Two papers on consistent estimation of a route choice model and link speed using sparse GPS data

MASOUD FADAEI OSHYANI
Licentiate Thesis in Transport Science
Stockholm, Sweden 2013

KTH Royal Institute of Technology
School of Architecture and the Built Environment
Department of Transport Science
Division of Transport and Location Analysis
SE-100 44 Stockholm
SWEDEN
Two papers on consistent estimation of a route choice model and link speed using sparse GPS data

TRITA-TSC-LIC 13-003

KTH Royal Institute of Technology
School of Architecture and the Built Environment
Department of Transport Science
Division of Transport and Location Analysis
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie licentiatexamen i transportvetenskap fredagen den 7 juni 2013 klockan 13.00 i sal Nash/Wardrop, Kungl Tekniska högskolan, Teknikringen 10, Stockholm.

© MASOUD FADAEI OSHYANI, June 2013

Tryck: Universitetsservice US AB
Abstract

Global Positioning System and nomad devices are increasingly used to provide data from individuals in urban traffic networks. In the two papers of this thesis we focus on consistent estimators of a route choice model and link speed.

In many different applications, it is important to predict the continuation of an observed path, and also, given sparse data, determine where the individual (or vehicle) has been. Estimating the perceived cost functions is a difficult statistical estimation problem, for different reasons. First, the choice set is typically very large. Second, it may be important to take into account the correlation between the (generalized) costs of different routes, and thus allow for realistic substitution patterns. Third, due to technical or privacy considerations, the data may be temporally and spatially sparse, with only partially observed paths. Finally, the position of vehicles may have measurement errors. We address all these problems using an indirect inference (II) approach. We demonstrate the feasibility of the proposed estimator in a model with random link costs, allowing for a natural correlation structure across paths, where the full choice set is considered.

In the second paper, we develop an estimator for the mean speed and travel time based on indirect inference when the data are spatially and temporally sparse. With sparse data, the full path of vehicles are not observed, which is typically addressed using map matching techniques. First, we show how speed can be estimated using an auxiliary model which includes map matching and a model of route choice. Next, we further develop the estimator and show how both speed and the route choice model can be jointly estimated by using iteration between an II estimator of speed and the II estimator of the route choice model (developed in Paper I). Monte Carlo evidence is provided which demon-
strates that the estimator is able to accurately estimate both speed and parameters of the route choice model.
Sammanfattning

GPS (Global Positioning System) och rörliga nomad-apparater blir alltmer vanligt förekommande för att samlta in data från individer i stadstrafik. Avhandlingens fokus är konsistenta estimatorer för hastighet på länkar och ruttvalsmodeller.


skatta både hastighet och ruttningsmodellens parametrar.
Acknowledgements
First I would like to sincerely appreciate my supervisors Anders Karlström and Marcus Sundberg. Their unsparing efforts to help me find the way to solve my research problems are unutterable.
I’m very grateful for all the valuable discussions with Per Olsson. 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 create a very friendly atmosphere. 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 science department was.
Finally I would like to thank my wife. She was always there for me with her support.
# Contents

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Route choice</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Travel time</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Data collection techniques</td>
<td>3</td>
</tr>
<tr>
<td>1.3.1 Route choice data collection</td>
<td>3</td>
</tr>
<tr>
<td>1.3.2 Travel time data collection</td>
<td>5</td>
</tr>
<tr>
<td>2 Experimental conditions / theoretical models</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Borlänge Network</td>
<td>10</td>
</tr>
<tr>
<td>2.2 The Model</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Indirect inference</td>
<td>12</td>
</tr>
<tr>
<td>3 Results and Conclusions</td>
<td>15</td>
</tr>
<tr>
<td>3.1 Paper I</td>
<td>15</td>
</tr>
<tr>
<td>3.2 Paper II</td>
<td>16</td>
</tr>
<tr>
<td>Bibliography</td>
<td>17</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Nowadays, transportation has an important role in people’s day life. There are different purposes for trips such as going to work, picking up children, shopping and so on. In the travel behavior analysis, several aspects of trips are considered such as the chosen path, the purpose, the destination, the time and mode of transport. The route choice is a concept analyzing travelers’ behavior concerning their preference of selecting routes.

Travelers’ preference in route choice could depend on different characteristics such as distance, road type, travel time, cost or number of traffic lights. In addition, individuals’ characteristics, such as gender, age, income influence the result of route choice models.

