.882) -6.84	(0.975) -6.41
CN free-flow time 30 min +	**-1.87**	**-2.16**
Scaled Estimate	*-1.87*	*-2.06*
(Rob. Std. Error) Rob. t-test	(0.331) -5.64	(0.384) -5.63
Main free-flow travel time 10 min +	**-2.03**	**-2.46**
Scaled Estimate	*-2.03*	*-2.35*
(Rob. Std. Error) Rob. t-test	(0.502) -4.05	(0.624) -3.95
Small free-flow travel time	**-2.16**	**-2.75**
Scaled Estimate	*-2.16*	*-2.63*
(Rob. Std. Error) Rob. t-test	(0.685) -3.16	(0.804) -3.42
Proportion of time on freeways	**-2.20**	**-2.31**
Scaled Estimate	*-2.20*	*-2.21*
(Rob. Std. Error) Rob. t-test	(0.812) -2.71	(0.865) -2.67
Proportion of time on CN	**0 fixed**	**0 fixed**
Proportion of time on main	**-4.43**	**-4.40**
Scaled Estimate	*-4.43*	*-4.21*
(Rob. Std. Error) Rob. t-test	(0.752) -5.88	(0.800) -5.51
Proportion of time on small	**-6.23**	**-6.02**
Scaled Estimate	*-6.23*	*-5.75*
(Rob. Std. Error) Rob. t-test	(0.992) -6.28	(1.03) -5.83

Table 4.2: Estimation Results

Coefficient	**PSL**	**Subnetwork**
ln(Path Size) based on free-flow time	**1.04**	**1.10**
Scaled Estimate	*1.04*	*1.05*
(Rob. Std. Error) Rob. t-test	(0.134) 7.81	(0.141) 7.78
Covariance		**0.217**
Scaled Estimate		*0.205*
(Rob. Std. Error) Rob. t-test		(0.0543) 4.00
Number of simulation draws	-	1000
Number of parameters	11	12
Final log likelihood	-1164.850	-1161.472
Adjusted rho square	0.145	0.147
Sample size: 780, Null log likelihood: -1375.851		
BIOGEME (Bierlaire, 2003, Bierlaire, 2007) has been used for all model estimations		

Table 4.3: Estimation Results (Continued)

dataset in its original form (sequences of GPS points) is at our disposal, the estimation based on this type of data is left for future research.

Chapter 5

Sampling of Paths

In this chapter we present a new paradigm for choice set generation and route choice modeling. Existing approaches assume that actual choice sets are generated. Empirical results suggest however that this is not true. Indeed, we are unaware of any real application where all observed paths are generated (see studies on coverage, discussed in Section 2.2.3) which clearly leads to the suspicion that not all alternatives are found by path generation algorithms.

Instead of focusing on finding alternatives actually considered by travelers, which is nearly impossible to evaluate (Section 2.2.3), we propose an approach where we focus on obtaining unbiased parameter estimates. We assume that actual choice sets are the sets of all paths connecting each origin-destination pair. Although this is behaviorally questionable, we expect this assumption to avoid bias in the econometric model. The sets of all paths are however impossible to generate explicitly and we propose a stochastic path generation algorithm based on an importance sampling approach. The path utilities must then be corrected according to the used sampling protocol in order to obtain unbiased parameter estimates. We derive such a sampling correction for the proposed algorithm. Furthermore, based on the assumption that actual choice sets contain all paths, we argue that Path Size (or Commonality Factor) attributes should be computed based on the universal choice set in order to reflect the true correlation structure. Since this is not possible in a real application, we propose a heuristic for computing an Extended Path Size attribute. We therefore use two different sets of paths: one for the model estimation and one, larger, for the Path Size computation.

In the following section we describe the proposed algorithm and we continue by deriving the sampling correction in Section 5.2. Note that existing stochastic path generation approaches (Section 2.2) may also be viewed as importance sampling approaches. We are however unaware of how to compute in a straightforward way the sampling correction for these algorithms.

In Section 5.3 we present numerical results based on synthetic data and describe the Extended Path Size heuristic.

5.1 A Stochastic Path Generation Approach

This stochastic path generation approach is flexible and can be used in various algorithms including those presented in the literature. We start by describing the general approach and then focus on a specific instance based on a biased random walk.

For a given origin-destination pair (s_o, s_d), the general approach associates a weight $\omega(\ell|b_1, b_2)$ with each link $\ell = (v, w)$ based on its distance to the shortest path according to a given generalized cost. More precisely, $\omega(\ell|b_1, b_2)$ is defined by the double bounded Kumaraswamy distribution (Kumaraswamy, 1980), that is

$$\omega(\ell|b_1, b_2) = 1 - (1 - {x_\ell}^{b_1})^{b_2}. \tag{5.1}$$

b_1 and b_2 are shape parameters and $x_\ell \in [0, 1]$ represents a measure of distance to the shortest path and is defined as

$$x_\ell = \frac{SP(s_o, s_d)}{SP(s_o, v) + C(\ell) + SP(w, s_d)}, \tag{5.2}$$

where $C(\ell)$ is the generalized cost of link ℓ, and $SP(v_1, v_2)$ is the generalized cost of the shortest path between nodes v_1 and v_2. Note that x_ℓ equals one if ℓ is part of the shortest path and $x_\ell \to 0$ as $C(\ell) \to \infty$. In Figure 5.1 we show the cumulative distribution function for different values of b_1 when $b_2 = 1$. The weights assigned to the links can be controlled by the definition of the distribution parameters. High values of b_1 when $b_2 = 1$ yield low weights for links with high cost. Low values of b_1 have the opposite effect.

Note that other distributions with suitable properties can be used. It is also worth mentioning that this idea presents similarities in its nature with the approach proposed by Dial (1971).

Once a weight has been assigned to each link, various methods can be applied. Bierlaire and Frejinger (2007) propose a gateway approach, used by Bierlaire and Frejinger (to appear) for modeling long distance route choice behavior in Switzerland. Note also that the method can be generalized to subpaths instead of links, in order to better reflect behavioral perceptions (see Frejinger and Bierlaire, 2007).

In this chapter, we use a biased random walk algorithm which is appropriate for an importance sampling approach. First, it generates any path in $\mathcal{U}$ with non-zero probability. Second, path selection probabilities can be computed in a straightforward way.

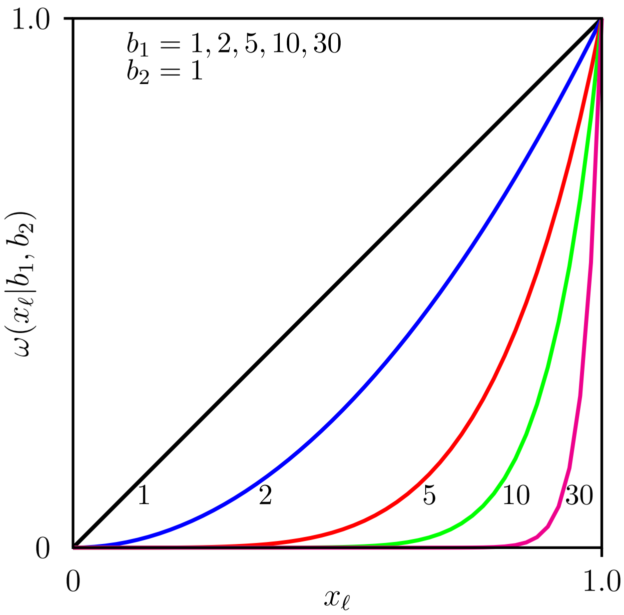

Figure 5.1: Kumaraswamy Distribution: Cumulative Distribution Function

Given an origin s_o and a destination s_d, an ordered set of links Γ is generated as follows:

Initialize $v = s_o$, $\Gamma = \emptyset$

Loop While $v \neq s_d$ perform the following

Weights For each link $\ell = (v, w) \in \mathcal{E}_v$, where $\mathcal{E}_v$ is the set of outgoing links from v, we compute the weights based on (5.1) where x_ℓ is defined by

$$x_\ell = \frac{SP(v, s_d)}{C(\ell) + SP(w, s_d)}. \tag{5.3}$$

Note that this is equivalent to (5.2) as v is considered the origin.

Probability For each link $\ell = (v, w) \in \mathcal{E}_v$, we compute

$$q(\ell|\mathcal{E}_v, b_1, b_2) = \frac{\omega(\ell|b_1, b_2)}{\sum_{m \in \mathcal{E}_v} \omega(m|b_1, b_2)} \tag{5.4}$$

Draw Randomly select a link (v, w^*) in $\mathcal{E}_v$ based on the above probability distribution.

Update path $\Gamma = \Gamma \cup (v, w^*)$

Next node $v = w^*$.

The algorithm biases the random walk towards the shortest path in a way controlled by the parameters of the distribution. The algorithm corresponds to a simple random walk if a uniform distribution (special case of Kumaraswamy distribution with $b_1 = 1$ and $b_2 = 1$) is used. Note however that a simple random walk does not generate a simple random sample of paths.

The probability $q(j)$ of generating a path j is the probability of selecting the ordered sequence of links Γ_j

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell|\mathcal{E}_v, b_1, b_2), \tag{5.5}$$

where $q(\ell|\mathcal{E}_v, b_1, b_2)$ is defined by (5.4).

With this algorithm, it is easy to compute path selection probabilities and it is not computationally demanding since at most $|\mathcal{V}|^2$, where $\mathcal{V}$ is the number of nodes in the network, shortest path computations are needed for any number of observations.

5.2 Sampling Correction

As discussed in Section 2.4, the correction terms $q(\mathcal{C}_n|j)\ \forall\ j \in \mathcal{C}_n$ must be defined for this type of sampling protocol in order to obtain unbiased parameter estimates. It is worth mentioning that if alternative specific constants are estimated, all parameter estimates except the constants would be unbiased even if the correction is not included in the utilities (Manski and Lerman, 1977). In a route choice context it is in general not possible to estimate alternative specific constants due to the large number of alternatives and the correction for sampling is therefore essential.

We define a sampling protocol for path generation as follows: a set $\widetilde{\mathcal{C}}_n$ is generated by drawing Ψ_n paths with replacement from the universal set of paths $\mathcal{U}$ using the biased random walk method described before, and then adding the chosen path to it ($|\widetilde{\mathcal{C}}_n| = \Psi_n + 1$). We assume without loss of generality that $\mathcal{U}$ is bounded with size J. Note that J is unknown in practice. Each path $j \in \mathcal{U}$ has sampling probability $q(j)$ defined by (5.5).

The outcome of this protocol is $(\widetilde{k}_{1n}, \widetilde{k}_{2n}, \ldots, \widetilde{k}_{Jn})$ where $\widetilde{k}_{jn}$ is the number of times alternative j is drawn ($\sum_{j\in\mathcal{U}} \widetilde{k}_{jn} = \Psi_n$). Following Ben-Akiva (1993) we derive $q(\mathcal{C}_n|j)$ for this sampling protocol. The probability of an outcome is given by the multinomial distribution

$$P(\widetilde{k}_{1n}, \widetilde{k}_{2n}, \ldots, \widetilde{k}_{Jn}) = \frac{\Psi_n!}{\prod_{j\in\mathcal{U}} \widetilde{k}_{jn}!} \prod_{j\in\mathcal{U}} q(j)^{\widetilde{k}_{jn}}. \tag{5.6}$$

The number of times alternative j appears in $\widetilde{\mathcal{C}}_n$ is $k_{jn} = \widetilde{k}_{jn} + \delta_{jc}$, where c denotes the index of the chosen alternative and δ_{jc} equals one if $j = c$ and zero otherwise. Let $\mathcal{C}_n$ be the set containing all alternatives corresponding to the Ψ_n draws ($\mathcal{C}_n = \{j \in \mathcal{U} \mid k_{jn} > 0\}$). The size of $\mathcal{C}_n$ ranges from one to $\Psi_n + 1$; $|\mathcal{C}_n| = 1$ if only duplicates of the chosen alternative were drawn and $|\mathcal{C}_n| = \Psi_n + 1$ if the chosen alternative was not drawn nor were any duplicates.

The probability of drawing $\mathcal{C}_n$ given the chosen alternative i (randomly drawn $k_{in} - 1$ times) can be defined using Equation (5.6) as

$$q(\mathcal{C}_n|i) = q(\widetilde{\mathcal{C}}_n|i) = \frac{\Psi_n!}{(k_{in}-1)! \prod_{\substack{j \in \mathcal{C}_n \\ j \neq i}} k_{jn}!} q(i)^{k_{in}-1} \prod_{\substack{j \in \mathcal{C}_n \\ j \neq i}} q(j)^{k_{jn}} \tag{5.7}$$

where the products now are over all elements in $\mathcal{C}_n$ since the terms for alternatives that are not drawn ($k_{jn} = 0$) equal one. Equation (5.7) can be reformulated as

$$q(\mathcal{C}_n|i) = \frac{\Psi_n!}{\frac{1}{k_{in}} \prod_{j \in \mathcal{C}_n} k_{jn}!} \frac{1}{q(i)} \prod_{j \in \mathcal{C}_n} q(j)^{k_{jn}} = K_{\mathcal{C}_n} \frac{k_{in}}{q(i)} \tag{5.8}$$

where

$$K_{\mathcal{C}_n} = \frac{\Psi_n!}{\prod_{j \in \mathcal{C}_n} k_{jn}!} \prod_{j \in \mathcal{C}_n} q(j)^{k_{jn}}.$$

Note that the positive conditioning property is trivially verified, that is

$$q(\mathcal{C}_n|i) > 0 \;\Rightarrow q(\mathcal{C}_n|j) > 0 \;\forall\; j \in \mathcal{C}_n.$$

We can now define the probability (2.6) that an individual chooses alternative i in $\mathcal{C}_n$ as

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in} + \ln\left(\frac{k_{in}}{q(i)}\right)}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn} + \ln\left(\frac{k_{jn}}{q(j)}\right)}}, \tag{5.9}$$

where $K_{\mathcal{C}_n}$ in Equation (5.8) cancels out since it is constant for all alternatives in $\mathcal{C}_n$. When using the previously presented biased random walk algorithm we consequently only need to count the number of times a given path j is generated as well as its sampling probability given by Equation (5.5) which are both straightforward to compute.

5.3 Numerical Results

The numerical results presented in this section aim at evaluating the impact on estimation results of

- the sampling correction,
- the definition of the Path Size (PS) attribute and
- the biased random walk algorithm parameters.

Synthetic data are used for which the true model structure and parameter values are known. Based on these data we then evaluate different model specifications with the t-test values of the parameter estimates with respect to (w.r.t.) their corresponding true values. In the following we refer to a parameter estimate as biased if it is significantly different from its true value at 5% significance level (critical value: 1.96).

