Estimating route choice models using low frequency GPS data

MASOUD FADAEI OSHYANI

Master’s Thesis
Supervisor: Anders Karlström
Documentation page

Title Estimating route choice models using low frequency GPS data

Keywords Route choice, Indirect inference, Shortest Path, Auxiliary model, Multinomial logit, Binding function

Author Masoud Fadaei Oshyani
masoodfo@kth.se

Committee members

Prof. Anders Karlström
Royal Institute of Technology (KTH)
Division for Transport and Location Analysis (TLA)
anders.karlstrom@abe.kth.se

Dr. Marcus Sundberg
Royal Institute of Technology (KTH)
Division for Transport and Location Analysis (TLA)
marcus.sundberg@abe.kth.se

Date of publication September 26, 2011
Acknowledgments
First I would like to sincerely appreciate my professor Anders Karlström’s kind considerations and being an all-time seconder of mine and his being everything that a supervisor is supposed to be. His unsparing efforts to help me find the way to solve my research problems are unutterable. He has always been an inspirer for me and has encouraged me even when the results have not been satisfying. His patience made me enthusiastic to ask all my questions and I was always sure that he knew the answer. He was the one to grant me the opportunity of conducting my research in TLA. It is a really great honor for me to be a member of a research group conducted by a smart and erudite leader. I couldn’t imagine any better start than this for my career in .Net programming and I will forever be grateful to him because of this favor.

I also wish to thank Marcus Sundberg for all his useful comments helping the improvement of results during my thesis. And I’m very grateful for all the valuable discussions with Per Olsson. His patience in helping me solve programing problems is admirable. He always carries a smile on his face which makes the working environment very peaceful and pleasant.

I would also like to thank all the members of TLA, VTI and CTS who make a very friendly atmosphere especially Shiva and Xiang whom I share the office with. During this period not only did I conduct my research, but I also worked as a member of a really friendly team and felt how interesting working at a transport modeling department was. Special thanks to professor Haris N. Koutsopoulos for allowing me to have access to Stockholm’s navigation data and all his helpful comments during my thesis.

Finally I would like to thank my wife. She was always there for me with her support. She realized how important this thesis was for me and patiently withstood the conditions when I almost always came back home late at night.
Abstract

GPS data are increasingly available to be used in transportation planning. Route choice models are estimated to address the behavior of individuals choosing a route in a given network. When data is collected with low frequency, it is unknown which path was traversed between the GPS data points. Furthermore, GPS data has measurements error. In this thesis we design an algorithm to consistently estimate a given route choice model in the presence of sparse GPS data and measurement errors.

We present an extension on a new method presented by Kalström et al. (2011) to estimate a route choice model. This method focuses on a given simple way to estimate the true parameter of a model. For this purpose the indirect inference method is employed as a structured procedure. In our context, a simple multinomial logit model is used as the auxiliary model with the simulated data sets and in a structured way returns the estimated parameter.

This version of discrete choice model is simple and fast which qualifies it as an appropriate auxiliary model. We estimate a model with random link costs which allows for a natural correlation structure across paths and is also useful for simulating paths in order to make choice sets.

In this study Monte Carlo evidence is provided to show the feasibility and accuracy of the proposed algorithm using a real world network from Borlänge, Sweden.

The main conclusion is that indirect inference is an exciting option in the tool box for route choice estimation which can be used for estimating route choice models using low frequency GPS sampling data.
Contents

Contents vi

List of Figures viii

List of Tables ix

1 Introduction 1
  1.1 Background ................................................. 1
  1.2 Research purpose ........................................... 2
  1.3 Research scope & limitations ............................... 3
  1.4 Research context ............................................ 4
  1.5 Report outline .............................................. 4

2 Literature Review 7
  2.1 Route Choice Models ........................................ 7
    2.1.1 Multinomial Logit ..................................... 9
    2.1.2 Multinomial Probit (MNP) ................................ 12
    2.1.3 Multivariate Extreme Value (MEV) .................... 12
    2.1.4 Error Component (EC) .................................. 13
  2.2 Estimation .................................................. 13
  2.3 Sources and Data Collection ............................... 15
  2.4 Choice Set Generation ..................................... 16
  2.5 Evaluation of Generated Choice Sets ....................... 21

3 Methodology 23
CONTENTS

3.1 Introduction ................................................. 23
3.2 The Model ................................................. 24
3.3 Indirect inference ............................................ 26
3.4 Universal choice set generation ....................... 30
   3.4.1 Monte Carlo (MC) view ........................... 30
   3.4.2 Sampling of Alternatives ......................... 31
3.5 Detailed description of the estimator ............... 31

4 Results ......................................................... 35
4.1 Estimation and Sensitivity analysis ................. 35
   4.1.1 Number of observations ......................... 36
   4.1.2 Choice set size .................................. 38
   4.1.3 Number of sample points for the OLS .......... 39
   4.1.4 Sensitivity with regard to beta zero ............ 41
   4.1.5 Frequency of GPS data .......................... 42
4.2 On computational efficiency ......................... 44
   4.2.1 Binding function ................................. 44
   4.2.2 Estimation with more stages .................... 45
4.3 Estimation for five OD-pairs ......................... 47

5 Conclusions .................................................. 49

Bibliography ................................................... 51
List of Figures

3.1 The applied II estimator: overview ........................................... 27
3.2 Unmapped Simulated GPS Points to Nodes ................................. 29
3.3 Binding Function: 0 to 5 ....................................................... 33

4.1 Sensitivity Analysis: Number of Observations ............................ 37
4.2 Sensitivity Analysis: Choice Set Size ....................................... 39
4.3 Sensitivity Analysis: Number of Sample Points ......................... 40
4.4 Sensitivity Analysis: Different beta zero .................................... 42
4.5 Sensitivity Analysis: GPS Points Frequency ............................... 43
4.6 Binding Function: 0 to 100 ..................................................... 44
4.7 Improvement in results through stages ...................................... 45
List of Tables

4.1 Estimation Results: Different Number of Observations . . . . . . . . . . 37
4.2 Estimation Results: Different Choice set Size . . . . . . . . . . . . . 38
4.3 Estimation Results: Different Number of Sample Points . . . . . . . 39
4.4 Estimation Results: Different Beta Zero Coefficient . . . . . . . . . 41
4.5 Estimation Results: Different frequency of GPS data . . . . . . . . . 43
4.6 Estimation Results: 3 stages with 3 sample points . . . . . . . . . . 46
4.7 Estimation Results: one OD to five ODs . . . . . . . . . . . . . . . . . 47
Chapter 1

Introduction

1.1 Background

Currently, transportation plays an essential role in many people’s everyday life. The purpose of the trips varies such as going to work, picking up children, shopping and so on. Different aspects of traveling such as the chosen path, the purpose, the destination, the time and mode of transport are in focal point of interest in order to analyze the travel behavior. Route choice models are applied in order to analyze travelers’ behavior regarding their preference of choosing routes.

Travelers’ preference for choosing routes can be based on different characteristics like distance, road type, travel time, cost or number of traffic lights. Other parameters which have effects on the result of route choice models is individuals’ characteristics, such as gender, age, income and etc.

In route choice analysis the main concern is to identify which route would be chosen by travelers in a transportation network. A greater understanding of route choice would be useful for predicting behavior under different scenarios. For example, a project for implementing a congestion charging system may be considered. Travelers which use routes crossing the charging gates should accept a more costly trip compared to previous situation. In return, their travel time may decrease. Route choice models can be used to analyze the
congestion charging scenario and predict future travelers’ behavior.

Route choice models are also a powerful tool in Vehicle Routing Problem (VRP), where a number of vehicles need to choose their routes in order to distribute their products among different destination nodes. Another application of route choice models is in traffic assignment concept where the more knowledge about route choice and the factors that influence the network user would be beneficial to develop an appropriate traffic assignment.

Route choice models can provide valuable knowledge for GPS device manufacturers whereas they would be able to suggest more efficient route to their customers to get to their destinations.

1.2 Research purpose

Currently, the department of transport science has got access to a database which contains valuable data from taxis’ GPS navigators in Stockholm. All taxis send their location to this database frequently. The average time interval between every two transferred signals is about 60 seconds. Since in this period of time, a taxi may pass a long distance and in most cases there are several paths between two consecutive collected points, this 60-seconds interval of data collecting may be categorized as low frequency GPS data collection.

These GPS records report the vehicle ID, time-stamp of measurement and a binary variable which shows the status of the taxi and carries the value of one if the taxi is in service and zero otherwise. The database containing such these fields, provides a great opportunity to do vast researches on the travelers behavior in Stockholm, however the low frequency GPS data has two obvious challenging characteristics from the route choice modeling point of view.

The first, in the most presented route choice models in literatures, known paths are considered as observed data, whereas the reported GPS data records consists of sets of points. In addition, it is unknown which path was traversed between the phrase GPS data points when low frequency data is collected. The Second problem is that GPS data has measurements error which should
be considered in proposed route choice models.