1.1 Route choice

In route choice analysis the main concern is to identify a route which would be selected by travelers in a transportation network. The better understanding of route choice could be helpful for predicting behavior under different circumstances. For instance, in a congestion charging system, routes which are taken by travelers crossing the charging gates should accept a more costly trip compared to a previous situation of not having charges. Route choice
models could be applied to analyze the congestion charging scenario in order to predict forthcoming travelers’ behavior. Furthermore, the route choice models are used in the Vehicle Routing Problem (VRP), where a number of vehicles should find their routes to distribute their products among several destinations.

Route choice models are generally based on a link-by-link investigation of observed paths which can be obtained by collecting data by asking travelers about the reason of their route choice or by passive monitoring which is the implementation of the Global Positioning System (GPS).

1.2 Travel time

Travel time is one of the most critical aspects of all trips which is usually considered by people in their route choice. Travel time is defined as a required time to traverse a route between two specific geographical points. The term of travel time is well known and understandable for a wide range of audiences, such as transportation researchers, politicians, media and most people. It provides a key aspect in transportation planning and appraisal, and to be able to accurately measure travel time is of paramount importance. For instance, travel time is related to other key factors such as congestion and pollution, and also has a significant impact in social cost benefit analysis, both directly and indirectly.

Turner et al. [12] mention three factors of interest in the travel time field studied in the 1990s. The first, travel time is a performance factor which represents traffic congestion in the congestion management systems. The second, travel time is a common factor which is defined for all transportation modes. Therefore it could be a meaningful parameter to compare different transport modes and distribute a common funding source among them. The third, travel time could increase involvement in transportation decision by parties which
are not expert in the field. As mentioned, the term is simple and easy to understand, yet precise enough for transportation analyses.

1.3 Data collection techniques

There are several data collection techniques in transportation studies. Based on the purpose of the study, the possible techniques are different.

1.3.1 Route choice data collection

In route choice research telephone, mail and more recently web-based surveys are the traditional data collection methods. Through these methods travelers explain their taken paths. Ramming [10] applies data collected by asking travelers to explain a chosen path with a set of road links and he uses the shortest path concept. Prato [8] also uses a conventional method of data collection. He applies web-based survey results in which people were asked to describe their chosen routes on a map of a city center. Vrtic et al. [13] uses trip data in Switzerland which was collected by telephone interviews.

The advent of passive monitoring of route choice caused different authors to compare these two different means of data collection (conventional and the new one). There is a lot of literature comparing data collected by conventional survey methods to GPS data. (See [6]; [3]). There is a common point in almost all literature reviews that passive monitoring methods have quite a lot of benefits compared to the conventional surveys. For instance, collected data is directly accessible in electronic layout. Additionally, trip data could be collected repeatedly for several days of trip ([14], and [15] for detailed discussions).

However, using GPS data has its own restrictions. For example, inaccuracy in data could happen because of receiver’s noise and clock errors. Depending on the number of available satellites, the atmospheric conditions and the local
environment (bridges, tunnels), the GPS devices may report an inaccurate location or miss the location of some points causing gaps in the data. Wolf et al. [14] specify that an accuracy level of 10 meters is needed for map-matching GPS points in metropolitan areas with a high level of certainty. Wolf et al. [14] verified data collected in Atlanta and realized that the best performance receivers achieved this accuracy level of 10 meters for 63% of the GPS points on average. Nielsen [7] presented that 90% of the journeys collected in the Copenhagen region had missing data.

Another restriction of passive monitoring techniques is that the data is saved in a series of GPS points and data processing such as map-matching and trip end identification is important for restructuring the journeys. Furthermore, as mentioned, there is missing data or gaps which should be measured by the user. Marchal et al. [5] proposed a map-matching process for huge choice sets. They assessed the performance in terms of the computational time and highlighted the difficulty of assessment of accuracy since the real selected routes were unknown. Quddus et al. [9], provided an overview of map-matching methods. Du and Aultman-Hall [1] worked with journey end identification methods and manually recognized journey ends in a GPS data series and tested the performance of the methods. Lastly they concluded that the method with the best fit recognized 94% of the journey ends and there was no shadow of doubt that the data processing was extremely dependent on the accuracy of the geographical information system data base that was used.