5.3.1 Synthetic Data

The network is shown in Figure 5.2 and is composed of 38 nodes and 64 links. It is a network without loops and the universal choice set $\mathcal{U}$ can therefore be enumerated ($|\mathcal{U}| = 170$). The length of the links is proportional to the length in the figure and some links have a speed bump (SB).

Observations are generated with a postulated model. In this case we use a Path Size Logit (PSL) model (Ben-Akiva and Ramming, 1998, and Ben-Akiva and Bierlaire, 1999b), and we specify a utility function for each alternative i and observation n: $U_{in} = \beta_{\text{PS}} \ln \text{PS}_i^{\mathcal{U}} + \beta_{\text{L}}\text{Length}_i + \beta_{SB}\text{NbSB}_i + \varepsilon_{in}$, where $\beta_{\text{PS}} = 1$, $\beta_{\text{L}} = -0.3$, $\beta_{\text{SB}} = -0.1$ and ε_{in} is distributed Extreme Value with scale 1 and location 0. The PS attribute is defined by

$$\text{PS}_i^{\mathcal{U}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum\limits_{j \in \mathcal{U}} \delta_{aj}} \tag{5.10}$$

where Γ_i is the set of links in path i, L_a is the length of link a, L_i the length of path i and δ_{aj} equals one if path j contains link a, zero otherwise. Note that we explicitly index $\mathcal{U}$ to emphasize on which path set it is computed. 3000 synthetic observations have been generated by simulation, associating a choice with the alternative having the highest utility.

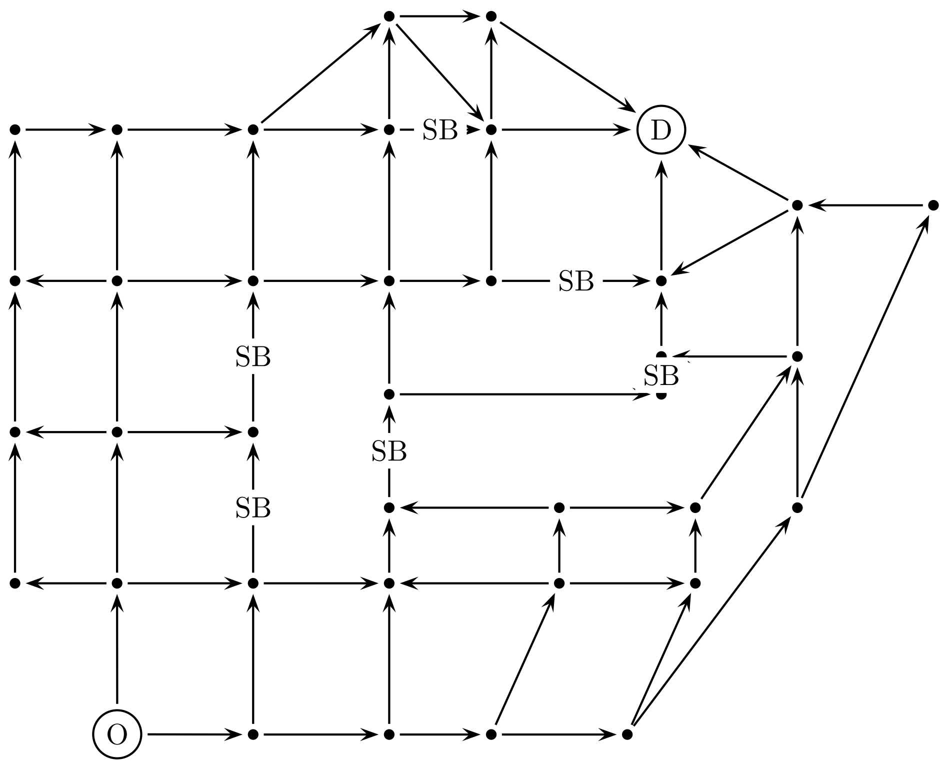

Figure 5.2: Example Network

		Sampling Correction	
		Without	With
Path Size	$\mathcal{C}$	$M^{\text{NoCorr}}_{PS(\mathcal{C})}$	$M^{\text{Corr}}_{PS(\mathcal{C})}$
	$\mathcal{U}$	$M^{\text{NoCorr}}_{PS(\mathcal{U})}$	$M^{\text{Corr}}_{PS(\mathcal{U})}$

Table 5.1: Model Specifications

5.3.2 Model Specifications

Table 5.1 present the four different model specifications that are used in order to evaluate both the PS attribute and the sampling correction. For each of these models we specify the deterministic term of the utility function as follows

$$\begin{aligned}
M_{PS(\mathcal{C})}^{\text{NoCorr}} \quad V_{in} = &\ \mu \left(\beta_{\text{PS}} \ln \text{PS}_{in}^{\mathcal{C}} - 0.3\text{Length}_i + \beta_{SB}\text{NbSB}_i\right) \\
M_{PS(\mathcal{C})}^{\text{Corr}} \quad V_{in} = &\ \mu \left(\beta_{\text{PS}} \ln \text{PS}_{in}^{\mathcal{C}} - 0.3\text{Length}_i + \beta_{SB}\text{NbSB}_i\right) + \ln(\frac{k_{in}}{q(i)}) \\
M_{PS(\mathcal{U})}^{\text{NoCorr}} \quad V_{i} = &\ \mu \left(\beta_{\text{PS}} \ln \text{PS}_{i}^{\mathcal{U}} - 0.3\text{Length}_i + \beta_{SB}\text{NbSB}_i\right) \\
M_{PS(\mathcal{U})}^{\text{Corr}} \quad V_{in} = &\ \mu \left(\beta_{\text{PS}} \ln \text{PS}_{i}^{\mathcal{U}} - 0.3\text{Length}_i + \beta_{SB}\text{NbSB}_i\right) + \ln(\frac{k_{in}}{q(i)}).
\end{aligned}$$

The PS attribute based on sampled paths is defined by

$$\text{PS}_{in}^{\mathcal{C}} = \sum_{a\in\Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j\in\mathcal{C}_n} \delta_{aj}}. \tag{5.11}$$

Note that the two first specifications are based on (5.11) and the two last on (5.10). β_{L} is fixed to its true value and we estimate μ, β_{PS} and β_{SB}. In this way the scale of the parameters is the same for all models and we can compute the t-tests w.r.t. the corresponding true values.

5.3.3 Estimation Results

For a specific parameter setting of the biased random walk algorithm (10 draws, Kumaraswamy parameters $b_1 = 5$ and $b_2 = 1$, length is used as generalized cost for the shortest path computations), we generate one choice set per observation and estimate the models. The corresponding estimation results are reported in Table 5.2. The t-test values show that only the model including a sampling correction and PS computed based on $\mathcal{U}$ ($M_{PS(\mathcal{U})}^{\text{Corr}}$) has unbiased parameter estimates.

The models including sampling correction have smaller variance of the random terms compared to the models without correction. (Recall that μ^2 is inversely proportional to the variance.) The standard errors of the parameter estimates are also in general smaller indicating more efficient estimates. Moreover, the model fit is remarkably better for the models with correction compared to those without. Despite of this the model with PS computed

	True PSL	$M^{\text{NoCorr}}_{PS(\mathcal{C})}$ PSL	$M^{\text{Corr}}_{PS(\mathcal{C})}$ PSL	$M^{\text{NoCorr}}_{PS(\mathcal{U})}$ PSL	$M^{\text{Corr}}_{PS(\mathcal{U})}$ PSL
$\beta_{\mathbf{L}}$ fixed	**-0.3**	**-0.3**	**-0.3**	**-0.3**	**-0.3**
$\widehat{\mu}$	**1**	**0.182**	**0.923**	**0.141**	**0.977**
standard error		0.0277	0.0246	0.0263	0.0254
t-test w.r.t. 1		-29.54	-3.13	-32.64	-0.91
$\widehat{\beta}_{\mathbf{PS}}$	**1**	**1.94**	**0.308**	**-1.02**	**1.02**
standard error		0.428	0.0736	0.383	0.0539
t-test w.r.t. 1		2.20	-9.40	-5.27	0.37
$\widehat{\beta}_{\mathbf{SB}}$	**-0.1**	**-1.91**	**-0.139**	**-2.82**	**-0.0951**
standard error		0.25	0.0232	0.428	0.024
t-test w.r.t. -0.1		-7.24	-1.68	-6.36	0.20
Final log likelihood		-6660.45	-6147.79	-6666.82	-5933.62
Adj. rho-square		0.018	0.093	0.017	0.125

Null log likelihood: -6784.96, 3000 observations
Algorithm parameters: 10 draws, $b_1 = 5$, $b_2 = 1$, $C(\ell) = L_\ell$
Average size of sampled choice sets: 9.66
BIOGEME (Bierlaire, 2007, and Bierlaire, 2003) has been used for all model estimations

Table 5.2: Path Size Logit Estimation Results

based on sampled choice sets ($M^{\text{Corr}}_{PS(\mathcal{C})}$) has biased estimates of the scale and PS parameters. Hence, these results support the hypothesis that the PS should be computed based on the true correlation structure, otherwise the attribute biases the results. In a real application it is however not possible to compute PS based on the true correlation structure since $\mathcal{U}$ cannot be explicitly generated. This is further discussed in the following section.

We now analyze the estimation results as a function of two of the biased random walk algorithm parameters: the Kumaraswamy distribution parameter b_1 and the number of draws. First we note from Figure 5.3 that, as expected, the number of generated paths increase with the number of draws but decrease as b_1 increase. Recall from Figure 5.1 that the higher the value of b_1 the more the biased random walk is oriented towards the shortest path. Figure 5.4 shows the absolute value of the t-tests w.r.t. the true values for the $M^{\text{Corr}}_{PS(\mathcal{U})}$ model. With one exception the parameters are unbiased for both 10 and 40 draws and for all values of b_1. (A line is shown at the critical value 1.96.) These results indicate that for this example the estimation results are robust w.r.t. to the algorithm parameter settings.

With only two exceptions, the other three model specifications ($M^{\text{NoCorr}}_{PS(\mathcal{C})}$,

$M^{\text{Corr}}_{PS(\mathcal{C})}$ and $M^{\text{NoCorr}}_{PS(\mathcal{U})}$) have biased estimates for at least one parameter for all values of b_1 and for all number of draws. $M^{\text{Corr}}_{PS(\mathcal{C})}$ has unbiased estimates for 5 and 10 draws when $b_1 = 30$. For these two models the parameter estimates are however not efficient having high standard deviations. The detailed t-test values are presented in the Appendix (from page 101).

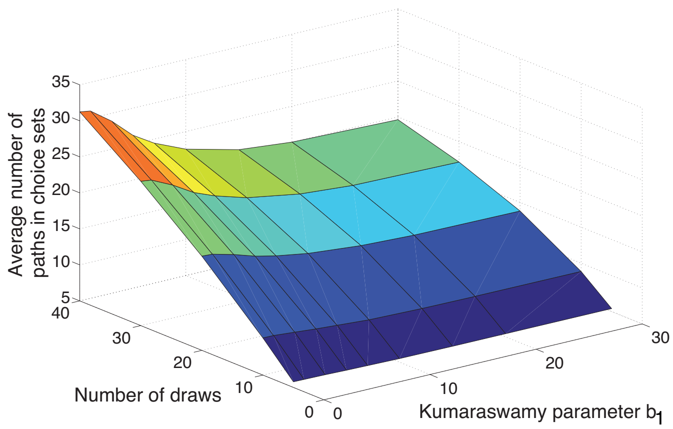

Figure 5.3: Average Number of Paths in Choice Sets

5.3.4 Heuristic for Extended Path Size

In a real application where $\mathcal{U}$ cannot be generated it is not possible to compute the PS attribute on the true correlation structure. It is however important to compute it based on a set of paths larger than the sampled set $\mathcal{C}_n$. It is therefore interesting to first study, for the previous example, how many paths are needed in order to obtain unbiased parameter estimates. Second, we propose an heuristic for computing a PS attribute that approximates the true correlation structure.

We generate an *extended choice set*, $\mathcal{C}_n^{\text{extended}}$, for each observation in the network shown in Figure 5.2. This choice set is only used for computing the PS attribute. In addition to all paths in $\mathcal{C}_n$ we randomly draw (uniform distribution) a number of paths from $\mathcal{U}\backslash\mathcal{C}_n$ and add these to $\mathcal{C}_n^{\text{extended}}$. The deterministic utilities for a model including sampling correction are now defined as

$$V_{in} = \mu\left(\beta_{\text{PS}} \ln \text{PS}_{in}^{\mathcal{C}^{\text{extended}}} - 0.3\text{Length}_i + \beta_{SB}\text{NbSB}_i\right) + \ln\left(\frac{k_{in}}{q(i)}\right) \quad (5.12)$$

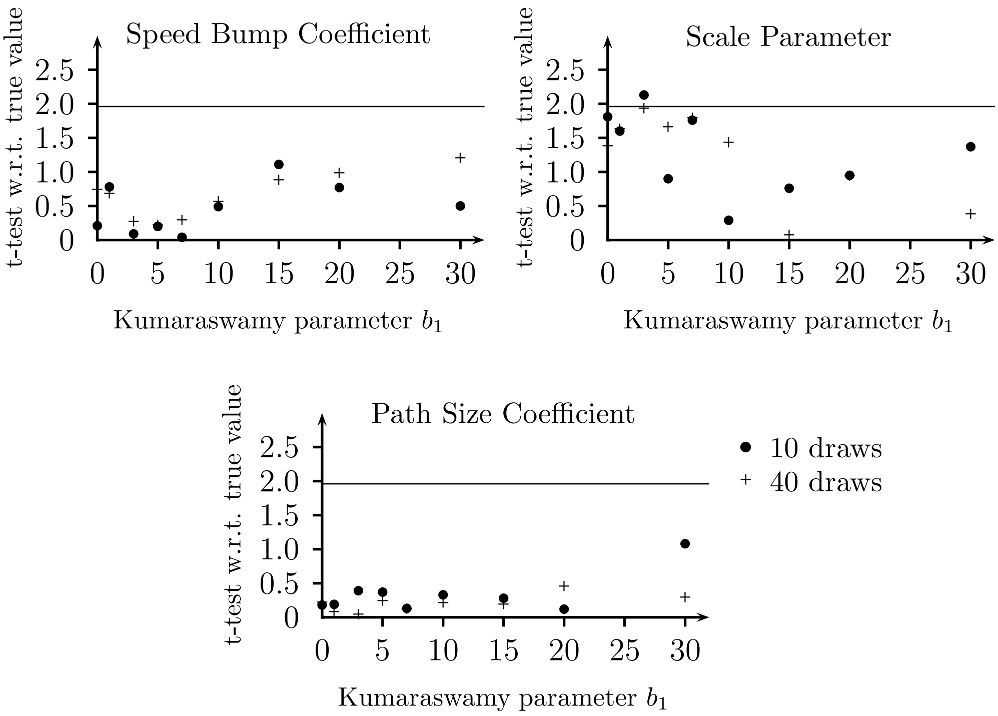

Figure 5.4: t-test Values w.r.t. True Values for the Coefficients of $M^{\text{Corr}}_{PS(\mathcal{U})}$

where

$$\mathrm{PS}_{in}^{\mathcal{C}^{\text{extended}}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in \mathcal{C}_n^{\text{extended}}} \delta_{aj}}.$$

The estimation results as a function of the average size of $\mathcal{C}_n^{\text{extended}}$ are shown in Figure 5.5 where the x-axis ranges from the average number of paths in $\mathcal{C}_n$ (9.66) up to $|\mathcal{U}| = 170$. For each parameter estimate we report the absolute value of the t-test w.r.t. its true value. An important improvement of the t-test value for the PS parameter can be noted after only 20 additional paths in $\mathcal{C}_n^{\text{extended}}$ and it is unbiased from 70 additional paths. The scale parameter is unbiased from 80 additional paths (except for the sample with 140 additional paths). Even though many paths (average number in $\mathcal{C}_n^{\text{extended}}$ approximately $0.5|\mathcal{U}|$) are needed in order for all parameter estimates to be unbiased, the estimates can be significantly improved by using an extended choice set for the PS computation.