Moreover, there is an ongoing research on the route choice modeling in transport and location analysis division focusing on Borlänge city. A number of studies have been done in order to estimate the parameters of the cost function assigned to the links in the network. Previous researches considered travel time and the existing speed bump on roads as cost function attributes and were based on the real observed paths whereas in this study the crucial assumption is we do not have paths rather, trips data are presented as low frequency GPS points.

The main objective of this study is to propose a consistent estimator for a route choice model based on phrase GPS data. The proposed estimator may then be applied to data such as the data Stockholm region.

### 1.3 Research scope & limitations

In this study we develop a method for a route choice model with low frequency GPS points data. We provide Monte Carlo evidence in order to validation our proposed estimator. In other words, we use simulated data as real data. Then we apply our estimator on this real data and try to retrieve the original parameters of the generated data. Results from method validation are useful to evaluate the quality, reliability and consistency of analytical results.

A simple and fast algorithm has been designed in order to generate GPS points of the desirable frequency on the simulated paths with specified link cost function parameters $\beta$. In this case we just consider the link length parameter ($\beta = \beta_l$). The procedure of converting a route to GPS points with a specific frequency is explained in section 3.3.

The objective of this research is to test the possibility of using low frequency GPS data for a proposed method to estimate route choice models. The implemented method is based on the indirect inference method to estimate parameters of a true model (which is difficult to estimate) considering an auxiliary model (which is easy to estimate). The research is meant to achieve
CHAPTER 1. INTRODUCTION

the answers to these two questions:

- How is it possible to estimate a route choice model based on low frequency GPS point data?
- Is the predicted value from the proposed algorithm accurate?

In this study we use mapped GPS points, in other words they are simulated in a way that they will be located on the links of a proposed digital map. As a further research area GPS points noises and the errors in mapping them on digital maps can be considered in our estimation process.

1.4 Research context

The study is conducted in the division of ”Transport and location analysis” at KTH university. A research group is working on route choice modeling and this study is an effort to contribute to the ongoing research. The proposed model in this paper has been applied to the same network (Borlänge) before, considering the same assumptions. The main difference of this study is that it considers the trip data collected from the low frequency GPS data sets, whereas the previous studies were based on real and exact path observations.

1.5 Report outline

This research is structured as follows. Chapter 2 presents a literature review on the concept of route choice and contains discussions regarding the choice set generation and route choice modeling. In chapter 3 the proposed model is explained in different steps; simulating the GPS points, the definition for the auxiliary model, computing the binding function, applying the indirect inference based estimator and detailed explanations of the algorithm.

In the section of chapter 4, the final results of running the model for different situations are presented. These results contain statistical parameters
1.5. REPORT OUTLINE

regarding the estimated value of $\beta_l$ (the attribute parameter of the link length) with different number of observations, choice set size, number of sample points (for estimating binding function), number of OD-pairs and periods of GPS points collecting. Sensitivity of the model is analyzed in this section regarding changes in the mentioned parameters of the model. During this discussion, a heuristic method will be presented leading to a significant reduction in the computation time, whereas the reliability of results will be kept.

Finally, in chapter 5 the conclusions from the findings in the previous chapters are discussed and an outline on the possibilities for further research is given.
Chapter 2

Literature Review

2.1 Route Choice Models

According to the definition by Bierlarie, M. et al. (2005), the route choice problem is a discrete choice problem with the following main characteristics. In general the universal choice set is very large and people do not consider all possible choices. Furthermore the presence of overlapping paths cause high correlation among some alternatives which has to be captured during the study.

A set of general fundamental concepts can present the framework for a discrete choice model (Ben-Akiva and Lerman, 1985). The basic features that will be discussed are listed below:

1. The decision-maker: Who the decision-maker is and what characteristics he/she has.

2. The alternatives: The feasible options that the decision maker might face.

3. The attributes: The pros, cons and costs of each alternative.

4. The decision rule: The process that the decision-maker has undergone to evaluate and choose an alternative.
Train (2003) introduces three requirements as an essentiality of the route choice problem’s fitting within the framework of discrete choice models. The first is that, choices must be unique from the decision maker’s point of view and passengers choose just one alternative from the set. Secondly, all possible choices must be included in the choice set and the decision maker must be able to choose her/his desirable choice from the set. Finally, the choice set must be finite.

In the route choice models, utility is a term referring to the total satisfactions received by users from passing a route to get their destination from their origin nodes. Utility is often modeled to be affected by characteristics of the decision maker and specifications of the selected choice. Decision makers can increase their utility by minimizing generalized cost of their travels may consider travel distance, time, number of traffic lights on the path and etc.

There are different kinds of route choice models and we have tried to give an overview of route choice models in general with special emphasize on the ones which have been used in real applications and discuss pros and cons of each model. We will examine some technical details of the underlying discrete choice models. As the brief reviews, Ben-Akiva and Bierlaire (2003) give a comprehensive description of discrete choice methods, and Prashker and Bekhor (2004) focus on route choice models used in stochastic equilibrium problems.

The path-based approach using discrete choice methods are elaborated in a large proportion of the literatures. A main problem with path-based approach is that the universal choice set is very large, and often unknown. Different formulations have addressed this problem like C-Logit approach.

In response to the problem mentioned above, multivariate extreme value (MEV) models have been developed. These are all logit-based models in which the correlation across paths is represented by introducing a nesting structure. Another approach is to use link specific errors, such that the errors are associated to the links. It is assumed that the link costs are random and the summation of link costs gives the cost of the paths.
In practice, Karlström et al. (2011) use a link based cost model and assign the random errors to the links rather than the paths whereas they use the trip data in path format. This is the approach implemented in this study as well where low frequency GPS point sampling is used.

### 2.1.1 Multinomial Logit

The Multinomial Logit (MNL) model is considered as a simple model but with restrictive assumptions in order to meet the condition of the error term’s being identically and independently distributed (i.i.d.). It means that the unobserved factors should be uncorrelated over alternatives, as well as having the same variance for all alternatives. The popularity of the MNL model is due to the i.i.d., whereby may be achieved an easy form of the model in term of the choice probability. Nevertheless, this assumption can be inappropriate in some situations where unobserved factors related to one alternative might be similar to that related to another alternative. (Train, 2009)

However, the assumption of independence does not hold in the context of route choice due to overlapping paths. Overlap in route choice is a popular problem in travel analysis leading alternatives are no longer independent. Indeed, overlapping among alternatives causes to a statistical correlation between alternatives that should be considered for estimation.

Train (2003) explains that, as a main assumption, decision makers maximize their utility through choice selecting. For individual $n$ who chooses alternative $i$, the utility is denoted by $U_{i}^{n}$, $i = 1, \cdots, I$. Alternative $k$ is chosen if and only if $U_{k}^{n} > U_{i}^{n}, \forall i \neq k$. The amount of this utility is known to individuals but not to researchers. The main point is that, only a couple of all the attributes affecting the utility are observable and possible to measure.

In the route choice concept, for a proposed OD-pair $(r, s)$ there is a set of unique routes $I_{rs}$ connecting these two points. The general utility function that user $n$ chooses route $k$ from the choice set is defined by

$$U_{k}^{n} = V_{k}^{n} + \epsilon_{k}^{n}$$  \hspace{1cm} (2.1)
where $V^n_k$ is the representative of the deterministic part of the utility and $\epsilon^n_k$ represents the parameters which have effects on the utility but are not included in $V^n_k$. Based on these assumptions, the probability for individual $n$ to choose alternative $k$ is:

$$P^n_k = \text{Prob}(U^n_k > U^n_i, \forall i \in I_{rs})$$

$$= \text{Prob}(V^n_k + \epsilon^n_k > V^n_i + \epsilon^n_i, \forall i \in I_{rs})$$

$$= \text{Prob}(\epsilon^n_k - \epsilon^n_i > V^n_i - V^n_k, \forall i \in I_{rs}) \quad (2.2)$$

Consequently

$$P^n_k = \text{Prob}(\epsilon^n_k > \epsilon^n_i + V^n_i - V^n_k, \forall i \in I_{rs}) \quad (2.3)$$

As already mentioned, in logit models $\epsilon$s are assumed to be independently, identically distributed and following Gumbel distribution. Then the most popular Multinomial Logit model for route choice modeling is:

$$P^n_k = \frac{e^{V^n_k}}{\sum_i e^{V^n_i}} \quad (2.4)$$

In order to relax the restriction mentioned above by making a deterministic correction of the utility for overlapping paths in route choice models, lots of efforts have been made.