While the GPS data has some errors as mentioned, it is commonly used for route choice study. For instance Nielsen [7] used 100,000 observations in the GPS data set in Copenhagen in order to realize route choice behavior and responses to road pricing alternatives. He highlighted the difficulties related to missing data and technical problems in his study.
1.3. DATA COLLECTION TECHNIQUES

1.3.2 Travel time data collection

There are several possible data collection techniques for travel time measurement. Turner et al. [12] introduces four different categories for the travel time data collection techniques:

1. "active” test vehicle techniques,
2. license plate matching techniques,
3. "passive" ITS probe vehicle techniques, and
4. emerging and non-traditional techniques.

In order to choose the best technique for a study, there are several criteria that should be considered. It is crucial that the required finance of the chosen technique would be affordable. The efficiency of the technique is another considerable parameter which could be evaluated by ratio of cost per unit of data. The level of required skill for the people involved in the data collection process. Turner et al. [12] compare the first three categories of travel time data collection techniques, illustrated in Table 1.1. They also summarized the advantages and disadvantages of these three categories, see in Table 1.2.
<table>
<thead>
<tr>
<th>Technique</th>
<th>Initial or Capital Costs</th>
<th>Operating Costs (per unit of data collected)</th>
<th>Required Skill or Knowledge Level</th>
<th>Data Reduction and/or Processing</th>
<th>Route Flexibility</th>
<th>Accuracy and Representativeness</th>
<th>Sampling Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Time Space Veh</td>
</tr>
<tr>
<td>Test Vehicle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual</td>
<td>low</td>
<td>high</td>
<td>low</td>
<td>poor</td>
<td>excellent</td>
<td>fair</td>
<td>low moderate low</td>
</tr>
<tr>
<td>DMI</td>
<td>moderate</td>
<td>moderate</td>
<td>moderate</td>
<td>good</td>
<td>excellent</td>
<td>good</td>
<td>low high low</td>
</tr>
<tr>
<td>GPs</td>
<td>moderate</td>
<td>moderate</td>
<td>moderate</td>
<td>good</td>
<td>excellent</td>
<td>good</td>
<td>low high low</td>
</tr>
<tr>
<td>License Plate Matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual</td>
<td>low</td>
<td>high</td>
<td>low</td>
<td>poor</td>
<td>fair</td>
<td>low</td>
<td>low moderate</td>
</tr>
<tr>
<td>Portable Computer</td>
<td>moderate</td>
<td>moderate</td>
<td>moderate</td>
<td>good</td>
<td>good</td>
<td>moderate</td>
<td>low high</td>
</tr>
<tr>
<td>Video with Manual Transcription</td>
<td>low</td>
<td>moderate</td>
<td>low</td>
<td>fair</td>
<td>fair</td>
<td>excellent</td>
<td>high low high</td>
</tr>
<tr>
<td>Video with Character Recognition</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>good</td>
<td>fair</td>
<td>excellent</td>
<td>high low high</td>
</tr>
<tr>
<td>ITS Probe Vehicle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signpost-Based</td>
<td>high</td>
<td>moderate</td>
<td>high</td>
<td>good</td>
<td>poor</td>
<td>good</td>
<td>moderate low low</td>
</tr>
<tr>
<td>AVI</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>good</td>
<td>poor</td>
<td>good</td>
<td>excellent high low moderate</td>
</tr>
<tr>
<td>Ground-based Radio Navigation</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>fair</td>
<td>good</td>
<td>good</td>
<td>moderate</td>
</tr>
<tr>
<td>GPs</td>
<td>moderate</td>
<td>low</td>
<td>high</td>
<td>fair</td>
<td>good</td>
<td>good</td>
<td>moderate moderate</td>
</tr>
<tr>
<td>Cellular Phone Tracking</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>fair</td>
<td>good</td>
<td>good</td>
<td>high moderate</td>
</tr>
</tbody>
</table>

Rating scales are relative among the techniques: [high, moderate, low] or [excellent, good, fair, poor].