Note that the purpose of the results presented in Figure 5.5 is to have an indication of the parameter estimates when the PS attribute is computed on more paths than those in $\mathcal{C}_n$. Each data point correspond to one random sample of paths. More samples would be needed in order to perform a deeper analysis, but this is already a clear indication on the need for using larger sets for computing the PS attribute.

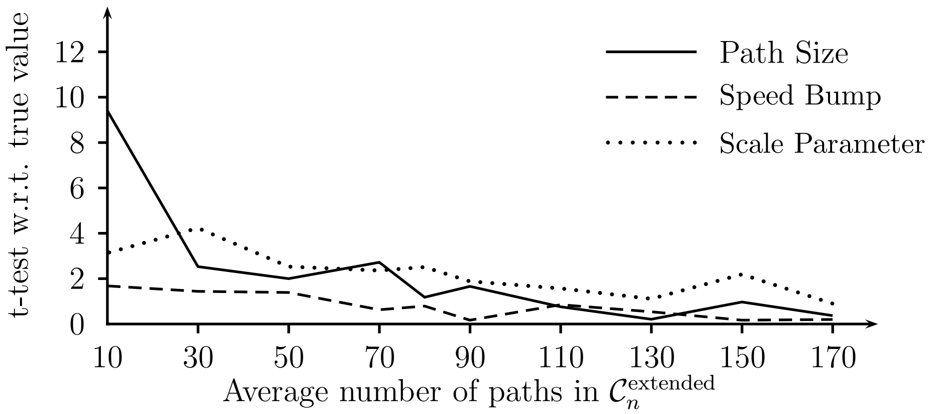

Figure 5.5: Estimation Results for Corrected Model as a Function of the $\mathcal{C}_n^{\text{extended}}$ Average Size

In order to use an extended choice set for the PS computation in a real network, we need to generate paths such that the true correlation structure is approximated. That is, the number of paths in the extended choice set using each link in the network should reflect the number of paths in $\mathcal{U}$ using each link. For this purpose we propose a *recursive gateway* algorithm that uses the general stochastic approach presented in Section 5.1. An extended choice set $\mathcal{C}^{\text{extended}}$ is defined for each origin-destination pair as follows:

- For each link in the network we generate a path and add it to $\mathcal{C}^{\text{extended}}$ if it is not already present.
- A path is generated by recursively drawing links based on probabilities defined by the Kumaraswamy distribution using Equation (5.2).
- In order to avoid selecting links scattered over the network, we update s_o, s_d, v and w in Equation (5.2) each draw so that higher probabilities are assigned to links close to already selected links than those further away.

The Extended PS attribute for alternative j and observation n is then computed based on $\mathcal{C}_n^{\text{extended}} = \mathcal{C}^{\text{extended}} \cup \mathcal{C}_n$.

We illustrate the heuristic with a small network in Figure 5.6 where we generate a path (dashed links in part IV) for link $(2, D)$ (bold link in part I). The weight for a link $\ell = (v, w)$ in the first iteration is given by

$$\omega(\ell) = \frac{SP(O,2)}{SP(O,v) + C(\ell) + SP(w,2)}$$

and the first link to be drawn is $(O, 3)$ (part II). The weights are then updated according to

$$\omega(\ell) = \frac{SP(3,2)}{SP(3,v) + C(\ell) + SP(w,2)}$$

where only one link is possible, namely $(3, 2)$ (part III).

The heuristic has been tested on the example network (Figure 5.2) and the average size of $\mathcal{C}_n^{\text{extended}}$ is 57 paths using $b_1 = 1$ and $b_2 = 1$ for the Kumaraswamy distribution. The estimation results, of deterministic utility specifications given by Equation (5.12), are reported in Table 5.3 where the reference model $M^{\text{corr}}_{\text{PS}(\mathcal{C})}$ from Table 5.2 is also shown. $\widehat{\mu}$ and $\widehat{\beta}_{\text{SB}}$ are unbiased and improved in $M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$ compared to $M^{\text{corr}}_{\text{PS}(\mathcal{C})}$. The PS coefficient is biased, this is however expected since $\mathcal{C}_n^{\text{extended}}$ is only an approximation of $\mathcal{U}$. Moreover, this approximation does not have the nice properties of a simple random sample and poorer $\widehat{\beta}_{\text{PS}}$ than the results reported in Figure 5.5 seems reasonable. Finally we note that the model fit is remarkably better for $M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$.

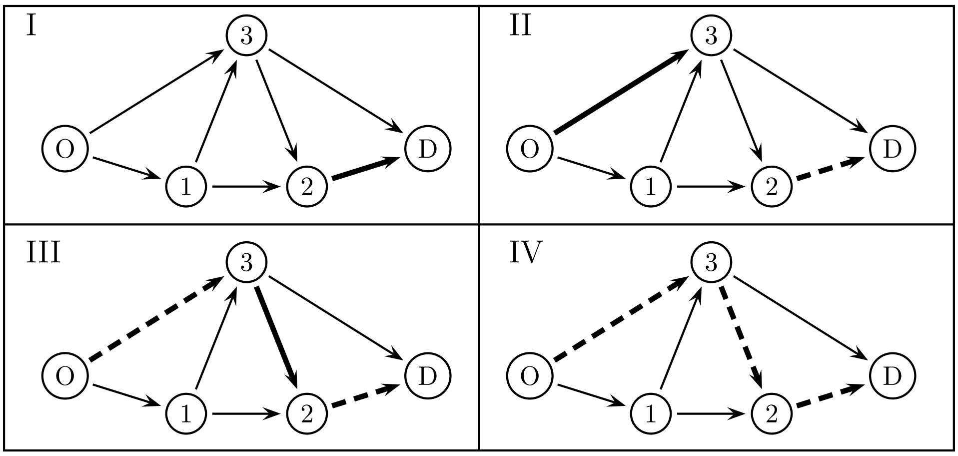

Figure 5.6: Illustration of Heuristic for Extended Path Size

	True PSL	$M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$ PSL	$M^{\text{corr}}_{\text{PS}(\mathcal{C})}$ PSL
$\widehat{\beta}_{\mathbf{L}}$ fixed	**-0.3**	**-0.3**	**-0.3**
$\widehat{\mu}$	**1**	**1.02**	**0.923**
Standard error		0.0275	0.0246
t-test w.r.t. 1		0.073	-3.13
$\widehat{\beta}_{\mathbf{PS}}$	**1**	**1.62**	**0.308**
Standard error		0.106	0.0736
t-test w.r.t. 1		5.85	-9.40
$\widehat{\beta}_{\mathbf{SB}}$	**-0.1**	**-0.076**	**-0.139**
Standard error		0.0253	0.0232
t-test w.r.t. -0.1		0.95	-1.68
Adj. Rho-Squared		0.113	0.093
Final Log-likelihood		-6015.94	-6147.79

Table 5.3: Estimation Results for Extended Path Size

5.4 Conclusions and Future Work

This chapter presents a new paradigm for choice set generation and route choice modeling. We view path generation as an importance sampling approach and derive a sampling correction to be added to the path utilities. We hypothesize that the true choice set is the set of all paths connecting an origin-destination pair. Accordingly, we propose to compute the Path Size attribute based on an approximation of the true correlation structure.

We present numerical results based on synthetic data which clearly show the strength of the approach. Models including a sampling correction are remarkably better than the ones that do not. Moreover, unbiased estimation results are obtained if the Path Size attribute is computed based on all paths and not on generated choice sets. This is completely different from route choice modeling praxis where generated choice sets are assumed to correspond to the true ones and Path Size (or Commonality Factor for the C-Logit model proposed by Cascetta et al., 1996) is computed on these generated path sets. Since it is not possible in real networks to compute these attributes on all paths we study how many paths are needed in order to obtain unbiased estimates and we propose an heuristic for generating *extended choice sets*.

It is important to note that the proposed sampling approach can be used with Multinomial Logit (MNL) based models (Path Size Logit and C-Logit). A consistent estimator for mixture of MNL (MMNL) models based on samples of alternatives does not exist but is available for Multivariate Extreme Value models (see Nerella and Bhat, 2004, for an empirical study of the bias in MMNL models when estimated on samples of alternatives).

Since the purpose of this chapter is to illustrate the proposed methodology, it is appropriate to use synthetic data for which the actual model is known. This allows to test the parameter estimates against their true values. A natural next step is to test the approach on real data.

Route choice models are often used for prediction. This has not been addressed in this chapter. Ben-Akiva and Lerman (1985) discuss prediction when samples of alternatives have been used for estimation. The correction for prediction is rather straightforward but it is important to take into account if chosen alternatives have been added to the sampled choice sets.

While we in this chapter focus on obtaining unbiased parameter estimates, the goal in prediction is to e.g. obtain representative traffic flows. It raises the issue of how many paths should be sampled for prediction. Furthermore, in a dynamic network as opposed to the static setting assumed here, the generalized cost values vary over time and consequently so do the sampling corrections. These two topics are important to investigate in future research.

Chapter 6

Adaptive Route Choice Models in Stochastic-Time Dependent Networks

This chapter is based on joint work with Moshe Ben-Akiva and Song Gao (Gao et al., 2007 and Gao et al., 2008).

Travel time and traffic conditions are inherently uncertain in transportation networks. Some of the uncertainty is introduced by design (e.g. traffic lights) and other due to disturbances such as incidents, vehicle breakdowns, weather conditions, special events and so forth. Real-time information in various formats are available, from personal observations, websites, variable message signs (VMS), radio broadcasts and cell phones to personal in-vehicle systems. Such information can reduce the uncertainty and therefore potentially help travelers make better route choice decisions. Information about traffic conditions can usually be obtained at various decision points during a trip, and route choice decisions can be updated according to the perception of the network state. This dynamic process of a series of route choices is of great interest since it is crucial for the evaluation of real-time information systems. Existing route choice models (such as the ones discussed in previous chapters) assume that travelers make their complete path choice at the origin. The fact that route choices can be adjusted during trips in response to revealed traffic conditions is therefore ignored. Throughout this chapter we refer to these models as *non-adaptive* path choice models.

In this chapter we study adaptive route choice models and the estimation of such models based on observations of chosen paths. Two types of adaptive route choice models are explored: an *adaptive path* model where a sequence of (non-adaptive) path choice models are applied at intermediate decision nodes, and a *routing policy* choice model where the alternatives correspond to routing policies rather than paths at the origin. (See Section 2.5 for a description of the routing policy concept.) We propose an estimator for the routing policy choice model and demonstrate the estimation feasibility based on synthetic data.

In the following section, we give a background on adaptive path choice and routing policy choice together with an example to illustrate the difference between the two. In Section 6.2 we present the model specifications and in Section 6.3 numerical results. Finally, we give some conclusions and discuss future work.

6.1 Background

The adaptive path choice model assumes that, at any given intermediate decision node, travelers choose a non-adaptive path from a choice set. When the traveler arrives at the next decision node (with random arrival time), he/she makes another choice out of a new set of non-adaptive paths, and so forth until the destination is reached. The choice sets as well as the path attributes are time dependent. This model therefore appears superior to a non-adaptive path choice model which ignores information on actual arrival times at intermediate nodes, but yet the choice is short-sighted. At each decision node, the next link is chosen based on a non-adaptive path, and thus the fact that travelers can be adaptive at subsequent decision points is not taken into account.

On the contrary, the routing policy choice model fully considers future adaptive choices. In this chapter we use the most simple definition of a routing policy assuming that the only available information to travelers are the arrival times at each node. The routing policy is therefore a mapping $(v, t) \to \ell$ from any node v at any arrival time t to next link $\ell \in \mathcal{E}_v$ where $\mathcal{E}_v$ is the set of outgoing links of node v.

In order to clarify the concepts, we illustrate in the following section adaptive path choice and routing policy choice with a small example. The same example is used for the numerical results presented in Section 6.3.

6.1.1 Illustrative Example

Figure 6.1 gives the topology of a small stochastic and time-dependent network. $\widetilde{T}_\ell$ denotes the random variable of the travel time on link ℓ, and T_ℓ a travel time. The travel times on links 0 and 1 at departure time $t = 0$ are random variables and a probability mass function (PMF) is given for each of the links. For example, the travel time for link 0 is x_0 with probability (w.p.) $1 - P_0$ and y_0 w.p. P_0. It is assumed that the two random variables, $\widetilde{T}_0$ and $\widetilde{T}_1$, are independent of each other. The travel times on links 2 and 4 are deterministic but dependent on the arrival time at the source node of the links. Links 3 and 5 both have deterministic travel times which are independent of the arrival times at the source nodes.

Travelers are going from A to D at departure time 0. The possible (node, time) pairs a traveler could encounter during the trip are $(A, 0)$, (B, x_0), (B, y_0), (C, x_1) and (C, y_1). Furthermore, the sets of outgoing links for each decision node are $\mathcal{E}_A = \{0, 1\}$, $\mathcal{E}_B = \{2, 3\}$ and $\mathcal{E}_C = \{4, 5\}$.