For instance, Cascetta et al. (1996) were the first to suggest such a deterministic correction. To clarify, they defined a factor called Commonality Factor (CF), in the deterministic part of the utility obtaining the C-Logit model.

$$V^n_k = V^n_k - CF_k \quad (2.5)$$

Regarding this modification the logit choice probability equation will redefine as:

$$P^n_k = \frac{e^{V^n_k - CF_k}}{\sum_i e^{V^n_i - CF_i}} \quad (2.6)$$
2.1. ROUTE CHOICE MODELS

the Commonality factor of path \( k \) is proportional to its overlap with other paths in the choice set belonging to \( I_{rs} \). One possible way to specify this factor is

\[
CF_k = \beta_o \ln \sum_{i \in I_{rs}} \left[ \frac{L_{ik}}{L_i^{1/2} L_k^{1/2}} \right]^{\gamma}
\]

(2.7)

where \( L_{ik} \) is the length of links shared among path \( i \) and \( k \), since \( L_i \) and \( L_k \) are summation of link lengths of paths \( i \) and \( k \). \( \gamma \) is a positive parameter.

Cascetta et al. (1996) present three different formulas to calculate CF value, but there is no guidance which of the formulations to use. Thus, Ben-Akiva and Ramming (1998) and Ben-Akiva and Bierlaire (1999) propose the Path Size Logit (PSL) model. The idea is like the C-Logit model. They define a new factor and add it to the deterministic part of the utility. This factor is called Path size (PS) factor and used to correct utility in order to relax the restriction which is caused by overlapping paths. The original PS utility formulation for path is

\[
U^n_i = V^n_i + \beta_{PS} \ln PS^n_i + \epsilon^n_i
\]

(2.8)

for individual \( n \) and path \( i \) where \( V^n_i \) is the deterministic part of utility function and \( \epsilon^n_i \) notes random part. The \( PS^n_i \) attribute is defined as

\[
PS^n_i = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}}
\]

(2.9)

where \( \Gamma_i \) is the set of all links of path \( i \), \( L_a \) and \( L_i \) represent the length of link \( a \) and path \( i \). \( \delta_{aj} \) is one if link \( a \) is located on path \( j \) and zero otherwise. \( C_n \) denotes the considered choice set.

Ramming (2001) compares the results of the C-Logit and PSL models with a different formulation. Having realized some flaws of the MNL model, the researchers started to seek for some other models which result in more complex models. However, rather few of these models have been implemented on real size networks with large choice sets.
2.1.2 Multinomial Probit (MNP)

The main characteristic of this model is that the error terms are normally distributed which permits an arbitrary covariance structure specification (Burrell, 1968, and Daganzo, 1977). It is well applicable for simulation when utilities are link additive. However its evaluation requires a great deal of computational time. Thus, it is less applicable for real applications with very large networks.

Yai et al. (1997) suggested a MNP model with covariance matrix for route choice in the Tokyo rail network. This method considerably limits the number of covariance parameters to be estimated. An efficient method for MNP model is suggested by Bolduc (1999). He estimates a model with 9 alternatives. Needless to say that, choice set sizes in reality especially in route choice are often much larger.

2.1.3 Multivariate Extreme Value (MEV)

This model, also called Generalized Extreme Value (GEV), is proposed by McFadden (1978) and includes the MNL and Nested logit models. In comparison to the MNL model, the MEV model allows for some correlation.

Vovsha and Bekhor (1998) suggested Link-Nested Logit (LNL) model, which is a Cross-Nested Logit (CNL) formulation. Each path of the network corresponds to an alternative and one nest is defined related to each link. Consequently, it allows for a rich correlation structure. However, due to the huge number of nests, the nesting parameters cannot be estimated. Therefore, Vovsha and Bekhor (1998) propose using the network’s lengths of links and paths to approximate the nesting parameters.

Abbé et al. (2007) analyze the CNL model and define the exact correlation structure. The nesting parameters can be estimated by solving a system of equations involving numerical integration. There are two approximating method for the nest parameters which have been introduced by Prashker and Bekhor (1998) and Gliebe et al. (1999) based on the network topology.
2.1.4 Error Component (EC)

This model is a Normal Mixture of MNL (MMNL) and was introduced by Bolduc and Ben-Akiva (1991). It was designed to be a compromise between the MNL and MNP models. Utilities have Normal and Extreme Value distributions of error terms simultaneously; thus, a flexible correlation structure can be defined while it keeps the form of a MNL model. The estimation for EC is easier in comparison with MNP but simulated maximum likelihood estimation is required.

The EC model can be supplemented by a factor analytic specification where some structure is explicitly specified in the model in order to decrease its complexity (Ben-Akiva and Bolduc, 1996).

There are a number researches on using MMNL for real sized network. For instance, Paag et al. (2002) and Nielsen et al. (2002) present a MMNL model considering random coefficient and keeping the error component structure to estimate route choice models in Copenhagen.

2.2 Estimation

Maximum-likelihood approach

Train (2003) illustrated the estimation of the attribute parameter where \( N \) individuals are involved in the estimation process. Based on the logit probability function and the maximum-likelihood approach, the probability of an alternative’s being chosen by the decision maker \( n \) is

\[
\prod_i (P_{ni})^{y_{ni}}
\]

(2.10)

Where \( y_{ni} = 1 \) if individual \( n \) has chosen alternative \( i \) and zero otherwise. Hence, the probability of a situation that all the decision makers choose the observed choices is:

\[
L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}}
\]

(2.11)
\( \beta \) is a vector representing the parameters of the model. Consequently, the log-likelihood function will be:

\[
LL(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} \ln P_{ni}
\]  

(2.12)

According to the mentioned assumption of the representative utility function’s being linear, the value of \( \beta \) is estimated where it maximizes the function (3.12). This is fulfilled putting the derivation of the likelihood function with respect to each parameter equal to zero.

\[
\frac{dLL(\beta)}{d\beta} = 0.
\]

(2.13)

The values of \( \beta \) that satisfy this equation are the estimations of the parameter.

**Indirect inference approach**

Indirect inference is a simulation based method for estimating or making inference of parameters of economic models (Smith, 2008). It is most applicable in estimating models with too difficult to evaluate or analytically intractable likelihood functions. Like other simulation-based methods, a major prerequisite of the indirect inference approach is that it should be possible to simulate data from the economic model for different values of the parameters involved in the model.

The main characteristic of the indirect inference method is that it uses an approximate or auxiliary model in order to form a criterion function. The number of parameters of auxiliary models have to be more or at least equal to parameters of the real models. There are two requirements for choosing an auxiliary model. First, it should be easy to estimate since we want to get help from an auxiliary model to estimate the auxiliary parameters and run the auxiliary model repeatedly. Secondly, the auxiliary model has to be flexible enough to capture the variation of the observed data.

The aim of the indirect inference is to select parameters of economic model such that the simulated and observed data look the same from the auxiliary model’s point of view.
\[ \hat{\beta} = \arg\max_{\beta} \mathcal{L}(y; x, \tilde{\theta}(\beta)) \] (2.14)

Where \( \hat{\theta} \) is the estimation of the auxiliary model parameter for the observed data

\[ \hat{\theta} = \arg\max_{\theta} \mathcal{L}(y; x, \theta) \] (2.15)

and \( \tilde{\theta}(\beta) \) is the auxiliary model estimation of the simulated data.

\[ \tilde{\theta}_m(\beta_m) = \arg\max_{\theta} \mathcal{L}(\tilde{y}(\beta_m); x, \theta) \] (2.16)

2.3 Sources and Data Collection

In this section we review a number of route choice modeling applications. Telephone, mail and more recently web-based surveys are the traditional methods of trip data collection. Travelers would be asked to describe the chosen paths and the related information.

Different collection methods are suggested in different literatures like Mahmassani et al. (1993) and Abdel-Aty et al. (1995). Ben-Akiva et al. (1984) presents one of the first applications of route choice modeling by using the data collected in 1979 for a road in Netherlands. Ramming (2001) uses data collected by asking travelers to describe a selected path with a set of segments and he implements the shortest path concept. Another literature describing the conventional method of collecting data is Prato (2004). He used data collected in a web-based survey in which travelers were asked to identify their selected routes on a map of city center. Vrtic et al. (2006) uses collected trip data in Switzerland based on telephone interviews.

The advent of passive monitoring of route choice caused different authors to compare these two different means of data collection (conventional methods and the GPS data).

For instance Murakami and Wagner (1999) and Jan et al. (2000) compared data collected by conventional survey methods to GPS data.
In passive monitoring, data is collected automatically and in electronic format. These characteristics are the advantages of this new generation of data collection methods compared to traditional surveys (Wolf et al., 1999, and Zito et al., 1995, for detailed discussions). However, there are a few restrictive weaknesses for using GPS data. For instance Bierlaire et al. (2007) discussed that inaccuracy in data may be introduced depend on the number of available satellites and receiver’s noise.

Wolf et al. (1999) indicate that an accuracy level of 10 meters is necessary for map matching GPS points in urban areas with a high level of certainty. Wolf et al. (1999) tested data collected in Atlanta and found out that the best performance receivers obtained this accuracy level of 10 for 63% of the GPS points on average. Nielsen (2004) showed that 90% of the trips collected in the Copenhagen region had missing data.