Notes:

1 Assumes that adequate quality control procedures are used.
2 Assumes that necessary equipment is purchased (as opposed to contracting data collection services).
3 Assumes that vehicle-to-vehicle communication infrastructure does not exist.
### 1.3. DATA COLLECTION TECHNIQUES

Table 1.2: Advantages and Disadvantages of Travel Time Data Collection Techniques, Turner et al. [12]

<table>
<thead>
<tr>
<th>Data Collection Technique</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Vehicle</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Manual                    | - low initial cost  
- low required skill level | - high operating cost per unit of data  
- limited travel time/delay information available  
- limited sample of motorists |
| Electronic DMI            | - moderate initial cost  
- very detailed speed/delay data available | - lacks geographical referencing (e.g., GIS)  
- limited sample of motorists |
| GPS                       | - moderate initial cost  
- data easily integrated into GIS  
- detailed speed/delay data available  
- can provide useful data for travel surveys | - reception problems in urban “canyons”, trees  
- limited sample of motorists |
| **License Plate Matching**|            |               |
| Manual                    | - low initial cost | - high operating cost per unit of data  
- accuracy may be questionable  
- data reduction time-consuming |
| Portable Computer         | - low operating cost per unit of data  
- travel times from large sample of motorists  
- continuum of travel times during data collection | - accuracy problems with data collection, spurious matches  
- limited geographic coverage on single day |
| Video with Manual Transcription | - travel times from large sample of motorists  
- continuum of travel times during data collection | - data reduction time-consuming  
- limited geographic coverage on single day |
| Video with Character Recognition | - low operating cost per unit of data  
- travel times from large sample of motorists  
- continuum of travel times during data collection | - high initial costs (if equipment purchased)  
- limited geographic coverage on single day |
| **ITS Probe Vehicle**     |            |               |
| Signpost-based Transponders | - low operating cost per unit of data | - typically used for transit vehicles (includes loading/unloading times)  
- sample dependent on equipped vehicles |
| AVI Transponders          | - low operating cost per unit of data  
- very accurate | - very high initial cost for AVI infrastructure  
- limited to instrumented locations  
- sample dependent on equipped vehicles |
| Ground-based Radio Navigation | - available consumer product | - typically used for transit vehicles  
- sample dependent on equipped vehicles |
| GPS                       | - increasingly available consumer product  
- low operating cost per unit of data | - sample dependent on equipped vehicles  
- privacy issues |
| Cellular Phone Tracking   | - widely available consumer product | - accuracy questionable for detailed applications  
- privacy issues |
Chapter 2

Experimental conditions / theoretical models

Typically, the points collected through the Global Positioning System do not match the digital maps. In our studies, we need to use an efficient method to map them on the network. There are several literatures presenting different methods called map matching.

Finally, we need to use an algorithm to map the collected GPS points on the digital network. Krumm et al. [4] present a simple closest link matching as a applied solution in order to map reported GPS points to the network. Through the algorithm, some paths, near to the GPS points, which connect the origin and the destination, are considered. The sum distance among the reported points and the closest point on each considered path is calculated. The path with the minimum computed distance is identified the matched path. In this study, we use this method for map-matching with a few modifications in the second paper.

In these two papers, we need to develop a method to estimate our proposed models with positive random link costs. Rather than doing hard computations to find the maximum likelihood estimate, we use an indirect inference based
method. Assuming the true model can be easily simulated. In this approach, by selecting the parameters of the true model such that the simulated and real-world data sets look similar from the auxiliary model point of view. Therefore, we will be able to consistently estimate the parameters of the true model.

2.1 Borlänge Network

In the both paper we were working on Borlänge city network. As it is shown in Figure 2.1, a river passes the city and it leads to have a number of bridges. The network consists of 7459 links defined by [2].

Figure 2.1: The Borlänge road network

The first paper focuses on the length of the links as the route choice attribute. In Table 2.1 a summary is presented for the descriptive statistics of the length value in the network.
2.2. The Model

In this section the model used in the both papers is presented. As already mentioned, we have the network \( \mathcal{N} \) which consists of sets of nodes \( v \) and links \( l \). Each link connects a source node \( v^o \) to a destination node \( v^d \). A path between two nodes can be specified by a sequence of links \( \{x_{l_1}, \cdots, x_{l_n}\} \), where

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

Thus, the path could be identified by the index of links \( \pi = \{l_1, \cdots, l_n\} \). A vector presents the characteristics of each link and denotes by \( x_{l_i} \). All links have their own strictly positive cost function \( c(x_{l_i}; \epsilon_{l,i}; \beta) \). This cost is defined as the cost related to each link \( l \) for each individual \( i \). It should be mentioned that the cost function consists of different components. \( \epsilon_{l,i} \) presents an individual specific random link cost and \( \beta \) is the vector of coefficients for the links, should be estimated. In these two papers, the cost function is assumed to have a linear deterministic component.