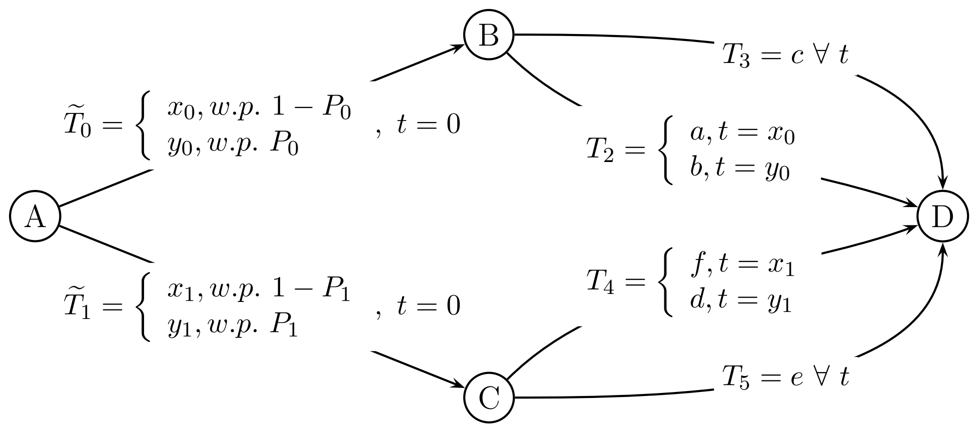

Figure 6.1: General Representation of Example Network

Recall that a routing policy is defined as a mapping from states to decisions. The concept therefore allows for mappings that are not interesting from a practical point of view. For this small example there are theoretically 2^5 routing policies since there are 5 possible (node, time) pairs and each pair can be mapped to two possible next links. However, once a traveler is at node B, the mapping at node C does not affect the remaining trip and therefore does not need to be specified. The same argument can be made at node C where the mapping at node B is not needed. There are hence 8 routing policies as shown in Figure 6.2. Note that a path is a special case of a routing policy such that the mapping from a (node, time) pair is the same regardless of the arrival time. For this example, four of the routing policies correspond to paths.

We now analyze a specific setting of this example, shown in Figure 6.3. The situation where link 0 or 1 has a travel time of x is referred to as the normal case and that where link 0 or 1 has a travel time of y as the incident

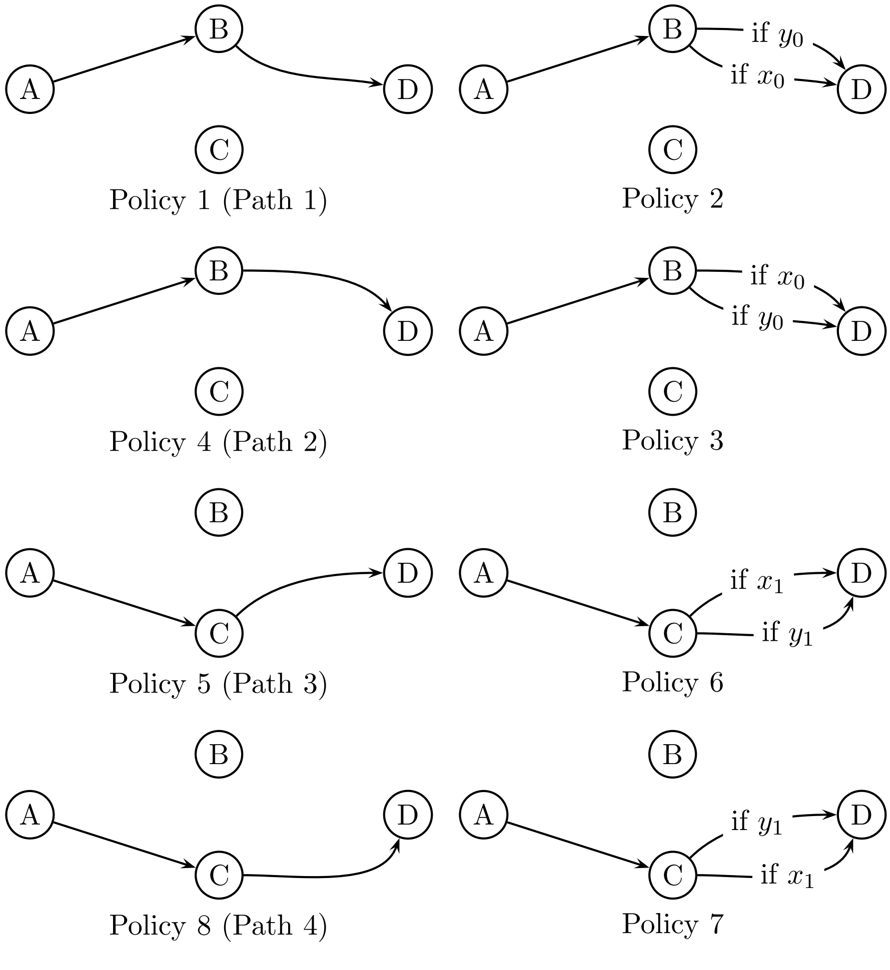

Figure 6.2: Routing Policies corresponding to Example Network

case. A later arrival time at node B (alternatively C) leads to longer travel time on link 2 (alternatively 4). This could be due to the fact that travelers who arrive late ($t = y$) are caught in peak traffic, while those with an earlier arrival ($t = x$) avoid it. Moreover, the peak traffic condition on link 4 is more severe than that on link 2. However, both links have diversions. Link 5 is the diversion of link 4 and is superior to link 3 which is the diversion of link 2.

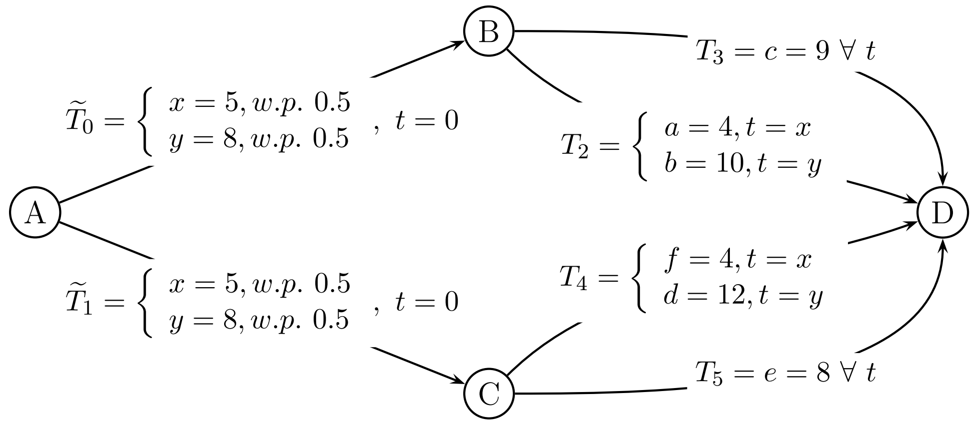

Figure 6.3: Specific Setting of Example Network

We now study the choice process for the two adaptive models in this network. A traveler is assumed to have a priori knowledge of the time-dependent link travel time PMFs of all links in the network before a trip starts. During the trip, he/she obtains additional online information on the actual arrival time at the second node (x or y). Depending on the arrival time, the traveler chooses the next link that minimizes his/her expected travel time.

Consider first the route choice process for an adaptive path model. At node A, four paths are available each with an expected travel time (ETT) as follows: ETT(Path1) $= (x + a + y + b)/2 = 13.5$, ETT(Path2) $= (x + y)/2 + c = 15.5$, ETT(Path3) $= (x + f + y + d)/2 = 14.5$ and ETT(Path4) $= (x + y)/2 + e = 14.5$. A traveler minimizing expected travel time takes link 0 which is the first link in path 1. The traveler then arrives at node B at either time x or y, each with probability 0.5. If the arrival time is x (off peak), the traveler takes link 2 with a travel time of $a = 4$, and if the arrival time is y (peak), the traveler takes the detour link 3 with a travel time of $c = 9$. The minimum expected travel time from node A to node D by making successive path choices is therefore $(x + a + y + c)/2 = 13$.

Next we consider the choice process for a routing policy model. At node

A, the traveler compares the attractiveness of links 0 and 1. The traveler knows that once arriving at the next node, the choice is based on realized arrival time and therefore considers all the possible diversions. The optimal routing policy from node B is to take the faster of links 2 and 3: if arrival time is x, take link 2 with a travel time $a = 4$ and if arrival time is y, take link 3 with a travel time $c = 9$. Similarly, the optimal routing policy at node C is to take the faster of links 4 and 5: if arrival time is x, take link 4 with a travel time $f = 4$ and if arrival time is y, take link 5 with a travel time $e = 8$. With this at hand, taking link 1 at node A is optimal since $(x + f + y + e)/2 = 12.5 < (x + a + y + c)/2 = 13$ (expected travel time of routing policy 6 compared to routing policy 2). Recall that the expected travel time for adaptive path choice is 13. The optimal routing policy is thus more efficient as a result of considering future adaptive possibilities.

We refer the reader to Gao and Chabini (2006) and Gao (2005) for a detailed discussion on optimal routing policy problems in stochastic time-dependent networks.

6.2 Model Specifications

In this section we present discrete choice model specifications for the previously discussed adaptive path and routing policy choice models. The models are designed for estimation based on path observations where each observation i is defined by an ordered set of links $\mathcal{I}_i$. We assume that the departure time as well as the arrival time at the source node of each link $\ell \in \mathcal{I}_i$ are known. Such information are available, for example, from Global Positioning System (GPS) data.

6.2.1 Adaptive Path Choice Model

Recall that this model assumes that a traveler chooses at the source node v of each observed link $\ell \in \mathcal{I}_i$ a path p from v to the destination. We therefore define an individual and time specific choice set $\mathcal{C}_{vtn}$ of paths from v to the destination. Hence, for each observation there are as many choice sets as there are links in the observed path.

The probability of an observation is defined as the product of the probabilities of choosing each link ℓ in the observed path, conditional on the arrival time t at v

$$P_n(i) = \prod_{\ell \in \mathcal{I}_i} P_n(\ell|t, v) = \prod_{\ell \in \mathcal{I}_i} \sum_{p \in \mathcal{C}_{vtn}} P(\ell|p)P(p|\mathcal{C}_{vtn}; \beta). \tag{6.1}$$

$P_n(\ell|t,v)$ is defined by the sum of the probabilities for each path that begins with ℓ. The path choice model $P(p|\mathcal{C}_{vtn};\beta)$, where β denotes the vector of parameters to be estimated, is therefore multiplied with a binary variable $P(\ell|p)$ that equals one if the first link in path p is ℓ and zero otherwise. Note that any existing non-adaptive path choice model can be used in this context.

The estimation of this type of model is rather straightforward. We view each link in the observed paths as an individual observation. Model (6.1) is then a special case of model (4.1) (developed for modeling network-free data in Chapter 4) and BIOGEME (Bierlaire, 2007) can be used for the estimation.

6.2.2 Routing Policy Choice Model

As opposed to the model for adaptive path choice which is sequential, the routing policy choice model is global. However, the choice of routing policy is latent and only the manifested path is observable. Although a traveler is not aware of the realized support point at the origin, we may assume that it is known to the modeler. Recall, that a support point is fully defined by the realized travel times on all random links. This information could be obtained through, for example, adequately dispersed GPS observations or probe vehicles that cover all random links. The probability of a path observation i of individual n, conditional on support point r and choice set of routing policies $\mathcal{G}_n$ is defined as

$$P(i|r,\mathcal{G}_n) = \sum_{\gamma \in \mathcal{G}_n} P(i|\gamma,r)P(\gamma|\mathcal{G}_n), \tag{6.2}$$

where γ is a routing policy. A routing policy is manifested as a path for a given support point. However, several different routing policies can be manifested as the same path. We therefore sum over all routing policies in $\mathcal{G}_n$ and multiply the routing policy choice model $P(\gamma|\mathcal{G}_n)$ with a binary variable $P(i|\gamma,r)$ that equals one if i corresponds to γ for support point r and zero otherwise. $P(\gamma|\mathcal{G}_n)$ can be modeled with the Policy Size Logit model (Gao, 2005) described in Section 2.5. In the same way as the adaptive path choice model, this model is a special case of (4.1) and BIOGEME can be used for the estimation.

For prediction the support point may be unknown to the modeler. In this case, a path cannot be unambiguously matched with a given routing policy. The model presented in Equation (6.2) can then be generalized to

$$P(i|\mathcal{G}_n) = \sum_{\gamma \in \mathcal{G}_n} P(i|\gamma)P(\gamma|\mathcal{G}_n) = \sum_{\gamma \in \mathcal{G}_n} \sum_{r=1}^{R} P(i|\gamma,r)P(r)P(\gamma|\mathcal{G}_n) \tag{6.3}$$

taking all possible support points into account.

6.3 Numerical Results

In this section we present numerical results of the proposed adaptive route choice models on a hypothetical network. There are two main objectives. First, to demonstrate the feasibility of estimating the two adaptive route choice models. Second, to gain insights into the adaptive route choice models by analyzing prediction results and comparing them with the results of a non-adaptive path choice model.

We use the example network shown in Figure 6.1 for which there are eight routing policies and four paths (see Figure 6.2). Three attributes can be computed for routing policies: expected travel time, standard deviation of travel time and Policy Size (PoS). The travel time for each routing policy is a random variable with two possible values. For example, the travel time of routing policy 6 is either $x_1 + f$ or $y_1 + e$, with probability $1 - P_1$ and P_1 respectively. The routing policies' expected travel time and standard deviation are therefore straightforward to compute.

The calculation of the PoS attribute is more involved because it is necessary to know how many routing policies use each link. Recall from Equation (2.7) that PoS is the expected value of Path Sizes over all support points of the random network

$$\mathrm{PoS}_{\gamma n} = \sum_{r=1}^{R} \left(\sum_{a \in \mathcal{I}_\gamma^r} \left(\frac{T_a^r}{T_\gamma^r} \right) \frac{1}{M_{an}^r} \right) P(r).$$

In order to compute M_{an}^r we define, for each support point, the path that corresponds to each routing policy. Since there are two random links in the network each with two possible realizations of travel times, there are altogether four support points. Let (T_0, T_1) represent a support point where T_0 and T_1 are realized travel times on links 0 and 1 respectively. The four support points are then (x_0, x_1), (x_0, y_1), (y_0, x_1), (y_0, y_1). As an example, routing policy 2 takes link 2 at node B for support point (x_0, x_1) and is therefore manifested as path 1. In the same way, we obtain the manifestation of all routing policies for support point (x_0, x_1): $\gamma_1 \to p_1$, $\gamma_2 \to p_1$, $\gamma_3 \to p_2$, $\gamma_4 \to p_2$, $\gamma_5 \to p_3$, $\gamma_6 \to p_3$, $\gamma_7 \to p_4$, $\gamma_8 \to p_4$. Hence, for this support point, four routing policies $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ use link 0, four routing policies $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ use link 1 and so forth. This is done for all support points and the PoS attribute can then be computed.