Another considerable point in using GPS data is that the data is stored in one set of GPS points and data processing such as map matching and trip end identification is essential for identifying the trips. In addition, there is missing data which should be considered by the researcher. Marchal et al. (2005) suggested a map matching algorithm for large choice sets. They consider calculation time and report that accuracy evaluating is difficult when the real traversed paths were unknown.

Even though the GPS data has some flaws as already mentioned, it is frequently used for route choice analysis. For instance Nielsen (2004) used 100,000 observations in the GPS dataset in Copenhagen in order to realize route choice behavior and responses to road pricing scenarios. He emphasized on the problems related to missing data and technical problems in his study.

2.4 Choice Set Generation

In reality, in large networks there are a huge number of routes which connect two points of origin (O) and destination (D). In fact there may be an infinite number of choices between origin and destination which can be the chosen
2.4. CHOICE SET GENERATION

route of travelers. This set, referred to as the universal choice set, can not in
general be enumerated. That is why a subset of paths needs to be defined in
order to estimate a route choice model and path generation algorithms are used
to meet this requirement. There are two different approaches for generation of
paths: deterministic and stochastic.

Deterministic approaches refer to algorithms always generating the same
set of paths for a given origin-destination pair, while a unique or observation
specific subset is generated by stochastic approaches.

Another approach is defining choice sets in a probabilistic way. More
details on probabilistic choice set models are discussed by Manski (1977),
Swait and Ben-Akiva (1987), Ben-Akiva and Boccara (1995) and Morikawa
(1996). Cascetta et al. (2002) suggest that in order to simplify probabilistic
choice set models, the choice set can be considered as a fuzzy set in a model.

In the following, we give a brief overview of existing deterministic and
stochastic path generation algorithms. Further information about these path
generation algorithms can be found in Fiorenzo-Catalano (2007) and Bovy
(2007).

Deterministic Approaches

As a reviewer, I found that most of existing path generation algorithms are
deterministic approaches. A majority of them are based on repeated shortest
path search. This type of approach is computationally appealing due to the
efficiency of shortest path algorithms.

There are different methods to obtain the shortest path in order to generate
choice sets. One of them is the link elimination approach. To clarify, one or
some links belonging to the shortest path are eliminated and a new shortest
path in the modified network is calculated and introduced in the choice set.
(Azevedo et al., 1993)

Another method proposed by de la Barra et al. (1993) is to increase the
generalized cost on the links on the shortest path and then run the modified
network to get a shortest path for the new cost structure, instead of eliminating
links. In this method we can kill two birds with one stone. One is that the link penalty approach allows for essential links (e.g. bridges) to be used and a connected network is guaranteed. On the other hand, the same path can be generated repeatedly depending on how the cost structure is updated.

Furthermore, considering time as a factor to find the shortest path, Ramming (2001) infers that the computational time is preventively large and disregards it for further consideration in his work. Even though the mentioned methods compute the shortest paths, they may generate paths which are very similar to each other; thus, another method which is using a constrained K-shortest paths approach was introduced. It is another variant of repeated shortest path search (Van der Zipp and Fiorenzo-Catalano, 2005).

In addition, Ben-Akiva et al. (1984) suggest an approach of considering specific criteria such as fastest, shortest or most scenic paths, for generating choice set paths. Thus, shortest paths will be repeatedly calculated based on different generalized cost functions.


**Stochastic Approaches**

As already mentioned, the majority of route choice generations are based on the deterministic approach; however, most of the deterministic approaches can turn into stochastic if we use random generalized cost for the shortest path estimations. In other words, the shortest path is calculated according to the randomly distributed generalized cost and introduced to the choice set.

Ramming (2001) suggests a simulation method that generates alternative routes by drawing link costs from different probability distributions. Bovy and Fiorenzo- Catalano (2006) present a choice set generation method which is named *doubly stochastic* approach. The method is similar to the simulation ones but generalized cost functions are specified as well as utilities where the
parameters and attributes are stochastic. A filtering process is presented by them in order to select only choices satisfying some constraints are kept in the choice set.

**Shortest path**

The shortest path concept is one of the most fundamental concepts in the combinatorial optimization since it plays the major role in theoretical optimization problems. In other words, in order to solve most combinational optimization, the shortest path concept is used either directly or indirectly. In the route choice modeling there is a need for the generation of the feasible alternatives, the choice set and the simulated observations for each OD-pair. For the choice set generation part, we would like to add a number of highly probable paths to the choice set.

As we know shorter paths are more probable to be chosen. The same holds for the simulated observations. The first algorithm in the history of shortest path methods was suggested by Dijkstra and all other algorithms proposed afterwards are somehow implementations of Dijkstra’s original method. Firstly, the notation relevant to Dijkstra’s method should be defined. \( G = (N, A : l) \) is a network(graph) with sets of nodes \( N \) and links \( A \), and a length function \( l : A \); thus, \( l_{ij} \) is defined as the length of link \((i, j)\). For each node \( u \), a list containing all links going out from \( u \) is defined. This set is defined by:

\[
FS(u) = \{(u, j) \in A\}
\]

One essential requirement to find the shortest path from a given node \( r \) to a specific destination is finding a shortest path tree \( T(r) \) including all the shortest paths from the node \( r \) to every \( v \in N \). Forming the shortest path tree is needed since there is no algorithm to solve the problem independently. \( d_v \) is assumed to be the length of the shortest path from \( r \) to \( v, v \in N \) in \( T(r) \).

\( T \) is a shortest path tree with the origin \( r \) if and only if:

\[
d_i + l_{ij} - d_j \geq 0, (i, j) \in A.
\]
The final notation demonstrates the candidate nodes (Q). This set shows a list of nodes sharing a mutual link with a given node.

If \( l_{ij} \geq o, (i,j) \in A \), then each node is removed from Q exactly once.

This is because of the fact that at each step, assuming no negative lengths for the links, \( d_u \) will be the shortest distance from \( r \) to \( u \), if \( u \) is a minimum element of Q.

The simplest version of Dijkstra’s algorithm is SPT. The following codes are presented by Gallo et al. (1984) for SPT. They seem to have been written in Pascal (programming language).

<table>
<thead>
<tr>
<th>Algorithm1: Procedure Dijkstra-SPT (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. begin</td>
</tr>
<tr>
<td>2. initialize ( p, d, Q; )</td>
</tr>
<tr>
<td>3. repeat</td>
</tr>
<tr>
<td>4. comment: selection of the minimum label node ( u; )</td>
</tr>
<tr>
<td>5. scan Q to retrieve and delete ( u; )</td>
</tr>
<tr>
<td>6. comment: exploration of ( FS(u); )</td>
</tr>
<tr>
<td>7. foreach ( (u, v) \in FS(u) ) suchthat ( d_u + l_{uv} &lt; d_v ) do</td>
</tr>
<tr>
<td>8. begin</td>
</tr>
<tr>
<td>9. ( d_u := d_u + l_{uv}; )</td>
</tr>
<tr>
<td>10. ( p_v := u; )</td>
</tr>
<tr>
<td>11. if ( v \notin Q ) then insert ( v ) into ( Q; )</td>
</tr>
<tr>
<td>12. end</td>
</tr>
<tr>
<td>13. until ( Q = 0 )</td>
</tr>
<tr>
<td>14. end</td>
</tr>
</tbody>
</table>

\( p_v \) is the predecessor of node \( v \) in the tree.
2.5. EVALUATION OF GENERATED CHOICE SETS

The Dijkstra algorithm is an applicable and very helpful algorithm to find the shortest path, whereas it is time consuming when trying to find the shortest path between two specific nodes. Aiming the reduction of the computational time, the shortest path algorithm should be broken while it reaches the considered destination point and then a new run should be started to find the next route.

2.5 Evaluation of Generated Choice Sets

It is really difficult to evaluate the generated choice sets because the actual choice sets are unknown in general. Ramming (2001) proposed the following criteria to evaluate the generated choice set.

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

In addition, Prato and Bekhor (2006) use reduced choice set in estimating models based on subsets of the generated choice sets to examine the effect of number of paths in the choice set. Fiorenzo-Catalano (2007) emphasize on choice set generation to interpret what a reasonable route is. They provide definitions of reasonable routes. These definitions and interpretations are based on different criteria at route level and overlap, size and etc.
Chapter 3

Methodology

3.1 Introduction

As a natural characteristic, the points reported through the Global Positioning System do not match directly to a network on a digital map. This is caused by several error sources not relevant to this study. Therefore, we have to apply an efficient method to map the reported GPS points on the network. There are several literatures presenting different methods called map matching methods.

In order to simplify the project, we ignore the map matching part but as a replacement; simulated matched GPS points are used. Moreover, the lack of knowledge about the chosen paths among consecutive points (In this study we call paths between GPS points sub-path in order to distinguish them from main paths between the origin and destination) is the major problem for our study. We need to know the exact path to evaluate the efficiency of the model; therefore in our Monte Carlo simulation, the true parameter is assumed to be $\beta = 1$ (we call it $\beta_{true}$) and "observed" data set will simulated.