\[
c(x_{l_i}, \epsilon_{l,i}; \beta) = \beta x_{l_i} + \epsilon_{l,i}, \quad (2.1)
\]

As mentioned, each path contains a number of links; therefore, it is assumed that the cost function of each path \( \pi \) is additive in link costs. Hence, the cost of a path can be calculated from the summation of all the link costs over the path. In other words, the cost for individual \( i \) to travers a path \( \pi \) is
calculated by

\[ C_i(\pi) = \sum c(x_l, \epsilon_{li}) \]  \hspace{1cm} (2.2)

Furthermore, it is assumed that the travelers know both the link characteristics and their idiosyncratic random utility \( \epsilon_{li} \) related to their traversed links. Since the travelers want to maximize their utility, they will select the path with the lowest generalized cost based on the proposed model.

\[ \pi_i = \arg \min_{\pi \in \Omega(v^o_i, v^d_i)} C_i(\pi) \]  \hspace{1cm} (2.3)

Assuming \( v^o_i \) as the origin and \( v^d_i \) as the destination, \( \Omega(v^o_i, v^d_i) \) will represent all the possible paths between the traveler’s origin and destination forming the choice set of this case.

In these two papers, it is assumed that the random part \( \epsilon_{li} \) has a truncated normal distribution. Additionally, we define a constraint on the values that the cost function can return to avoid having negative link costs. In practice for our study \( \epsilon_{li} \) is assumed to follow 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.

### 2.3 Indirect inference

In our both paper, we implement a simulation based method called Indirect Inference. In this approach we inference the parameters of economic models with too difficult to evaluate or analytically intractable likelihood functions [11]. As a simulation-based method, a main requirement of the indirect inference approach is that the model should be able to simulate data for different values of the involved parameters in the model. In addition, the indirect inference method applies an auxiliary model to formulate a criterion function.
There are two prerequisite for selecting an auxiliary model. First, since this model should help us to simulate datasets and be capable to be run repeatedly, the auxiliary model should be easy to estimate. Second, this model should be flexible enough to capture the variation of the observed data. Practically in most cases, auxiliary models have more parameters than the real models. However in our papers, we use the auxiliary models with the same number of parameters to the true models.

The auxiliary model does not required 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. Briefly, it chooses features of the data which are points of interest in our analysis.

We have a relationship which is defined as a binding function introducing the smoothed relation between different values of parameter for the true model and their corresponding estimated auxiliary parameters’ through the auxiliary model.

The purpose of the indirect inference is to choose parameters of economic model such that the simulated and observed data look the similar from the auxiliary model’s viewpoint.
Chapter 3

Results and Conclusions

The highest conclusion from these studies is that indirect inference is a valuable tool to estimate a route choice model and speed values on all the links in a network. Our indirect inference based method can be applied for estimating speeds using sparse GPS data with measurement errors. The Monte Carlo evidence confirms that, using the indirect inference based method for estimating route choice and link speed is a worthwhile solution.

3.1 Paper I

In the first paper, a method is presented to consistently estimate a flexible route choice model with sparse GPS data and measurement errors. The indirect inference approach is used as a structured algorithm to estimate a model with random link costs, for which the likelihood function is hard to estimate. Rather, we utilize a much simpler likelihood function related to the auxiliary model. Then estimation is done over simulation. It is required that the same data transformations (map-matching) have been performed on both real and simulated data, the introduced estimator is powerful enough to correct the estimates for bias otherwise caused by such transformations.

The results in the first paper show that the proposed indirect inference
based method works well to estimate route choice models. Although we just considered length as the parameter of our route choice model, the method is applicable for more parameters.

### 3.2 Paper II

In the second paper we have designed an iterative method to estimate the parameters of the route choice model and link speed both which have been brought together into an algorithm. First, the links are classified based on their speed limits into a number of classes. Then, the average speed of each class is consistently estimated. We have applied a route choice model which is link based with random costs, and where the path cost is additive in link costs. The route choice model describes how routes are chosen in the network, which is crucial to understand when the paths are not known, using sparse GPS data.

The results in the second paper also show that the proposed indirect inference based method simultaneously and consistently estimate the link length and route choice model parameters.
Bibliography


List of papers


II. Fadaei Oshyani, M.; Sundberg, M.; Karlstrom, A. (2013), 'Consistently estimating link speed using sparse GPS data with measured errors'.

The author of this thesis is the main contributor to these two papers.