The non-adaptive and adaptive path choice models use a Path Size attribute given by Equation (2.1) that is computed based on expected travel times.

6.3.1 Observation Generation

A synthetic dataset of 6000 path observations is generated with a postulated model. Since the Policy Size Logit is the most complex model that best takes adaptive route choice behavior into account, we use this model to generate the observations. The probability of a routing policy γ is then defined as

$$P(\gamma|\mathcal{G}_n) = \frac{e^{V_{\gamma n}}}{\sum\limits_{k \in \mathcal{G}_n} e^{V_{kn}}} \tag{6.4}$$

where $V_{\gamma n} = \beta_{\text{PoS}} \ln \text{PoS}_{\gamma n} + \beta_{\text{ExpTime}}\text{ExpTime}_{\gamma n} + \beta_{\text{StdTime}}\text{StdTime}_{\gamma n}$ with $\beta_{\text{PoS}} = 1$, $\beta_{\text{ExpTime}} = -0.4$ and $\beta_{\text{StdTime}} = -0.1$. The choice set $\mathcal{G}_n$ contains the same eight routing policy alternatives for all observations but the link travel times vary.

Each path observation is generated in three main steps. First we sample link travel times (x_0, y_0, x_1, y_1, a, b, c, d, e and f) from a uniform distribution between 10 and 40 as well as link travel time probabilities (P_0 and P_1) from a uniform distribution $[0, 1]$. Second, we compute the probability $P(\gamma|\mathcal{G}_n)\ \forall\ \gamma \in \mathcal{G}_n$ using Equation (6.4) and randomly draw one routing policy that is labeled as chosen. Third, we sample a support point from the set of all support points based on their probabilities and associate a path with the chosen routing policy.

6.3.2 Estimation

Three models are estimated based on the generated path observations:

1. a routing policy model (6.2) with Policy Size Logit,
2. an adaptive path model (6.1) with Path Size Logit, and
3. a non-adaptive path model with Path Size Logit.

The deterministic utility functions have a linear-in-parameters specification of the same attributes as the true model including expected travel time, travel time standard deviation and a Path (or Policy) Size attribute. This corresponds to an ideal setting where all explanatory variables of the true utility function are included in the model. It is however convenient to validate the model estimation. For this case, the coefficient estimates in the routing policy model should not be significantly different from the postulated coefficient values.

The estimation results are shown in Table 6.1. The results confirm that a routing policy choice model can be estimated based on path observations. Indeed, the t-test values with respect to the true values are 0.07, -0.25 and -0.93 for $\widehat{\beta}_{\text{PoS}}$, $\widehat{\beta}_{\text{ExpTime}}$ and $\widehat{\beta}_{\text{StdTime}}$ respectively. The coefficient estimates of the other two models have their appropriate signs and are significantly different from zero.

In the following section we compare the three models in terms of prediction performance.

	Routing Policy	Adaptive path	Non-adaptive path
$\widehat{\beta}_{\text{PoS}}/\widehat{\beta}_{\text{PS}}$	1.03	1.23	2.75
std error	0.452	0.437	0.344
t-test	2.28	2.80	8.00
$\widehat{\beta}_{\text{ExpTime}}$	-0.402	-0.28	-0.265
std error	0.00805	0.00467	0.0049
t-test	-49.97	-60.00	-54.02
$\widehat{\beta}_{\text{StdTime}}$	-0.108	-0.071	-0.0451
std error	0.00857	0.00923	0.00643
t-test	-12.60	-7.69	-7.02
Final log likelihood	-3257.097	-3536.324	-3932.998
Adj. rho-square	0.608	0.574	0.527
Number of observed paths: 6000 Null log likelihood: -8317.766 BIOGEME Bierlaire, 2003, and Bierlaire, 2007, has been used for all model estimations			

Table 6.1: Estimation Results

6.3.3 Prediction

The three estimated models are applied to predict route choices in the same topological network, but with a fixed set of hypothetical link travel times as shown in Figure 6.4. The value of P is a parameter of the prediction test and varies from 0 to 1, with an increment of 0.1. Similar to the network setting discussed in Section 6.1.1, path 1 is the minimum expected travel time path. Links 2 and 4 have the same travel time under normal condition and link 4 is more congested than link 2 under incident condition. Moreover, links 3 and 5 are diversions for links 2 and 4, where link 5 is a better diversion than link 3.

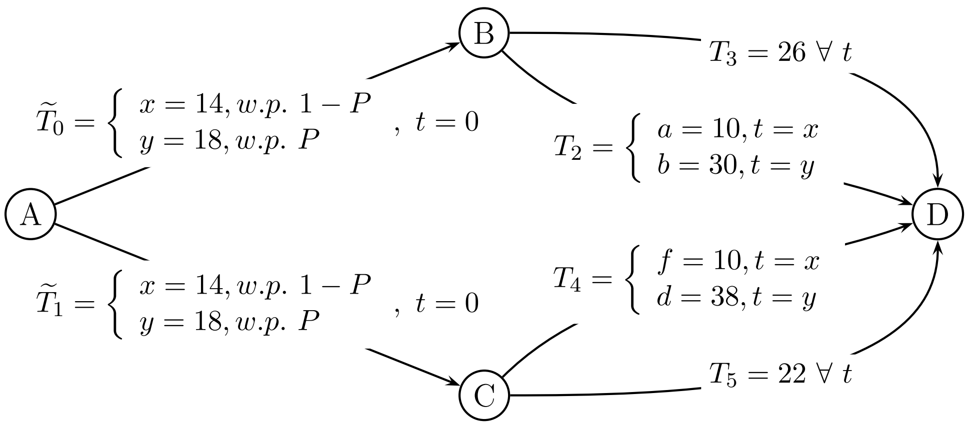

Figure 6.4: Network used for Prediction

Since the network is stochastic with all the support points known, we obtain distributions of variables such as path shares, path travel times, origin-destination travel time and so forth. We take expectations of these variables over the four support points, where the probability of each support is a function of P. In the following, we present the summary statistics (mean and/or standard deviation) to gain a clear understanding of the results.

Figure 6.5 shows the expected shares for each of the paths where the results from the three models are plotted as functions of incident probability P. Recall that paths 1 and 3 contain the links that can be affected by incidents due to the time-dependency of their travel times, while paths 2 and 4 contain the respective diversion links that are not affected by the incidents. It is therefore intuitive that, for all three models, the shares of paths 1 and 3 are decreasing functions of P, while shares of paths 2 and 4 are increasing functions of P.

In order to better appreciate the differences between the three models, we aggregate the results to obtain the expected shares for a left and a right turn at the origin. The shares of paths 1 and 2 respectively paths 3 and 4 are therefore put together and the corresponding results are reported in Figure 6.6. Recall that the right branch has a better diversion (link 5) than the left branch. In the routing policy model, as P increases, the importance of diversion becomes more important and consequently more flow goes to the right. In the two path based models, as P increases, the left share first increases and then decreases. This is because when $P = 0$, both paths 1

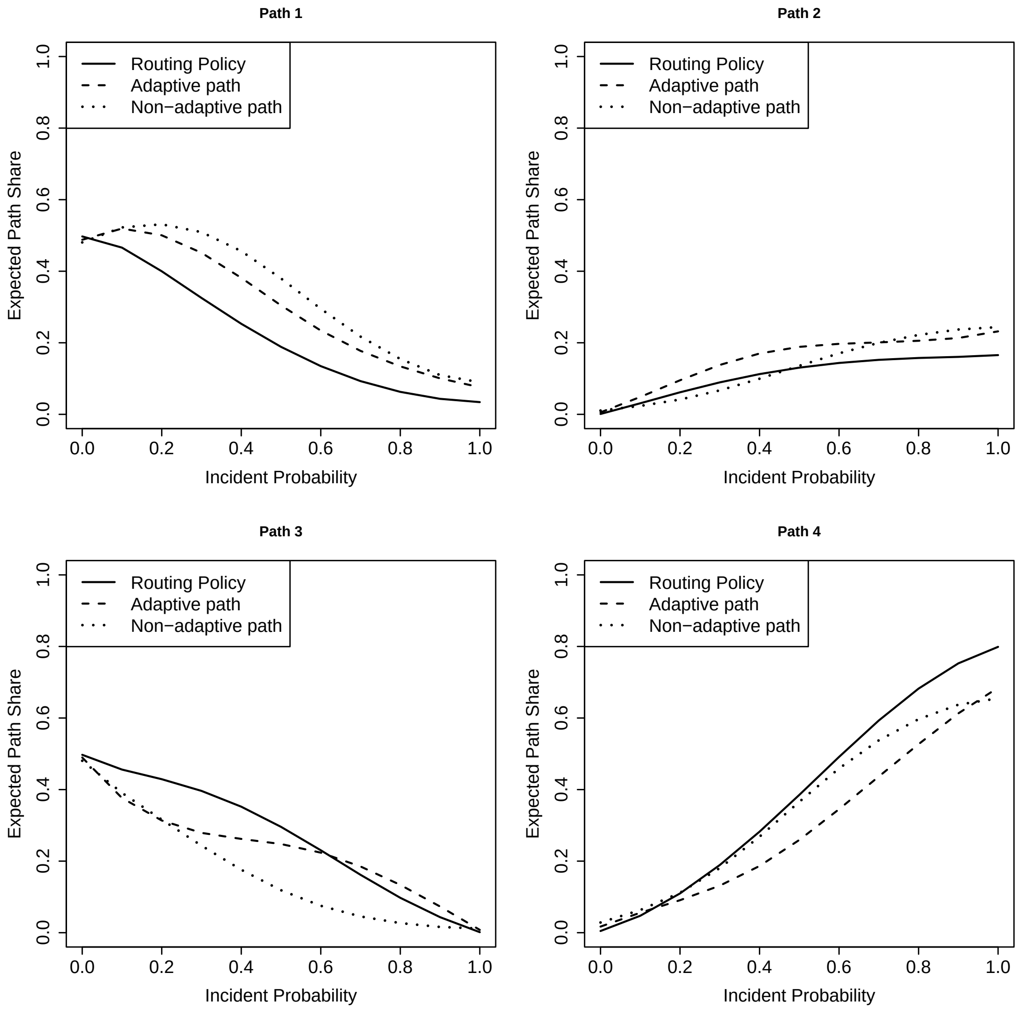

Figure 6.5: Expected Path Shares

and 3 (belonging to the left and right branches respectively) have the same minimum travel time and zero standard deviation. When P increases, path 1 has higher utility than path 3 and therefore gains a higher share. However, when P is larger than a certain value (approximately 0.3), path 4 has the highest utility, and thus the right share starts to increase.

Note that although the left-right shares of the non-adaptive and the adaptive path models are similar, the distribution of the flows at the second nodes are different. This is because the adaptive path model redistributes flows depending on the actual arrival times. Moreover, both the adaptive path and non-adaptive path model predict more flow taking the left branch than the routing policy model does. Future diversion possibility is not considered in either of the models and the branch with the least expected travel time (path 1) is therefore favored although link 3 is a worse diversion. In another words, the routing policy model better captures the option value of diversion than the adaptive path model.

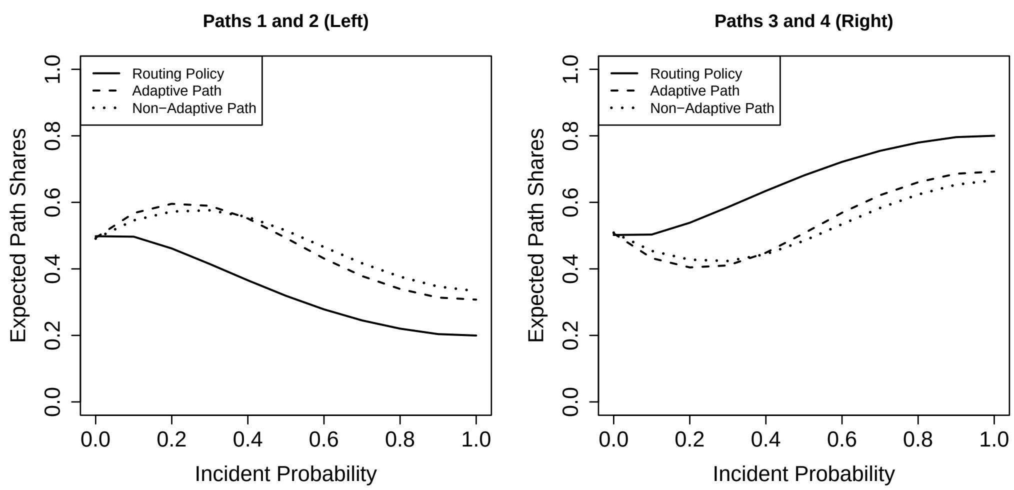

Figure 6.6: Expected Shares for Left and Right at Origin

Finally, we aggregate the results of all paths (weighted by path shares) and analyze the expected value and standard deviation of average path travel time. The results are reported in Figure 6.7 and show that, in terms of expected average travel time, the two adaptive models and the non-adaptive model are the most different when P is close to 0.5. This is in accordance with the intuition that being adaptive is of higher importance when the network is the most uncertain. We can also note that adaptive path and routing policy models have similar expected average travel time but their standard deviations are quite different. Since the adaptive path model predicts more flow to the left branch (worse diversion) than the routing policy model, it has longer travel time under incident condition but shorter travel time under normal condition. Hence both models predict roughly the same expected average travel time, but the adaptive path model has larger standard deviation.

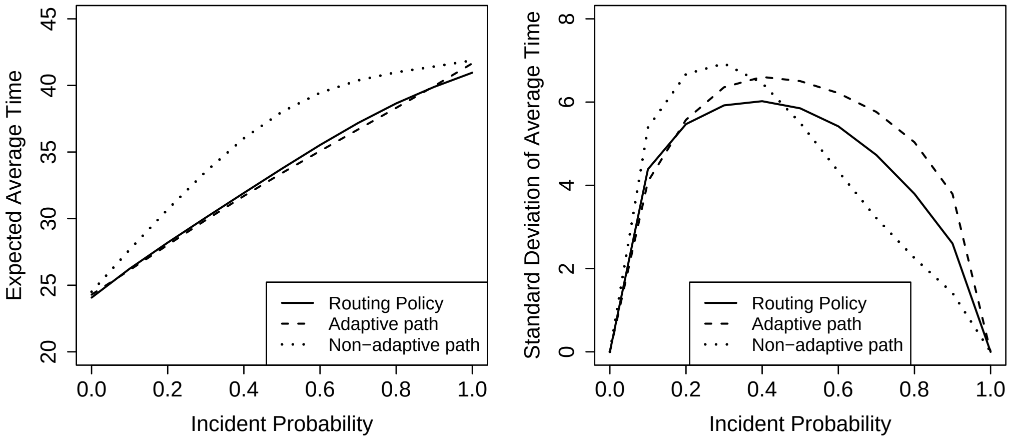

Figure 6.7: Average Time (Expected Value and Standard Deviation)

6.4 Conclusions

In this chapter we propose an estimator for a routing policy choice model and demonstrate the feasibility of estimating the model based on path observations. The concept of routing policy explicitly captures travelers' route choice adjustments according to information on realized network conditions in stochastic time-dependent networks. A routing policy can incorporate various assumptions about available information but in this chapter we assume that travelers only know the arrival times at intermediate nodes.