In this study we use a method to estimate a model with positive random link costs. Instead of doing hard computations to find the maximum likelihood estimate, our method is based on the principle of indirect inference. we assume that the true model can be easily simulated. In this approach, By choosing the parameters in the true model such that the simulated data set looks like
the real-world data set when we examine it through the lens of the auxiliary model, we will be able to consistently estimate the parameters of the true model. Through this process we will first specify the model that we want to estimate, and then our indirect inference based estimator will be introduced. We will define the properties of the estimator in a Monte Carlo simulation experiment.

3.2 The Model

In this part the model that is used for the proposed route choice problem is presented. We have the network $\mathcal{N}$ which is defined by sets of nodes (vertices) $v$ and links (edges) $l$. These two together indicate the direction of the link. Each link is defined as a connector of a source node (from-point) $v^o$ to a destination node (to-point) $v^d$. The path between a source node and a destination node could be seen as a sequence of links, where

$$s(l_1) = v^o, \quad d(l_j) = s(l_j + 1) \quad \text{for} \quad j = n - 1, \quad d(l_n) = v^d.$$  

Hence, a path may be defined by the index of links $\pi = \{l_1, \ldots, l_n\}$. Each link is associated with a vector of its own characteristics, represented as $x_l$, and a strictly positive cost function $c(x_l, \epsilon_l; \beta)$.

The cost function is defined as the cost associated with each link $l$ for each individual $i$. To clarify, the cost function includes different components. $\epsilon_l$ is an individual specific random link cost and $\beta$ is the vector of coefficients for the links, ought to be estimated. In this report, the cost function is assumed to have a linear deterministic component.

$$c(x_l, \epsilon_l) = \beta x_l + \epsilon_l$$  \hspace{1cm} (3.1)

It should be noted that, since the deterministic part and the random one are additively separable, we have the cost function as two separate parts showed in the formula above. So far, the cost function of each link can be computed by the procedure mentioned above. As we know, each path consists of a number of links; thus, another assumption is that the cost function of
each path $\pi$ is additive in link costs. In other words, the cost of a path can be attained from the summation of all the link costs through the path. Hence, the cost for individual $i$ to pass a path $\pi$ is computed by

$$C_i(\pi) = \sum_{l \in \pi} c(x_l, \epsilon_{li}) \tag{3.2}$$

Furthermore, we assume the travelers know both the link characteristics and their idiosyncratic random utility $\epsilon_{li}$ regarding their passed links. Since the choice makers tend to maximize their utility, they will rather choose the path with the lowest generalized cost in this model.

$$\pi_i = \arg\min_{\pi \in \Omega(v_{oi}^i, v_{di}^i)} C_i(\pi) \tag{3.3}$$

Assuming $v_{oi}^i$ as the origin and $v_{di}^i$ as the destination, $\Omega(v_{oi}^i, v_{di}^i)$ represents all the possible paths between the traveler’s origin and destination forming the choice set.

The random part $\epsilon_{li}$ can have an arbitrary distribution like normal, gamma or exponential distributions. Through implementing the model, in this report we assumed that the random cost component followed a truncated normal distribution. In order to avoid negative link costs, it is common to introduce a constraint on the values that the cost function can return. Here we assumed that $\epsilon_{li}$ follows a standard normal distribution with only the positive values.

Based on the above assumption, the prerequisites of Dijkstra algorithm (always necessitates a positive link value on the network) is satisfied. The dependent variables are the observed routes $y = \{\pi_i\}_{i=1}^N$ and the observed characteristic are the links costs $\{x_l\}_{l=1}^L$ regarding the network.

According to the definition of the project, the observed routes are in low frequency GPS point format. They are supposed to be translated into a path including a sequence of links which is suitable as an input for our model.
The proposed model specifications:

• A link-based model with random link cost is implemented.
• The random cost component $\epsilon_{li}$ follows a standard normal distribution with only the positive values retrieved.
• The observations are in low frequency GPS point format.

The final aim of these discussions is to estimate $\beta$ in equation 3.1. The likelihood function for this model is complicated and hard to estimate; therefore, a simple method is requisite.

3.3 Indirect inference

In this section, the indirect inference method is introduced to estimate the parameter of proposed route choice model $\beta$. In order to meet the simplicity requirements regarding choosing an approximate model, the simple multinomial logit (MNL) is used as an auxiliary model with the same number of parameters to the true models. In a situation that each individual $i$ is choosing from a set of alternative resources $r$, each $r$ is given a utility of $U_{ri}$ as follows:

$$U_{ri} = \theta X_r + \epsilon_{ri}$$  \hspace{1cm} (3.4)

where $\theta$ is the vector of the auxiliary parameters that the parameters in the basic model are inferred from, $X_r$ represents route characteristics and $\epsilon_{ri}$ is assumed i.i.d Gumbel distributed. The auxiliary model does not need to be an accurate description of the data generation process. This model operates as a window through which we can view both observed data and simulated data generated by the economic model. In brief, it selects aspects of the data on which we want to focus in the analysis.

The proposed auxiliary model:

$$U_{ri} = \theta X_r + \epsilon_{ri}$$

We have a relationship $\theta(\beta)$, which is defined as a binding function introduc-
3.3. INDIRECT INFERENCE

ing the smoothed relation between different $\beta$ values and their corresponding estimated $\theta$’s through the auxiliary model.

We use the proposed model in section 3.2 (the true model) to generate simulate simulated data sets.

![Diagram of the II based estimator](attachment:diagram.png)

**Figure 3.1.** The overview of the II based estimator

We need to point out that some arbitrary structural parameters $\beta_m, m = 1, \cdots, M$ are generated and $M$ different data sets are simulated with the same size $N$ as the size of the true data set based on the true model. All simulated choices $\tilde{y}(\beta_m)$ in the same data set is generated by the same parameter $\beta_m$. (Karlström et al., 2011)

It should be mentioned that the growth in the data set size $N$ causes the estimated $\tilde{\theta}_m(\beta_m)$ to converge to a non-stochastic limit $\theta(\beta_m)$. On the other hand, the estimated binding function $\tilde{\theta}_m(\beta_m)$ is discontinuous due to the discrete nature of the individual choices $\tilde{y}(\beta_m)$. A smooth binding function $\tilde{\theta}(\beta)$ is estimated by local regression or OLS, for $M$ different given values of $\beta_m$ and their corresponding $\tilde{\theta}_m(\beta_m)$. (Karlström et al., 2011)

In our case the true model is the route choice model. It is really useful but difficult to estimate. Yet, in order to simulate a path for a given origin and destination pair, we just need to draw random links’ cost and calculate the shortest path.

In brief, data sets with different $\beta$ are generated; then, an auxiliary model
is applied to compare the simulated and observed data. Now it is time to change parameters $\beta$ in the simulated data set, in a way that they look like the observed data, using the auxiliary model.

We use likelihood ratio approach in order to find the indirectly inferred estimate. Hence, according to the following equation, $\hat{\beta}$ is the indirectly inferred estimate of $\beta_{true}$. Targeting the maximization of the likelihood function of the auxiliary model, we choose parameters of the true model such that observed data meeting the constraint of estimated binding function (Figure 3.1).

$$\hat{\beta} = \arg \max_\beta L(y; x, \tilde{\theta}(\beta))$$  \hspace{1cm} (3.5)

**GPS data set generation**

As explained in the introduction part, the main factor that differentiates this study from the previous research on Borlänge is the fact that we considered low frequency GPS points as observed trip data. The real $\beta(\beta_{true})$ must be known for further evaluation of the model; hence, the following algorithm (Algorithm 2) is implemented to create a set of GPS points corresponding to each trip. The used procedure is defined through the following steps:

(i) $\beta_{true}$ is assumed to be equal to one. The random part of the link cost function should be redrawn and the shortest path must be calculated for the given origin and destination. The Dijkstra algorithm returns one path including links and nodes as a temporary observed path $y_{temp}^k(\beta_{true} = 1)$.

(ii) The average speed of cars in the city center is assumed to be 16 km/h. Based on this, the location of some virtual cars is determined every 60 seconds on the temporary observed paths.

(iii) In most locations, the calculated points are in the middle of the links. Therefore, in order to simplify the algorithm, the start nodes of the links holding the calculated points are stored as observed GPS points $gp_j(y_{temp}^k)$.

(iv) All GPS point sets must contain the initial origins and destinations; therefore the origin is taken as the first point and the last simulated GPS point is replaced by the destination. The process is illustrated in figure 3.2.
3.3. INDIRECT INFERENCE

Figure 3.2. Applied method for Converting unmapped simulated GPS points to nodes

Algorithm 2: GPS data set generation

Require: network \( \mathcal{N} \)
Require: Individual choice sets \( I_i(o,d) \)
1. Take \( \beta_{true} = 1 \)
2. for simulated true observation \( k = 1, \cdots, N^o \) do
3. Simulate a path \( y_{k}^{temp}(\beta_{true} = 1) \)
4. Create GPS dataset \( gp_j(y_{k}^{temp}) \)
5. end for
3.4 Universal choice set generation

Choice set is an inseparable part of discrete choice models. In order to generate it, a method is required to generate all feasible alternatives for a proposed OD-pair. In a real network, the number of possible routes among every two points is infinite; hence, the application of an efficient and fast method is necessary to generate a semi universal choice set including all observed routes. With regard our auxiliary model being the simple MNL, we need to generate a universal choice set corresponding to the proposed OD-pair in order to estimate the auxiliary parameter.

3.4.1 Monte Carlo (MC) view

Zantema et al. (2007) explained an application of Monte Carlo method in the choice set generation, repeating the shortest path search algorithm on the network after successive randomizations of the link attributes. The path search may be based on variant attributes in the cost function. The termination criterion in Monte Carlo search is reaching the maximum number of iterations. This approach can be modified by gradually increasing the variance of the randomization at each step of the iterations. This change can reduce calculation times, as longer routes are returned in fewer iterations. However, it can lead to the route set’s becoming larger than the desired size. For a description and example see Bliemer & Taale, (2006). Naturally, the number of generated unique routes depends on the number of iterations and the variance of the random part.

Labeling is another option for applying the Monte Carlo approach considering several search criteria, maybe for different user groups (such as vehicle types, trip lengths). Because of the variation in search criteria, the MC method is known to be a sufficient method to find paths that are spatially different corresponding to the variation of travelers’ preferences. Using the labeling approach may cause the generation of more routes than desired and excessive growth in calculation times.
3.5. DETAILED DESCRIPTION OF THE ESTIMATOR

In this paper, as already mentioned, one random variable is defined in the cost function regarding the unobserved link factors and it has been randomized using a positive random distribution. This implies that the random terms are always added to the link length (which stands for the observed attribute) making generated routes longer (or equally long) than the objectively shortest one. Experiments have demonstrated (see Bliemer et al., 2007) that this randomization methods provide better results compared to a symmetric (positive and negative) randomization.

In practice, we generate our universal choice set by using the true model with different randomly $\beta$ to find 800 unique paths.

3.4.2 Sampling of Alternatives

The large size of the proposed universal choice set encouraged researchers to work on methods to draw subsets of the alternatives as choice sets. McFadden (1978) gave this idea for estimating MNL. Frejinger et al. (2009) explained that, estimations based on a subset of universal choice sets could return biased results and developed a method to get a sample paths and calculate the correction. Note that this correction technique can only be used for MNL.

Regarding the proposed auxiliary model, estimations using choice set sampling are rather easy to gain, but misspecified. Still, we can apply this auxiliary model to the real data set and arrive at an estimate $\theta$ of the auxiliary parameters based on Karlsrtöm (2011) experience.

3.5 Detailed description of the estimator

In this section, the detailed assumptions and notations of the estimator are explained. As the model network, the Borlänge’s graph in which all the links are allowed to be double directed, has been used in Frejinger et al. (2009). The network has 3077 nodes and 7459 links. The main focus is just on one OD-pair and in the last part of the study, the model will be applied to five
ODs. The link length $L_l$ is the only attribute which is taken into account. Hence, the cost of link $l$ is given by

$$c(L_l, \epsilon_{li}) = \beta L_l + \epsilon_{li},$$

(3.6)

the utility of a path $r$ is given by

$$U_r = \sum_{l \in r} \theta L_l + \epsilon_{ri},$$

(3.7)

Since a simple MNL model is proposed as our auxiliary model and the input data sets for this model must be in path format (Our observed data is in the GPS point format), we need to guess the traversed paths based on the observed GPS data. For this purpose, we use the true model through Algorithm 3 in order to translate the GPS data to paths. When trying to obtain the shortest path among consecutive couple GPS points, the hypothetical length parameter used to create sub-paths is taken as $\beta = \infty$. Then the sub-paths are made and sum up to create the final observed path. The algorithm is described in the following steps:

(i) Take the true model parameter $\beta$ (We denote the true model parameter for creating sub-paths by $\beta_0$ and it is assumed to be infinite).

(ii) Use the true model to draw sub-paths between the consecutive GPS points $(gp_j, gp_{j+1})$.

(iii) Add sub-paths together and create the estimated path $y^{gp_j}(\beta_0)$ corresponding to the given GPS dataset $gp_j$.

Algorithm 3: Estimate traversed path based on GPS data

<table>
<thead>
<tr>
<th>Require:</th>
<th>network $\mathcal{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require:</td>
<td>GPS dataset $gp_j$</td>
</tr>
</tbody>
</table>

1. Take $\beta_0 = \infty$
2. for $j=1, \cdots, J-1$ do
3. \hspace{1em} Simulate a sub-path $y_j^{(gp_j, gp_{j+1})}(\beta_0)$
4. end for
5. Add sub-paths together and create $y^{gp_j}(\beta_0)$
Through our indirect inference estimator, $M$ points $\beta_m$ are drawn from the given domain $D \ [0, 5]$ and all route choices $N = N_o$ for them are simulated. The corresponding auxiliary parameter $\tilde{\theta}_m(\beta_m)$ is estimated for each point. The binding function is estimated by the data set $\{\tilde{\theta}_m, \beta_m\}_{m=1}^M$ using OLS regression. This function presents the relation between $\tilde{\theta}_m$ and $\beta_m$. Figure 3.3 shows that the binding function has the desired property of being monotone. The indirect inference method estimates the true parameter calculating the binding function and considering equation 3.5. The heart of the estimator algorithm is explained in the following steps:

(i) Drawn $M$ points $\beta_m$ from the given domain $D \ [0, 5]$

(ii) All route choices $N = N_o$ for each $\beta_m$ are simulated and $M$ data sets are generated. (Algorithms 2, 3)

(iii) Employ the auxiliary model to compute the corresponding auxiliary parameters $\tilde{\theta}_m(\beta_m)$. 

Figure 3.3. The drawn $\beta$ and corresponding estimates of $\tilde{\theta}$ with related estimated binding function
(iv) Estimate the binding function by the data \( \{ \tilde{\theta}_m, \beta_m \}_{m=1}^M \) using OLS regression.

(v) Estimate the traversed path for the observed GPS data \( y^{g_{obs}}(\beta) \) (Algorithm 3)

(vi) Plug the estimated binding function into equation 3.5 and find the indirect inference estimate of the true parameter \( \hat{\beta} \).

The final algorithm is defined below:

**Algorithm 4: Indirect inference route choice estimator**

**Require:** network \( \mathcal{N} \)

**Require:** Individual choice sets \( I_i(o,d) \)

**Require:** Observed GPS dataset \( g_{obs} \)

1. for \( m=1,\cdots,M \) do
2. Draw \( \beta_m \in \mathcal{D} \)
3. for simulated observation \( i = 1, \cdots, N \) do
4. Simulate GPS dataset \( g_{pj} \)
5. Estimate the traversed path for the simulated data set \( \tilde{y}^{g_{pj}}(\beta_m) \) (Algorithm 3)
6. end for
7. Estimate auxiliary parameters \( \tilde{\theta}_m(\beta_m : \beta_0) = \arg \max_{\theta} \mathcal{L}(\tilde{y}(\beta_m); x, \theta) \)
8. Given \( \{ \tilde{\theta}_m, \beta_m \}_{m=1}^M \), estimate \( \tilde{\theta}(\beta, \beta_0) \) using linear OLS
9. Estimate the traversed path for the observed data \( y^{g_{obs}}(\beta) \) (Algorithm 3)
10. Estimate \( \hat{\beta} = \arg \max_{\beta} \mathcal{L}(y; x, \tilde{\theta}(\beta, \beta_0)) \)
Chapter 4

Results

We use a network of Borlänge city as described by Frejinger et al. (2009). The network contains 3077 nodes and 7459 links. The length of links is denoted by $L_l$. Monte Carlo simulation approach is used by getting the true parameter $\beta_{true} = 1$ to simulate the observed data set. Then, the proposed indirect inference estimator is used to estimate the true model parameter. The estimated parameter will compare with the true value of $\beta$ ($\beta_{true} = 1$) in order to examine validation of the proposed estimator. Briefly, in this section consistency of the proposed estimator is evaluated to retrieve the “unknown” value of $\beta$.

4.1 Estimation and Sensitivity analysis

In this section, we will estimate and compare several different specifications of our proposed true method. First, we specify the initial form of the estimator and the implemented auxiliary model. Then we discuss different statistical measures that can be compared across the model, such as the mean and standard deviation of estimated values in different runs. Finally, we present a discussion to find out the optimum estimator method specifications. The value of link length attribute is estimated based on a binding function which is estimated using $M = 10$ sample points. The observed paths contains $N = 3000$ routes which are made based on $\beta_{true} = 1, \beta_0 = \infty$. The estimator engine
applies a pseudo-universal choice set of size 800, and the size of the individual specific choice sets is 201 ($CS = 201$).