The routing policy choice model is compared to an adaptive path model which is a sequence of non-adaptive path choice models applied at intermediate decision nodes. Prediction results show that the routing policy model better captures the option value of diversion than the adaptive path model because of the foresight of a routing policy. The difference between the two adaptive models and the non-adaptive model is larger in terms of expected travel time in a highly stochastic network compared to a less stochastic network. As expected, this result indicates that being adaptive is of greater importance in uncertain networks.

6.5 Future Directions

The results presented in this chapter are very promising. A number of issues must however be addressed in order to estimate a routing policy choice model on data collected in a real network. We are currently working on the estimation based on the Borlänge GPS dataset described Chapter 3. In the following we discuss the issues related to making the routing policy model operational for real applications.

First of all, the model is based on a *stochastic and time-dependent network*. The definition of such a network is not straightforward and requires a considerable amount of data.

The path *observations* must at least contain information about departure time as well as the arrival times at each intermediate node. Such information is available for GPS data but not for reported trips. In order to fully exploit the capacity of the routing policy concept the travelers' information access is of importance. For example, whether they use an in-vehicle information system, listen to radio broadcasts or pass by variable message signs. Depending on the type of information access, the modeler can make hypotheses about the realized link travel times that are known to a given traveler at a given time. For instance, a variable message sign only gives information about downstream links while radio broadcasts can give information about links scattered in the network. For future route choice data collection, it would be interesting to combine passive monitoring of vehicles with a traffic information system that can store information about disturbances in the network as well as information that is broadcasted to travelers though various channels.

Routing policy *choice set generation* has not been discussed in this chapter but is an important issue. Gao (2005) proposes algorithms to compute the optimal routing policy under various information access assumptions. Such algorithms can be used in a similar way as shortest path algorithms for path choice set generation (see Section 2.2 for a review). Since the goal of this research is to study travelers' responses to real-time information, choice sets corresponding to different *information access* assumptions should be generated. In this context it is important to note that the complexity of the optimal routing policy algorithm is dependent on the information assumption.

Finally, we note that the routing policy choice model estimator proposed in this chapter does not consider choice sets based on different information access assumptions. This must be addressed in future research.

Chapter 7

Conclusions

This thesis is based on a collection of papers and the chapters are therefore rather independent of each other. In this chapter we give an overview of the main results and some directions for future research.

Correlation in route choice models is discussed in Chapter 3. First, we present an in-depth analysis of the commonly used Path Size Logit model and show how the Path Size attribute can be derived from the theory on aggregate alternatives. Several different formulations have been used in the literature and this research motivates the use of the original Path Size formulation, or the recently proposed Path Size Correction factor (Bovy, 2007a). Second, we propose the Subnetwork approach that allows the analyst to control the trade-off between the simplicity of the model and the level of realism. This model can be viewed as an extension to the one presented by Bekhor et al. (2002) but where we make the assumption that all links in the paths are not equally important for the correlation structure. The empirical results, based on a real GPS dataset, are very promising and show that this approach outperforms the more simple Path Size Logit.

The framework for modeling route choice with network-free data, presented in Chapter 4, addresses the data processing issue. The proposed framework allows the estimation of any existing route choice model based on original descriptions of trip observations that are defined by sequences of pieces of data. The concept that bridges the gap between the data and the model is called Domain of Data Relevance and corresponds to a physical area in the network where a given piece of data is relevant. First, we illustrate the approach on small examples for two types of data: reported trips and GPS data. Second, a real dataset of reported long distance trips in Switzerland is used for estimation. The precision of the route choice descriptions in this dataset is very low. Moreover, the network is to our knowledge the largest one used in the literature for route choice analysis based on revealed preferences data. These observations are therefore particularly challenging for the proposed framework. We present estimation results for Path Size Logit and Subnetwork models that illustrate the feasibility of the approach. Note that it is necessary to adapt the precision of the network to the precision of the trip descriptions. The framework allows to estimate models based on even low precision observations but can of course not add information that is not in the data.

The methodology and results presented in Chapter 5 provide a novel paradigm for choice set generation in particular and route choice modeling in general. One difficult aspect of route choice analysis is the generation of the choice sets. It is known that different choice sets lead to different estimation results, it is however difficult to analyze which choice set is the most appropriate. This problem is addressed in this thesis by a sampling approach. We consider that the true choice sets contain all paths for each origin-destination pair. We propose a stochastic path generation algorithm based on an importance sampling protocol and derive the corresponding sampling correction to be added to the path utilities. Estimation results of Path Size Logit models based on synthetic data clearly show the strength of this approach. Models including a sampling correction are remarkably better than the ones that do not. The results also show that the Path Size attribute should be computed based on the true correlation structure, that is the set of all paths. As opposed to route choice modeling practice where the Path Size attribute is computed based on generated choice sets only. Since this is not possible for a real application we propose a heuristic for computing an Extended Path Size attribute that approximates the true correlation structure.

It is important to note that the proposed sampling approach can be used with Multinomial Logit (MNL) based models (Path Size Logit and C-Logit). A consistent estimator for mixture of MNL (MMNL) models based on samples of alternatives does not exist (but is available for Multivariate Extreme Value models). The Subnetwork approach is therefore not appropriate in this context.

In Chapter 6 we study adaptive route choice behavior in stochastic and time-dependent networks and present an estimator for a routing policy choice model. The concept of routing policy explicitly captures travelers' route choice adjustments according to information on realized network conditions. Modeling travelers' response to en-route information is of great interest since it is crucial for the evaluation of real-time information systems. Optimal adaptive routing problems have been studied in the literature but the estimation of such choice models based on disaggregate revealed preferences data is a new area. We estimate two models capturing adaptive behavior based on a synthetic data set: an adaptive path model where a sequence of non-adaptive path choice models are applied at intermediate nodes, and a routing policy model. We compare the prediction performance of these models to the performance of a non-adaptive path model in an example network. The results show that the routing policy model better captures the option value of diversion than the adaptive path model because of the foresight of the routing policy concept. The difference between the two adaptive models and the non-adaptive model is larger in terms of expected travel time in a highly stochastic network compared to a less stochastic network. As expected, this result indicates that being adaptive is of higher importance in uncertain networks.

In the following discussion we focus on future work related to extensions and improvements of the research presented in this thesis. Since route choice modeling based on GPS data involves a considerable amount of data processing, we believe that the network-free data modeling approach is particularly suitable in this context. It would therefore be interesting to test the approach on a GPS dataset where the original logging coordinates are available. Such a dataset could also be map matched so that a comparison in terms of estimation and prediction results between the two approaches can be performed.

The proposed sampling approach for choice set generation questions the traditional route choice modeling process; not only how choice sets are modeled but also how the correlation among alternatives should be defined. This opens up for research questions that are not addressed, or only partly addressed, in this thesis. We use synthetic data to test the proposed methodology which is an important step since the true model is known and the bias in parameter estimates hence can be evaluated. It is however essential to test the stochastic path generation algorithm and the heuristic for the Extended Path Size in a real network. Moreover, a comparison of models estimated with, respectively without, sampling correction based on real data should be performed.

Sampling paths for prediction has not been discussed in this thesis and is an important topic for future research. A sampling correction for prediction must be derived. Moreover, it is necessary to evaluate the characteristics and number of generated paths to be included in the choice sets used for prediction.

Given the uncertainty related to travel times and traffic conditions in transport systems and that real-time information systems become more and more available, we believe that route choice modeling in this context is an important direction for future research. This thesis, based on joint work with Moshe Ben-Akiva and Song Gao, presents some pioneer work on the estimation of such models based on revealed preferences route choice data. This research is however still in its starting phase. As discussed in Section 6.5, several issues need to be addressed in order to make the models operational for real applications. The models are also data demanding and advanced techniques for collecting trip data and combining them with various information sources on traffic conditions need to be developed.

The contributions of this thesis to the state of the art in route choice modeling for uni-modal networks are manifold. The developed methodology is general and it would be interesting to investigate how it can be extended for modeling route choice behavior in other contexts.

A first, closely related field of research is route choice modeling in multimodal networks where we believe that most of the approaches presented in this thesis can potentially be used.

Models of pedestrian movements is another research area of great importance. The main difference with car route choice is that pedestrians are not limited to a network but rather in a two or three dimensional environment. It would be interesting to investigate how the network-free data modeling approach can be extended to this context.

Notations

Shorthand

CF	:	Commonality Factor
CNL	:	Cross-Nested Logit
DDR	:	Domain of Data Relevance
EC	:	Error Component
GEV	:	Generalized Extreme Value
GPS	:	Global Positioning System
IIA	:	Independence from Irrelevant Alternatives
i.i.d.	:	independent and identically distributed
LNL	:	Link-Nested Logit
MEV	:	Multivariate Extreme Value
MMNL	:	Mixture of Multinomial Logit
MNL	:	Multinomial Logit
MNP	:	Multinomial Probit
OD	:	Origin-destination pair
PDF	:	Probability density function
PMF	:	Probability mass function
PoS	:	Policy Size
PS	:	Path Size
PSC	:	Path Size Correction
PSL	:	Path Size Logit
VMS	:	Variable message sign
w.r.t.	:	with respect to

Notations

n	:	index for individuals
N	:	number of individuals in sample
i,j,p	:	path alternative (in Chapter 4 i refers to an observation and in Chapter 6 to a path observation)

$U_{in} = V_{in} + \varepsilon_{in}$	:	utility of alternative i for individual n
V_{in}	:	deterministic part of U_{in}
ε_{in}	:	error term of U_{in}
β	:	vector of unknown parameters
$\widehat{\beta}$	:	vector of parameters estimates
ν_n	:	vector of Extreme Value distributed random variables
μ	:	scale parameter of the Extreme Value distribution
ζ_n	:	vector of Normal distributed random variables
σ	:	covariance parameter of the Normal distribution
F_n	:	factor loading matrix for individual n
$P(i\|\mathcal{C}_n)$	:	probability of individual n choosing alternative i within choice set $\mathcal{C}_n$
φ	:	parameter in Generalized Path Size formulation, Equation (2.3)
L_{ij}	:	length (generalized cost) of links common to paths i and j
Γ_j	:	ordered set of all links in path j

Network

$\mathcal{E}$	:	set of links in a network
$\mathcal{V}$	:	set of nodes in a network
$\ell = (v, w)$	:	a link with source node v and sink node w
a	:	index for links
$\mathcal{E}_v$	:	set of outgoing links from node v
L_ℓ, $C(\ell)$	:	generalized cost of link ℓ
s	:	origin-destination pair
s_o	:	origin node
s_d	:	destination node
$\mathcal{S}$	:	set of origin-destination pairs

Choice sets and paths

$\mathcal{U}$	:	universal choice set of paths
J	:	number of paths in $\mathcal{U}$
$\mathcal{M}$	:	subset of $\mathcal{U}$
$\mathcal{M}_n$	:	subset of $\mathcal{U}$ specific to individual n
$\mathcal{H}_n$	:	all non empty subsets of $\mathcal{M}_n$
$\mathcal{C}_n(s)$	:	choice set of paths of individual n and origin-destination pair s, $\mathcal{C}_n$ is used if there is no ambiguity as to which s it refers to
J_n	:	number of paths in $\mathcal{C}_n(s)$
M_n	:	number of links in $\mathcal{C}_n(s)$
M_{an}	:	number paths in $\mathcal{C}_n(s)$ using link a

Network-free Data

d	:	piece of data
e	:	network element (link or node)
i	:	observation (sequence of data)
$\delta(d, e)$	:	indicator function; equals one if e is related with the DDR of d and zero otherwise
p	:	path alternative

Sampling of Paths

b_1, b_2	:	parameters of the Kumaraswamy distribution
$\omega(\ell \vert b_1, b_2)$	:	weight associated with a link ℓ given by the Kumaraswamy distribution
$q(\ell \vert \mathcal{E}_v, b_1, b_2)$	:	probability of selecting link $\ell = (v, w)$ given a set of outgoing links $\mathcal{E}_v$ and Kumaraswamy distribution parameters b_1 and b_2
$q(j)$	:	probability of generating (sampling) path j
$q(\mathcal{C}_n \vert j)$	:	probability of generating (sampling) choice set $\mathcal{C}_n$ given that j is the chosen path
$\widetilde{\mathcal{C}}_n$	:	set of sampled paths including duplicates
Ψ_n	:	number of draws for sampling $\widetilde{\mathcal{C}}_n$

$\widetilde{k}_{jn}$	:	number of times alternative j is sampled
k_{jn}	:	number of times alternative j is in $\widetilde{\mathcal{C}}_n$ considering that the chosen alternative is added after sampling
δ_{jc}	:	binary variable; equals one if $j = c$ and zero otherwise, where c denotes the index of the chosen alternative

Adaptive Route Choice

$\mathcal{T}$	:	set of time periods in dynamic network
$\mathcal{P}$	:	probabilistic description of link travel times in a stochastic network
t	:	time (period)
$\widetilde{T}_{\ell t}$	:	random variable of travel time for link ℓ and time t given by a probability mass function
r	:	support point
R	:	number of support points
$T^r_{\ell t}$	:	travel time of link ℓ at time t for support point r
ϱ_r	:	matrix of size $(\lvert\mathcal{T}\rvert \times \lvert\mathcal{E}\rvert)$ where each element corresponds to a travel time $T^r_{\ell t}$
I	:	information; set of realized link travel times
γ	:	routing policy alternative
$\mathcal{G}_n$	:	choice set of routing policies for individual n
$\mathcal{I}$	:	ordered set of links
i	:	path observation

Detailed Estimation Results for Path Sampling

The following tables show the absolute value of t-test values for the four different models discussed in section 5.3.