Furthermore sensitivity of the results to the different effective variables of the estimator is investigated, such as the choice set size $CS$, number of observations $N$, number of sample points $M$ and frequency of GPS data.

### 4.1.1 Number of observations

In this part we are interested to investigate on required number of observations in our simulation based estimating method. Turning back to the method definition, we assumed that we can simulate how many observations we need. However, in practice it is costly to gather huge number of observations and on the other hand this parameter has a great effect on the computation time of the estimator. Therefore, the main aim will be to keep this parameter small.

Obviously, we want to compare how well the different number of observations perform in term of accuracy for the estimating model. However, comparing the closeness of the mean value, for the estimated $\beta$'s through the ten runs, to the true $\beta_{true} = 1$ does no reveal to how many observations are required. To get a feeling for it, we will consider variation of the estimated values. Naturally, growth in the number of observations should induce a reduction in standard deviation of the estimated values through different runs. Table 4.1. presents the relevant statistical indicators of the estimated $\beta$ for different numbers of observations $N = \{1000, 3000, 5000\}$. The mean values are close to $\beta_{true} = 1$ and there is a significant reduction in the standard deviation as $N$ increases. The RMSE decreases corresponding to increasing the number of observations.

All results from these simulations are close to the true parameter. One hypothesis is that it may be caused by considering a single OD-pair. While the correlation structure is represented by the network, and for one OD-pair the induced correlation structure across routes in the choice sets is identical across observations in the sense that they all have the same universal choice set.
4.1. ESTIMATION AND SENSITIVITY ANALYSIS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>1000</td>
</tr>
<tr>
<td>Mean</td>
<td>1.061</td>
</tr>
<tr>
<td>Std</td>
<td>0.062</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.085</td>
</tr>
<tr>
<td>ZETA</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 4.1. Results from 10 simulations, choice set size is 201 and number of sample points \( M = 10 \). \( \beta_{\text{true}} = 1 \)

In order to test our estimation process will be performed for five different OD pairs in section 4.3.

The results are compared for different number of observations \( N = \{1000, 3000, 5000\} \) with the same choice set size \( CS = 201 \) and number of sample points \( M = 10 \). As displayed in figure 4.1, each bar represents the mean and standard deviation of the estimated values for the mentioned amounts of number of observations.

Figure 4.1. Sensitivity analysis regarding number of observations

The result for the estimated \( \beta \) confirms that the sample mean and standard deviation of the estimated values significantly improve as the number of observations increase. More observations leads to more satisfactory results, however computation time grows.
4.1.2 Choice set size

According to the model definition, a pseudo-universal choice set with the size of 800 is used in the algorithm and individual specific choice set size ($CS$) is initially 201. In this part, trends of measured factors are investigated with regard to change in $CS$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>51</th>
<th>101</th>
<th>201</th>
<th>401</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice set size</td>
<td>β</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.012</td>
<td>1.005</td>
<td>1.006</td>
<td>1.011</td>
</tr>
<tr>
<td>Std</td>
<td>0.041</td>
<td>0.042</td>
<td>0.041</td>
<td>0.038</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.041</td>
<td>0.040</td>
<td>0.039</td>
<td>0.038</td>
</tr>
<tr>
<td>ZETA</td>
<td>0.292</td>
<td>0.126</td>
<td>0.150</td>
<td>0.274</td>
</tr>
</tbody>
</table>

Table 4.2. Results from 10 simulations, number of observation $N = 3000$ and number of sample points $M = 10$. $\beta_{true} = 1$

Table 4.2. shows estimated $\beta$ for $CS = \{51, 101, 201, 401\}$ where the number of sample points is fixed on 10 and 3000 observations are considered. Estimated parameter mean has a $U$ shape behavior which minimum is located on $CS = 101$, however all are close to one. The standard deviation varies frequently and the minimum value is for choice set size 401. The RMSE presents a straight trend to become smaller for bigger choice set size.

Note that this variable has a small effect on computation time as well as results accuracy and will get lower priority than other effective factors of the model.

According to the explained assumptions, the choice set size is representer of the number of chosen routes from the universal choice set as alternative set for an individual.

In figure 4.2 statistical parameters related to estimated parameter $\hat{\beta}$ are illustrated for choice set size $CS = \{50, 100, 200, 400\}$ with the number of observations $N = 1000$ and number of sample points $M = 10$. Each bar represents the mean and standard deviation of the estimated values for denoted choice set sizes.
4.1. ESTIMATION AND SENSITIVITY ANALYSIS

For the estimated $\beta$, the mean and standard deviation do not vary much as the size changes. The mean is approximately close to true value for all sizes and choice set contains 200 members returned the best average, whereas variation of samples are more or less same to each other. Hence, the model has low sensitivity to choice set size.

4.1.3 Number of sample points for the OLS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Sample Points</td>
<td>Mean</td>
<td>Std</td>
<td>RMSE</td>
<td>ZETA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.069</td>
<td>0.060</td>
<td>0.089</td>
<td>1.166</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.042</td>
<td>0.054</td>
<td>0.066</td>
<td>0.770</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.031</td>
<td>0.047</td>
<td>0.054</td>
<td>0.674</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.021</td>
<td>0.044</td>
<td>0.047</td>
<td>0.484</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.014</td>
<td>0.042</td>
<td>0.043</td>
<td>0.333</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.012</td>
<td>0.045</td>
<td>0.044</td>
<td>0.278</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.011</td>
<td>0.044</td>
<td>0.043</td>
<td>0.255</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.006</td>
<td>0.041</td>
<td>0.039</td>
<td>0.150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3. Results from 10 simulations, number of observation $N = 3000$ and choice set size $CS = 201$. $\beta_{true} = 1$

Here we will compare estimation results where number of sample points varies from 3 to 10. The choice set size is fixed on 201 and in this part, model
behavior is tested regarding change in number of sample points. It should be noted that these points are implemented in binding function estimation.

The results are improved by increasing the number of sample points; whereas, there is very small bias regarding ten sample points. Such an improvement is intelligible from deceasing trend of represented ZETA values in table 4.3.

As an input for the algorithm, we should determine the number of sample points using to estimate binding function. Comparison regarding the number of sample points \( M = \{3, 4, 5, 6, 7, 8, 9, 10\} \) is presented in figure 4.4 where number of observations is \( N = 3000 \) and with choice set size \( CS = 201 \). Each bar shows the mean and standard deviation of the estimated values for given number of sample points.

![Figure 4.3. Sensitivity analysis regarding number of sample points](image)

The results of sensitivity analysis regarding the different number of sample points is presented in figure 4.3. As it is shown, the mean value is approximating to the true value when the sample size increases. Hence, we can conclude that our results are improving by growth in sample size from the mean and standard deviation views. It is notable that all sizes except 3 cover the true value with confidence level 95%.
4.1. ESTIMATION AND SENSITIVITY ANALYSIS

4.1.4 Sensitivity with regard to $\beta_0$

In all previous sections, $\beta_0$ was assumed $\infty$ (In model codes $\beta_0 = 10000 \ast \beta$ as a very huge coefficient). In other words, objectively shortest path chosen for drawing sub-paths among consecutive GPS points. It means $\beta_0$ is independent from main $\beta$ on which GPS points are generated based. In this part of the study, effects of variation in $\beta_0$ is noteworthy purpose. Table 4.4 illustrates trends of estimated $\beta$ statistical parameters regarding changes in $\beta_0$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ coefficient</td>
<td>0.5</td>
</tr>
<tr>
<td>Mean</td>
<td>0.993</td>
</tr>
<tr>
<td>Std</td>
<td>0.038</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.037</td>
</tr>
<tr>
<td>ZETA</td>
<td>-0.193</td>
</tr>
</tbody>
</table>

*In model codes $\beta_0 = 10000 \ast \beta$ with a very huge coefficient

The best results are achieved when $\beta_0 = 1.5 \ast \beta$ from standard deviation view and in $\beta_0 = 3 \ast \beta$ from closeness of mean to true value ($\beta_{true} = 1$). Variation in standard deviation is very low. The point in this table is that the last column which presents values regarding to $\beta_0 = \infty$ are more or less close to others’, while it has a special difference with previous columns. $\beta_0$ is quantified independent from $\beta$ and equal to 10000.

In figure 4.4 statistical parameters related to estimated parameter $\hat{\beta}$ are illustrated for different values of $\beta_0$ with the number of observations $N = 3000$ and number of sample points $M = 10$. $\beta_0$ is defined as a product of a coefficient and the true model attribute $\beta$. Each bar represents the mean and standard deviation of the estimated values for denoted value for coefficient of $\beta_0$.

The trends of variables are not traceable, however all represented values for $\beta_0$ cover the true value with confidence level 95%. Generally, the estimator is not significantly sensitive to the value of $\beta_0$. 
4.1.5 Frequency of GPS data

In this part we are interested to investigate on effect of variation in the frequency of GPS data sampling on our estimates regarding unknown attribute. Based on our data simulation algorithm, we can simulate GPS data sets with different frequency.

We aim to compare how well the different frequency of GPS data sampling perform in term of accuracy for the estimating model. Naturally, growth in the frequency of GPS points should leads improvement in estimation results due to have more sample points introducing a trip and consequently better estimates for the unknown paths. Moreover, accuracy of the results declines for lower frequency.

Table 4.5 shows the relevant statistical indicators of the estimated $\beta$ for different values of data sampling frequency. More appropriate mean values and lower standard deviation for higher frequency are significant out come of this table. The RMSE decreases corresponding to increasing the frequency of GPS data.

In figure 4.5 statistical parameters related to estimated parameter $\hat{\beta}$ are illustrated for different frequency of GPS data sampling with the number of
4.1. ESTIMATION AND SENSITIVITY ANALYSIS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Mean</td>
<td>1.002</td>
</tr>
<tr>
<td>Std</td>
<td>0.039</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.037</td>
</tr>
<tr>
<td>ZETA</td>
<td>0.046</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequency</th>
<th>1</th>
<th>0.8</th>
<th>0.67</th>
<th>0.57</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.019</td>
<td>1.013</td>
<td>1.013</td>
<td>1.032</td>
<td>1.032</td>
</tr>
<tr>
<td>Std</td>
<td>0.045</td>
<td>0.042</td>
<td>0.039</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.047</td>
<td>0.042</td>
<td>0.049</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td>ZETA</td>
<td>0.428</td>
<td>0.305</td>
<td>0.835</td>
<td>0.734</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.032</td>
</tr>
<tr>
<td>Std</td>
<td>0.049</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.053</td>
</tr>
<tr>
<td>ZETA</td>
<td>0.734</td>
</tr>
</tbody>
</table>

Table 4.5. Results from 10 simulations, number of observation $N = 3000$, choice set size is 201 and number of sample points $M = 10$. $\beta_{true} = 1$

observations $N = 3000$ and number of sample points $M = 10$. Each bar shows the mean and standard deviation of the estimated values for denoted value for GPS data frequency.

![Figure 4.5](image)

Figure 4.5. Sensitivity analysis regarding GPS points frequency

The standard deviation for estimated $\beta$ do not vary much as GPS points frequency changes. Rather, the mean goes far away from the true value for lower values of frequency.
4.2 On computational efficiency

Running time and memory usage of an algorithm are two most critical and measurable parameters. Effort to revise algorithm codes and structure in order to become efficient from run-time point of view is of high interest in computer science. In this part, we try to decrease computation time of our proposed model as much as not missing acceptable accuracy.

4.2.1 Binding function

Figure 4.6 illustrates trend of 1500 randomly drawn $\beta_m$ in $D = [0, 70]$ and their corresponding $\theta$. Clearly it is comprehensive the estimated linear function has a huge error and is not acceptable, due to monotonic trend is first requirement of having a satisfactory estimation results.

![Figure 4.6](image)

**Figure 4.6.** The drawn 1500 $\beta \in [0, 70]$ and corresponding estimates of $\hat{\theta}$ with related estimated binding function

To present sufficient binding function as a linear function, a few sample
points response to $\beta_m$ and their corresponding $\theta$ are required. For this purpose, 10 sample points ($M = 10$) are taken from defined interval. These sample points are randomly selected from the initial domain. As was explained in chapter 3, in the simulations part, this domain is defined to be $[0, 5]$ (Figure 3.3).

### 4.2.2 Estimation with more stages

According to initial assumptions, the defined domain for $\beta$ is $[0, 5]$. As a heuristic solution, in order to decrease the computation time, the estimator is used in different stages to be defined below. The heart of the algorithm at each stage $s$ is described in chapter 3. We use a local linear approximation for our indirect inference estimator. Since our initial guess of the domain for $\beta_m$ is chosen arbitrarily, it may be the case that it does not cover the estimate of the true parameters. In addition, the nonlinearity of the true smooth binding function is not captured by the local regression, and our local estimate of the binding function may provide a poor estimate. (Karlström et al., 2011)

![Figure 4.7. Improvement in results through stages](image)

As figure 4.7 illustrates, we repeat the process in stages $s = 1, \cdots, S$ and in each stage, we arrive at an estimate $\tilde{\beta}_s$. At the next stage, we re-center
the domain $D_{s+1}$ on the current estimate $\hat{\beta}^s$. At the same time we shrink the domain to concentrate the samples on the current estimate. As a result, the interval of the domain shrinks by each stage. Thus, the first stages provide crude estimates. That is, to improve accuracy at each stage we re-center and shrink the interval. (Karlström et al., 2011)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>No. sample points</th>
<th>$\beta$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Stage</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Mean</td>
<td>1.127</td>
<td>1.053</td>
<td>1.056</td>
</tr>
<tr>
<td>Std</td>
<td>0.081</td>
<td>0.074</td>
<td>0.075</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.149</td>
<td>0.089</td>
<td>0.091</td>
</tr>
<tr>
<td>ZETA</td>
<td>1.561</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>Computation time (min)</td>
<td>4:05</td>
<td>7:49</td>
<td>11:16</td>
</tr>
</tbody>
</table>

Table 4.6. Results for two model definitions:
1. 10 simulations, number of observation $N = 1000$, choice set size is 201 and three sample points ($M = 3$) with three stages ($S = 3$), $\beta_{true} = 1$
2. 10 simulations, number of observation $N = 3000$, choice set size is 201 and ten sample points ($M = 10$) with one stage ($S = 1$), $\beta_{true} = 1$

* The cumulative computation time through the stages.

We examine our case study with the stage based estimator being mentioned above. Table 4.6 shows the results regarding two different situations: the first regarding 10 simulations with 1000 observations, 201 choices and 3 sample points through 3 separate stages where $\beta_{true} = 1$, the second, regarding 10 simulations with 3000 observations, 201 choices and 10 sample points but just for one stage where $\beta_{true} = 1$. Remarkable parameter in the table is the computation time showing significant improvement in implementation of the stage-based estimator. The results of the stage-based estimator are appropriate as much as those are gained from previous version of the estimator (without using stages), whereas the computation time for one simulation reduces from 34:32 minutes to 11:14. It is notable that there is no impressive improvement through stage two to three and it just increases the computation time. Thus, we will run the estimator for more stages till meet desirable
results. The number of stages may be vary for different cases depending on their binding function trends.

### 4.3 Estimation for five OD-pairs

Specified one couple origin and destination points in all discussed sections may decrease reliability of the results. Therefore, the model is run for the five different OD-pairs collected from various parts of the network with different distances. Table 4.5 contains corresponding results statistical parameters for one and five ODs.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>1 OD</th>
<th>5 ODs</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. sample points</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Stage</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mean</td>
<td>1.127</td>
<td>1.053</td>
</tr>
<tr>
<td>Std</td>
<td>0.081</td>
<td>0.074</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.149</td>
<td>0.089</td>
</tr>
<tr>
<td>ZETA</td>
<td>1.561</td>
<td>0.72</td>
</tr>
</tbody>
</table>

**Table 4.7.** Results for two model definitions:
1. 10 simulations, number of observation $N = 1000$, choice set size is 201 and three sample points ($M = 3$) with three stages ($S = 3$), $\beta_{true} = 1$ for one OD-pair
2. 10 simulations, number of observation $N = 1000$, choice set size is 201 and three sample points ($M = 3$) with three stages ($S = 3$), $\beta_{true} = 1$ for five OD-pairs

In a same situation (equal number of observations, number of sample points, choice set size and number of runs), five ODs return closer estimation mean to the true value with lower standard deviation and more desirable RMSE and ZETA values after running the estimator for two stages. Therefore, the estimator works well for multiple OD-pairs as for single one neither loss of accuracy nor precision.
Chapter 5

Conclusions

In this study we have estimated a route choice model using low frequency GPS data sampling. The route choice is a concept analyzing travelers’ behavior regarding their preference of choosing routes. When the travel data is collected with low frequency, it is unknown which path has been traversed between the GPS data points. Moreover, GPS data has measurements error. These characteristics may introduce bias into the estimates.

We have designed an algorithm to consistently estimate a given route choice model in the presence of sparse GPS data and measurement errors. The indirect inference method is applied as a structured procedure to estimate a model with random link costs when a likelihood function is analytically difficult to calculate. On the other hand, our proposed method corrects for bias being caused by GPS data measurement errors and lack of knowledge regarding traversed paths between GPS data points.

The results show that, applying the indirect inference approach to route choice estimation is a worthwhile solution, when the travel data is in a sparse GPS data format.

Considering GPS data errors is suggested as an extension to the proposed indirect inference method for future researches.

The main conclusion is that indirect inference is a useful option in the tool box for route choice estimation which can be used for estimating observed
path using low frequency GPS sampling data.
Bibliography


the 11th International Conference on Travel Behaviour Research, Kyoto, Japan.