		Kumaraswamy parameter b_1			
Parameter	Nb. Draws	0	1	3	5
	5	24.68	21.99	17.12	6.65
	10	24.20	20.61	16.68	7.24
$\widehat{\beta}_{\text{SB}}$	20	21.31	18.10	12.76	7.71
	30	19.11	15.03	10.52	6.93
	40	15.99	14.17	8.92	5.89
	5	5.17	5.11	0.22	2.46
	10	5.08	3.98	2.18	2.20
$\widehat{\beta}_{\text{PS}}$	20	6.93	5.23	0.30	3.52
	30	6.93	3.93	0.22	3.28
	40	4.97	5.12	0.10	3.38
	5	0.66	6.52	18.7	29.35
	10	0.27	6.47	18.34	29.54
$\widehat{\mu}$	20	0.06	5.92	18.01	27.49
	30	0.53	5.75	17.45	26.51
	40	0.31	5.38	16.93	25.66

Table 7.1: Model $M_{PS(\mathcal{C})}^{\text{NoCorr}}$ (no convergence for $b_1 > 5$ due to $\widehat{\mu}$ close to zero)

Parameter	Nb. Draws	Kumaraswamy parameter b_1 0	1	3	5
$\widehat{\beta}_{\mathrm{SB}}$	5	28.02	24.67	18.92	5.63
	10	29.06	25.26	19.90	6.35
	20	28.38	24.93	18.78	8.20
	30	28.02	23.96	17.71	9.31
	40	26.81	22.88	16.47	9.83
$\widehat{\beta}_{\mathrm{PS}}$	5	36.35	28.19	15.18	5.34
	10	37.07	28.12	14.69	5.29
	20	35.01	25.84	12.05	3.98
	30	32.31	23.04	9.81	2.26
	40	29.17	20.50	7.80	0.94
$\widehat{\mu}$	5	3.06	4.54	19.25	31.3
	10	3.69	4.65	19.23	32.64
	20	3.56	4.43	19.68	32.41
	30	3.75	4.41	19.15	31.65
	40	3.37	4.38	18.77	30.99

Table 7.2: Model $M_{PS(\mathcal{U})}^{\mathrm{NoCorr}}$ (no convergence for $b_1 > 5$ due to $\widehat{\mu}$ close to zero)

Parameter	Nb. Draws	Kumaraswamy parameter b_1								
		0	1	3	5	7	10	15	20	30
	5	4.86	3.73	2.99	3.15	1.73	1.26	0.58	0.02	0.43
	10	2.81	1.32	2.47	1.68	0.91	0.32	1.60	1.60	1.66
$\widehat{\beta}_{\text{SB}}$	20	0.00	0.55	1.07	0.89	0.04	0.98	2.39	2.11	2.39
	30	1.51	2.88	2.77	1.86	2.75	1.44	2.35	2.58	3.58
	40	3.76	3.35	3.88	2.87	1.65	2.38	2.73	2.71	3.52
	5	8.49	8.16	5.73	6.53	7.41	5.98	4.60	4.39	1.47
	10	8.40	7.53	9.62	9.41	8.38	6.88	5.09	4.13	1.22
$\widehat{\beta}_{\text{PS}}$	20	9.00	8.64	7.70	8.30	10.47	9.74	7.27	5.87	4.14
	30	7.95	6.49	6.82	8.39	8.28	10.27	8.79	6.98	6.22
	40	5.06	7.01	5.90	7.62	10.44	10.39	9.53	8.40	7.49
	5	6.80	5.60	4.43	2.58	0.65	1.88	3.27	2.22	1.19
	10	5.71	5.20	5.47	3.13	3.28	0.55	1.28	1.56	1.55
$\widehat{\mu}$	20	4.23	4.71	6.32	6.71	5.11	2.87	1.59	0.54	1.33
	30	3.59	4.10	6.85	8.62	9.51	3.92	2.26	0.39	0.55
	40	2.91	3.91	7.44	9.92	8.46	6.98	3.23	0.86	0.80

Table 7.3: Model $M_{PS(\mathcal{C})}^{\text{Corr}}$

Parameter	Nb. Draws	Kumaraswamy parameter b_1								
		0	1	3	5	7	10	15	20	30
$\widehat{\beta}_{\mathrm{SB}}$	5	0.02	0.66	0.44	0.42	0.77	0.22	1.48	0.57	0.49
	10	0.21	0.78	0.09	0.20	0.04	0.49	1.11	0.77	0.50
	20	0.97	0.65	0.39	0.27	0.27	0.89	1.03	0.57	0.77
	30	0.66	0.66	0.31	0.21	0.46	0.73	0.91	0.69	1.16
	40	0.72	0.66	0.25	0.20	0.27	0.54	0.86	0.96	1.18
$\widehat{\beta}_{\mathrm{PS}}$	5	0.49	0.34	0.53	0.00	0.83	0.29	1.28	0.23	0.35
	10	0.18	0.19	0.39	0.37	0.13	0.33	0.28	0.12	1.08
	20	0.19	0.00	0.40	0.00	0.00	0.00	0.06	0.64	0.89
	30	0.19	0.00	0.20	0.14	0.00	0.00	0.17	0.80	0.16
	40	0.20	0.06	0.02	0.22	0.10	0.19	0.17	0.43	0.27
$\widehat{\mu}$	5	1.46	1.45	1.34	0.90	0.69	0.68	1.93	0.97	0.71
	10	1.81	1.60	2.13	0.90	1.76	0.29	0.76	0.95	1.37
	20	1.50	1.95	1.73	1.36	1.80	1.24	0.19	1.25	1.31
	30	1.62	1.65	1.86	1.62	1.69	0.85	0.21	1.34	0.33
	40	1.36	1.61	1.91	1.64	1.77	1.41	0.05	0.92	0.36

Table 7.4: Model $M_{PS(\mathcal{U})}^{\mathrm{Corr}}$

Bibliography

Abbé, E., Bierlaire, M. and Toledo, T. (2007). Normalization and correlation of cross-nested logit models, *Transportation Research Part B: Methodological* **41**(7): 795–808.

Abdel-Aty, M. A. and Abdalla, M. F. (2006). Examination of multiple mode/route-choice paradigms under ATIS, *IEEE Transactions on Intelligent Transportation Systems* **7**(3): 332–348.

Abdel-Aty, M. A., Kitamura, R., Jovanis, P. P., Reddy, P. and Vaughin, K. M. (1995). New approach to route choice data collection: Multiphase, computer-aided telephone interview panel surveys using geographic information systems data base, *Transportation Research Record* **1493**: 159–169.

Axhausen, K. W., Schönfelder, S., Wolf, J., Oliveira, M. and Samaga, U. (2003). 80 weeks of GPS traces: Approaches to enriching the trip information, *Transportation Research Record* **1870**: 46–54.

Azevedo, J., Costa, M. S., Madeira, J. S. and Martins, E. V. (1993). An algorithm for the ranking of shortest paths, *European Journal of Operational Research* **69**: 97–106.

Balakrishna, R. (2006). *Off-line Calibration of Dynamic Traffic Assignment Models*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Balakrishna, R., Ben-Akiva, M. and Koutsopoulos, H. N. (2007). Off-line calibration of dynamic traffic assignment: Simultaneous demand-supply estimation, *Transportation Research Record* **2003**.

Bekhor, S., Ben-Akiva, M. E. and Ramming, S. (2006). Evaluation of choice set generation algorithms, *Annals of Operations Research* **144**(1).

Bekhor, S., Ben-Akiva, M. and Ramming, M. (2002). Adaptation of logit kernel to route choice situation, *Transportation Research Record* **1805**: 78–85.

Bekhor, S. and Prato, C. G. (2006). Effects of choice set composition in route choice modelling, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Ben-Akiva, M. (1993). Lecture notes on large set of alternatives. Massachusetts Institute of Technology.

Ben-Akiva, M., Bergman, M., Daly, A. and Ramaswamy, R. (1984). Modeling inter urban route choice behaviour, *in* J. Vollmuller and R. Hamerslag (eds), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, VNU Science Press, Utrecht, Netherlands, pp. 299–330.

Ben-Akiva, M. and Bierlaire, M. (1999a). Discrete choice methods and their applications to short term travel decisions. Chapter for the Transportation Science Handbook, Preliminary Draft.

Ben-Akiva, M. and Bierlaire, M. (1999b). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Kluwer, pp. 5–34.

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice, *in* R. Hall (ed.), *Handbook of Transportation Science, 2nd edition*, Operations Research and Management Science, Kluwer, pp. 7–38. ISBN:1-4020-7246-5.

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H. N. and Mishalani, R. (2001). Network state estimation and prediction for real-time traffic management, *Networks and Spatial Economics* **1**(3-4): 293–218.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets, *International Journal of Research in Marketing* **12**: 9–24.

Ben-Akiva, M. and Bolduc, D. (1996). Multinomial probit with a logit kernel and a general parametric specification of the covariance structure, *Paper presented at the 3rd Invitational Choice Symposium*, Columbia University.

Ben-Akiva, M., Choudhury, C. and Toledo, T. (2006a). Lane changing models, *Proceedings of the International Symposium of Transport Simulation*, Lausanne, Switzerland.

Ben-Akiva, M., Choudhury, C. and Toledo, T. (2006b). Modeling latent choices: Application to driving behavior, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

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

Ben-Akiva, M. and Ramming, S. (1998). Lecture notes: Discrete choice models of traveler behavior in networks. Prepared for Advanced Methods for Planning and Management of Transportation Networks. Capri, Italy.

Ben-Akiva, M. and Watanatada, T. (1981). Application of a continuous spatial choice logit model, *in* C. F. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data and Econometric Applications*, MIT Press, Cambridge.

Bierlaire, M. (1998). *Discrete Choice Models*, Vol. 166 of *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, Springer, pp. 203–227.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transport Research Conference*, Ascona, Switzerland.

Bierlaire, M. (2007). An introduction to BIOGEME version 1.5. http://biogeme.epfl.ch.

Bierlaire, M., Bolduc, D. and McFadden, D. (to appear). The estimation of Generalized Extreme Value models from choice-based samples, *Transportation Research Part B: Methodological* . Accepted for publication on September 20, 2007, doi:10.1016/j.trb.2007.09.003.

Bierlaire, M. and Frejinger, E. (2007). Technical note: A stochastic choice set generation algorithm, *Technical report TRANSP-OR 070213*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Bierlaire, M. and Frejinger, E. (to appear). Route choice modeling with network-free data, *Transportation Research Part C: Emerging Technologies* . Accepted for publication on July 23, 2007, doi:10.1016/j.trc.2007.07.007.

Bogers, E., Bierlaire, M. and Hoogendoorn, S. (to appear). Modeling learning in route choice, *Transportation Research Record* . Accepted for publication.

Bolduc, D. (1999). A practical technique to estimate multinomial probit models in transportation, *Transportation Research Part B: Methodological* **33**(1): 63–79.

Bolduc, D. and Ben-Akiva, M. (1991). A multinomial probit formulation for large choice sets, *Proceedings of the 6th International Conference on Travel Behaviour*, Vol. 2, pp. 243–258.

Bovy, P. H. L. (2007a). Derivation of a new path size factor for use in MNL route choice modelling. Delft University of Technology, Internal report.

Bovy, P. H. L. (2007b). Modeling route choice sets in transportation networks: A preliminary synthesis, *Proceedings of the Sixth Triennial Symposium on Transportation Analysis (TRISTAN)*, Phuket, Thailand.

Bovy, P. H. L., Bekhor, S. and Prato, C. G. (2008). The path size factor revisited: an alternative derivation, *Paper 08-1469 accepted for presentation at the 87th Annual Meeting of the Transportation Research Board*, Washington, DC.

Bovy, P. H. L. and Fiorenzo-Catalano, S. (2006). Stochastic route choice set generation: behavioral and probabilistic foundations, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Bovy, P. H. L. and Stern, E. (1990). *Route choice: wayfinding in transport networks*, Kluwer Academic Publishers.

Burrell, J. (1968). Multipath route assignment and its application to capacity restraint, *Proceedings of the 4th International Symposium on the Theory of Road and Traffic Flow*.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks, *in* J. B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France*.

Cascetta, E. and Papola, A. (2001). Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, *Transportation Research Part C: Emerging Technologies* **9**(4): 249–263.

Cascetta, E., Russo, E., Viola, F. and Vitetta, A. (2002). A model of route perception in urban road network, *Transportation Research Part B: Methodological* **36**: 577–592.

Choudhury, C. (2007). *Modeling Driving Decisions with Latent Plans*, PhD thesis, MIT.

Chu, C. (1989). A paired combinatorial logit model for travel demand analysis, *Proceedings of the fifth World Conference on Transportation Research*, Vol. 4, Ventura, CA, pp. 295–309.

Daganzo, C. (1977). Multinomial probit and qualitative choice: A computationally efficient algorithm, *Transportation Science* **11**: 338–358.

de la Barra, T., Pérez, B. and Añez, J. (1993). Mutidimensional path search and assignment, *Proceedings of the 21st PTRC Summer Meeting*, pp. 307–319.

de Palma, A. and Picard, N. (2005). Route choice decision under travel time uncertainty, *Transportation Research Part A* **39**(4): 295–324.

Dial, R. (1971). A probabilistic multipath traffic assignment algorithm which obviates path enumeration, *Transportation Research* **5**(2): 83–111.

Dougherty, M. (1995). A review of neural networks applied to transport, *Transportation Research Part C: Emerging Technologies* **3**(4): 247–260.

Du, J. and Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues, *Transportation Research Part A* **41**(3): 220–232.

Fiorenzo-Catalano, S. (2007). *Choice Set Generation in Multi-modal Transportation Networks*, PhD thesis, Delft University of Technology.

Frejinger, E. (2004). *Route choice analysis using GPS data*, Master's thesis, École Polytechnique Fédérale de Lausanne.

Frejinger, E. (2007). Random sampling of alternatives in a route choice context, *Proceedings of the European Transport Conference*, Leiden, The Netherlands.

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

Friedrich, M., Hofsäss, I. and Wekeck, S. (2001). Timetable-based transit assignment using branch and bound, *Transportation Research Record* **1752**.

Gao, S. (2005). *Optimal Adaptive Routing and Traffic Assignment in Stochastic Time-Dependent Networks*, PhD thesis, MIT.

Gao, S. and Chabini, I. (2006). Optimal routing policy problems in stochastic time-dependent networks, *Transportation Research Part B: Methodological* **40**(2): 93–122.

Gao, S., Frejinger, E. and Ben-Akiva, M. (2007). Adaptive route choice models in stochastic time dependent networks, *Technical Report TRANSP-OR 070730*, Transport and Mobility Laboratory.

Gao, S., Frejinger, E. and Ben-Akiva, M. (2008). Adaptive route choice models in stochastic time dependent networks, *Paper 08-1349 accepted for presentation at the 87th Annual Meeting of the Transportation Research Board*, Washington, DC.

Gliebe, J., Koppelman, F. and Ziliaskopoulos, A. (1999). Route choice using a paired combinatorial logit model, *78th Annual Meeting of the Transportation Research Board*, Washington, DC.

Han, B. (2001). *Analyzing car ownership and route choices using discrete choice models*, PhD thesis, Department of infrastructure and planning, Royal Institute of Technology, Stockholm. TRITA-IP FR 01-95.

Han, B., Algers, S. and Engelson, L. (2001). Accommodating drivers' taste variation and repeated choice correlation in route choice modeling by using the mixed logit model, *80th Annual Meeting of the Transportation Research Board*.

Henn, V. (2000). Fuzzy route choice model for traffic assignment, *Fuzzy sets and systems* **116**: 77–101.

Hoogendoorn-Lanser, S. (2005). *Modelling Travel Behaviour in Multi-modal Networks*, PhD thesis, Delft University of Technology.

Hoogendoorn-Lanser, S., Van Nes, R. and Bovy, P. H. L. (2005). Path size and overlap in multi-modal transport networks, a new interpretation, *in* H. S. Mahmassani (ed.), *Flow, Dynamics and Human Interaction*, Proceedings of the 16th International Symposium on Transportation and Traffic Theory.

Jan, O., Horowitz, A. and Peng, Z. (2000). Using GPS data to understand variations in path choice, *Transportation Research Record* **1706**: 145–151.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes, *Journal of Hydrology* **46**: 79–88.

Lotan, T. (1997). Effects of familiarity on route choice behvior in the presence of information, *Transportation Research Part C: Emerging Technologies* **5**(3): 225–243.

Lotan, T. and Koutsopoulos, H. (1993). Route choice in the presence of information using concepts from fuzzy control and approximate reasoning, *Transportation Planning and Technology* **17**: 113–126.

Mahmassani, H. (2001). Dynamic network traffic assignment and simulation methodology for advanced system management applications, *Network and Spatial Economics* **1**: 267–292.

Mahmassani, H. S., Joseph, T. and Jou, R.-C. (1993). Survey approach for study of urban commuter choice dynamics, *Transportation Research Record* **1412**: 80–89.

Manski, C. F. (1977). The structure of random utility models, *Theory and decision* **8**: 229–254.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples, *Econometrica* **45**(8): 1977–1988.

Marchal, F., Hackney, J. and Axhausen, K. (2005). Efficient map matching of large GPS data sets - Tests on speed-monitoring experiment in Zurich, *Presented at the 84th Annual Meeting of the Transportation Research Board*, Washington, DC, USA.

Marcotte, P. and Nguyen, S. (1998). *Equilibrium and Advanced Transportation Modeling*, Kluwer Academic Publishers, chapter Hyperpath Formulations of Traffic Assignment Problems, pp. 175–199.

Marzano, V. and Papola, A. (2004). A path-multilevel logit model for route choice which allows implicit path enumeration, *Proceedings of the European Transport Conference*, Strasbourg, France.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial Interaction*

Theory and Residential Location, North-Holland, Amsterdam, pp. 75–96.

Morikawa, T. (1996). A hybrid probabilistic choice set model with compensatory and noncompensatory choice rules, *Proceedings of the 7th World Conference on Transport Research*, Vol. 1, pp. 317–325.

Murakami, E. and Wagner, D. (1999). Can using Global Positioning System (GPS) improve trip reporting?, *Transportation Research Part C: Emerging Technologies* **7**(2-3): 149–165.

Nerella, S. and Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models, *Transportation Research Record* **1894**: 11–19.

Nielsen, O. A. (2004). Behavioral responses to road pricing schemes: Description of the Danish AKTA experiment, *Journal of Intelligent Transportation Systems* **8**(4): 233–251.

Nielsen, O. A., Daly, A. and Frederiksen, R. (2002). A stochastic route choice model for car travellers in the Copenhagen region, *Networks and Spatial Economics* **2**: 327–346.

Nielsen, O. A. and Joergensen, R. (2004). Map-matching algorithms for GPS-data – methodology and test on data from the AKTA road pricing experiment in Copenhagen, *World Conference on Transport Research Society (WCTRS), Proceedings paper 1435*, Istanbul, Turkey.

Paag, H., Daly, A. and Rohr, C. (2002). Predicting use of the Copenhagen harbour tunnel, *in* D. Hensher (ed.), *Travel Behaviour Research: The Leading Edge*, Pergamon Press, pp. 627–646.

Polychronopoulos, G. and Tsitsiklis, J. N. (1996). Stochastic shortest path problems with recourse, *Networks* **27**(2): 133–143.

Prashker, J. N. and Bekhor, S. (1998). Investigation of stochastic network loading procedures, *Transportation Research Record* **1645**: 94–102.

Prashker, J. N. and Bekhor, S. (2004). Route choice models used in the stochastic user equilibrium problem: A review, *Transportation Review* **24**(4): 437–463.

Prato, C. G. (2004). *Latent Factors and Route Choice Behaviour*, PhD thesis, Politecnico di Torio.

Prato, C. G. and Bekhor, S. (2006). Applying branch and bound technique to route choice set generation, *Presented at the 85th Annual Meeting of the Transportation Research Board*.

Quddus, M. A., Ochieng, W. Y., Zhao, L. and Noland, R. B. (2003). A general map matching algorithm for transport telematics applications, *GPS Solutions* **7**(3): 157–167.

Ramming, M. (2001). *Network Knowledge and Route Choice*, PhD thesis, Massachusetts Institute of Technology.

Rilett, L. and Park, D. (2001). Incorporating uncertainty and multiple objectives in real-time route selection, *Journal of Transportation Engineering* **127**(6): 531–539.

Schönfelder, S., Axhausen, K., Antille, N. and Bierlaire, M. (2002). Exploring the potentials of automatically collected GPS data for travel behaviour analysis - A Swedish data source, *in* J. Möltgen and A. Wytzisk (eds), *GI-Technologien für Verkehr und Logistik*, number 13 in *IfGIprints*, Institut für Geoinformatik, Universität Münster, Münster, pp. 155–179.

Schönfelder, S. and Samaga, U. (2003). Where do you want to go today? - More observations on daily mobility, *3rd Swiss Transport Research Conference, Ascona*.

Swait, J. and Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B: Methodological* **21**(2): 91–102.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press.

Train, K., McFadden, D. and Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model for residential calling patterns and service choice, *The RAND Journal of Economics* **18**(1): 109–123.

Ukkusuri, S. V. and Patil, G. R. (2006). Exploring user behavior in online network equilibrium problems, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Van der Zijpp, N. J. and Fiorenzo-Catalano, S. (2005). Path enumeration by finding the constrained k-shortest paths, *Transportation Research Part B: Methodological* **39**(6): 545–563.

Van Nes, R., Hoogendoorn-Lanser, S. and Koppelman, F. (2006). On the use of choice sets for estimation and prediction in route choice, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Vovsha, P. and Bekhor, S. (1998). Link-nested logit model of route choice Overcoming route overlapping problem, *Transportation Research Record* **1645**: 133–142.

Vrtic, M., Fröhlich, P., Schüssler, N., Dasen, S., Erne, S., Singer, B., Axhausen, K. and Lohse, D. (2005). Erzeugung neuer Quell-/Zielmatrizen im Personenverkehr. Bundesamt für Strassen and Bundesamt für Verkehr, IVT ETH Zurich, Emch und Berger und TU Dresden, Final report to the Bundesamt für Raumentwicklung, Zürich.

Vrtic, M., Schüssler, N., Erath, A., Axhausen, K., Frejinger, E., Bierlaire, M., Stojanovic, S., Rudel, R. and Maggi, R. (2006). Including travelling costs in the modelling of mobility behaviour. Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Communications, IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.

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

Yai, T., Iwakura, S. and Morichi, S. (1997). Multinomial probit with structured covariance for route choice behavior, *Transportation Research Part B: Methodological* **31**(3): 195–207.

Yamamoto, T., Kitamura, R. and Fujii, J. (2002). Drivers' route choice behavior: Analysis by data mining algorithms, *Transportation Research Record* pp. 59–66.

Zito, R., D'Este, G. and Taylor, M. A. P. (1995). Global positioning systems in the time domain: How useful a tool for intelligent vehicle-highway systems?, *Transportation Research Part C: Emerging Technologies* **3**(4): 193–209.

Emma FREJINGER

Avenue de Senalèche 27	Born on March 26, 1979
1009 Pully, Switzerland	Nationality: Swedish
E-mail: emma.frejinger@gmail.com	Languages: Swedish, French, English, German (basic)
Homepage: transp-or2.epfl.ch/frejinger	

Research Assistant (2004 - 2008)

Research assistant and PhD candidate at the Swiss Federal Institute of Technology, Lausanne (EPFL): 2004-2006 in the Operations Research Group (ROSO) and 2006-2008 in the Transport and Mobility Laboratory (TRANSP-OR).

During the fall of 2006, visiting researcher at Massachusetts Institute of Technology (MIT) in the Intelligent Transportation Systems Laboratory directed by Prof. Moshe Ben-Akiva.

Research projects

▷ **Thesis topic: Development of Route Choice Models**
Supervisor: Prof. Michel Bierlaire
Project sponsored by the Swiss National Science Foundation (May 2005 to April 2008)

▷ **Analysis of Travel Costs in Mobility Behavior Models**
A collaboration with Prof. Kay Axhausen and Prof. Rico Maggi. Project sponsored by the Swiss Federal Road Office (November 2005 to February 2007).

Development of Case Studies

▷ **Workbook on discrete choice models**
Case studies covering seven topics related to discrete choice analysis based on eight datasets from different fields. A collaboration with the MIT ITS Laboratory, Dr. Gianluca Antonini, Dr. Carmine Gioia and Dr. Michaël Thémans. The workbook is used in MIT and EPFL courses (undergraduate and postgraduate).

▷ **Computer Assisted and Project-Oriented Teaching**
Case studies for teaching optimization to students in mechanical engineering. Project sponsored by the EPFL funding program for teaching and learning (April 2004 to March 2006).

Teaching

▷ **Postgraduate courses**
Discrete Choice Analysis: Predicting Demand and Market Shares (one week yearly)
Responsibilities (2005-2007): lab sessions in the end of each day, coordination of course material and workbook (case studies)
Modeling and Simulation in Logistics (International Institute for the Management of Logistics, IML), one week twice a year. Responsibilities (2005-2007): Coordination and planning, teaching assistant each day, lecturer one day.

▷ **Teaching assistant EPFL undergraduate courses**
Five courses in optimization, operations research and discrete choice modeling.

▷ **Supervision of student projects**
Two master projects of students in mathematics and eight semester projects of students in mathematics and communication systems.

Education

2008	**Doctor of Philosophy in Mathematics** Swiss Federal Institute of Technology, Lausanne
2004	University degree obtained from Linköping Institute of Technology, Sweden: **Master of Science in Industrial Engineering and Management** – Specialization in optimization and computer science Master's thesis: **Route Choice Analysis using GPS Data** Supervisor: Michel Bierlaire, Swiss Federal Institute of Technology, Lausanne
2001 – 2003	**Master's Program in Industrial Engineering**, as an exchange student, at the Swiss Federal Institute of Technology, Lausanne
1999 – 2001	**Bachelor's Program in Industrial Engineering**, Linköping Institute of Technology, Sweden
1999	Pre-university studies, scientific program, Skara, Sweden
1996 – 1997	Exchange student with scholarship, Argelès-Gazost, France

Internships

▷ **Summer Internship Program for Female Undergraduate Students (2002)**
offered by National Center of Competence in Research, Mobile Information and Communication Systems (NCCR-MICS). Development (in Java programming language) of a simulator.

▷ **Non-academic Temporary Positions**
Sales representative for wholesale market, 2 months during summers 2000–2001, Swedish Meats, Skara, Sweden
Accountant's Assistant, 1 month during winter 2000, Swedish Meats, Skara, Sweden
Spinning (indoor biking) Instructor, 1999–2000, 2-3 times a week in the university sports organization, Linköping, Sweden
Secretary, 2 months during summers 1998–1999, Swedish Meats, Skara, Sweden

Computer skills

Languages	Java, Ada 95, C++, SQL, JSP, XML, XSLT, HTML, AMPL, LaTeX
Software	MS Office, TransCAD, MatLab, Mathematica, Adobe Illustrator
Operating systems	Mac OS X, Windows, Linux, Unix

List of Publications

▷ **Papers in International Journals**

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

Bierlaire, M., and Frejinger, E. (to appear). Route choice modeling with network-free data, Transportation Research Part C: Emerging Technologies, doi:10.1016/j.trc.2007.07.007.

▷ **Conference Proceedings**

Frejinger, E. (2007). Random Sampling of Alternatives in a Route Choice Context, Proceedings of the European Transport Conference, The Netherlands.
*Received the **Neil Mansfield Award** by the Association for European Transport for the best paper by a sole author aged 35 or under.*

Frejinger, E. and Bierlaire, M. (2007). Random Sampling of Alternatives for Route Choice Modeling, Proceedings of the Swiss Transport Research Conference, Ascona, Switzerland.

Frejinger, E. and Bierlaire, M. (2007). Stochastic Path Generation Algorithm for Route Choice Models, Proceedings of the Sixth Triennial Symposium on Transportation Analysis (TRISTAN), Phuket Island, Thailand.

Bierlaire, M., Frejinger, E. and Stojanovic, J. (2006). A latent route choice model in Switzerland, Proceedings of the European Transport Conference, Strasbourg, France.

Frejinger, E. and Bierlaire, M. (2006). Capturing correlation with subnetworks in route choice models, Proceedings of the 11th International Conference on Travel Behaviour Research, Kyoto, Japan.

Frejinger, E. and Bierlaire, M. (2006). Capturing correlation in route choice models using subnetworks, Proceedings of the 6th Swiss Transportation Research Conference, Ascona, Switzerland.

Bierlaire, M. and Frejinger, E. (2005). Route choice models with subpath components, Proceedings of the 5th Swiss Transportation Research Conference, Ascona, Switzerland.

▷ **Technical Reports**

Frejinger, E., and Bierlaire, M. (2007). Sampling of Alternatives for Route Choice Modeling. Technical report TRANSP-OR 071121. Transport and Mobility Laboratory, ENAC, EPFL.

Gao, S., Frejinger, E. and Ben-Akiva, M. (2007). Adaptive Route Choice Models in Stochastic Time Dependent Networks, Technical report TRANSP-OR 070730. Transport and Mobility Laboratory, ENAC, EPFL.

Bierlaire, M. and Frejinger, E. (2007). Technical note: A Stochastic Choice Set Generation Algorithm. Technical report TRANSP-OR 070213. Transport and Mobility Laboratory, ENAC, EPFL.

▷ **Project Report**

M. Vrtic, N. Schussler, A. Erath, K.W. Axhausen, E. Frejinger, M. Bierlaire, S. Stojanovic, R. Rudel and R. Maggi (2006). Including travelling costs in the modelling of mobility behaviour, Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Communications, IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.