Centre for Transport Studies
STOCKHOLM

Estimating flexible route choice models using sparse data

Masoud Fadaei Oshyani (KTH)
Marcus Sundberg (KTH)
Anders Karlström (KTH)

CTS Working Paper 2013:11

Abstract
GPS and nomad devices are increasingly used to provide data from individuals in urban traffic networks. In many different applications, it is important to predict the continuation of an observed path, and also, given sparse data, predict where the individual (or vehicle) has been. Estimating the perceived cost functions is a dicult 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 a indirect inference 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.

Keywords: GPS, route choice model, indirect inference, sparse data, statistical estimation problem.

Centre for Transport Studies
SE-100 44 Stockholm
Sweden
www.cts.kth.se

vti

KTH

Estimating flexible route choice models using sparse data

Masoud Fadaei Oshyani
Marcus Sundberg
Anders Karlström*

KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

May 3, 2013

Abstract

GPS and nomad devices are increasingly used to provide data from individuals in urban traffic networks. In many different applications, it is important to predict the continuation of an observed path, and also, given sparse data, predict 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 a indirect inference approach. We demonstrate the feasibility of the proposed estimator in a model with random link costs, allowing

*Corresponding author: masoodfo@kth.se

for a natural correlation structure across paths, where the full choice set is considered.

1 INTRODUCTION

A growing range of data gathering techniques is increasingly available to be used in transportation planning. For example, the Global Positioning System (GPS) and nomad devices are used for the collection of spatial-temporal data regarding movements of individuals. Due to both technical restrictions and privacy concerns, such data often have a sparse structure in space and time. Individuals' whereabouts are not registered at every instant; rather they are sampled in space and/or time. For many different applications, such as route guidance, traffic management for collaborative filtering applications, it is of major interest to predict the immediate future path, and also to be able to infer where the individual (or vehicle) has been. In this context, the estimation of a route choice model is of fundamental interest.

There are several problems that make route choice estimation difficult from a statistical and practical point of view. For instance, considering route choice as a choice among discrete alternatives, it is clear that the choice set can be very large. Another problem is that we would like to consider paths with many overlapping paths as close substitutes. This is typically achieved by allowing for a flexible correlation structure among paths. The focus in this paper is to address two further problems. First, the data may be sparse, being sparsely sampled in time and space. For route choice estimation, this is a problem since the path cannot be fully observed. Second, the sampled points may be sampled with errors. The purpose of this paper is to address all these problems using an indirect inference approach.

In this paper, we focus on GPS data, which has previously attracted

attention in the transportation science literature (see, e.g. [5], [9], [21], [19] and [20]). GPS technology allow for collection of large data sets regarding mobility and route choices of individuals (Jan et al. [13]). Methods and algorithms associated with the use of sparse GPS data has also previously been addressed by Lou et al. [17]. However, it should be noted that the advent of passive monitoring of route choices has provided different data collection methods (Murakami and Wagner [20]), and that the estimation methods proposed in this paper can equally well be applied to data sets collected using other technologies.

Considering the statistical estimation problem, route choice models can be formulated and estimated in a discrete choice framework. In this framework, the notion of choice sets is fundamental. Typically, in a route choice context, the potential choice set is very large, and it is infeasible to enumerate in realistic sized applications. Different approaches have been suggested to deal with this problem. Sometimes it is possible to arrive at statistically consistent estimates by sampling the choice set, for instance in the case of multinomial logit models (McFadden [18], Frejinger et al. [10]), or Multivariate Extreme Value models (Guevara [12]). To deal with the problem of correlation between overlapping paths, path-based approaches using discrete choice methods are elaborated in the literature, see, e.g., Ben-Akiva and Ramming [3], Ben-Akiva and Bierlaire [2], and Cascetta et al. [4]. Alternatively, one may view a route as a sequence of links (Dial [6]). Link costs are assumed to be random (at least to the modeler), and the path cost is assumed to be given by the summation of link costs. In practice, Fosgerau et al. [7] and Karlström et al. [14] use such link-based route cost models. In this approach, the random terms are associated to the links, rather than the paths, which induce a natural correlation structure between paths, since overlapping paths are sharing the same random components. Also, the ap-

proach allow for statistically consistent estimation with consideration of the full choice set.

The link-based approach is used in this paper as well, where it is assumed that the individual is choosing the shortest path. While route choice models traditionally are estimated using observations of chosen paths, the estimator proposed in this paper enables the estimation of flexible route choice models using sparse GPS data, with only partially observed paths. Most closely related to our work are Frejinger and Bierlaire [9] and Newman et al. [21], who also consider GPS data for route choice estimation. Yet, our approach allows for the estimation of models with a flexible correlation structure, full choice set consideration, and corrects for potential bias due to data manipulations.

The estimation procedure that is proposed in this paper is based on indirect inference, which is a simulation based estimation method (Gourieroux et al. [11], and Smith [22]). The methodology is useful even when the likelihood functions are intractable or even impossible to specify. Like other simulation-based methods, a major prerequisite of the indirect inference approach is that it should be possible to simulate data from the model of interest for different values of the parameters involved.

The main characteristic of the indirect inference method is the use of an approximate or auxiliary model in order to form a criterion function. The number of parameters of auxiliary models has to be more or at least equal to the number of parameters in 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 for the model

of interest such that the simulated and observed data look the same from the auxiliary model's point of view. Karlström et al. [14] proposed an indirect inference estimator for the estimation of flexible route choice models, in this paper the methodology is extended to address the problem of sparse GPS data with measurement errors.

2 METHOD

We propose a method to estimate a model with random link costs, i.e. a flexible route choice model. Instead of doing hard computations to find the maximum likelihood estimate, our method is based on the principle of indirect inference and we use the fact that the true model can be easily simulated. In this indirect 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 an auxiliary model, we will be able to consistently estimate the parameters of the true model. First, we specify the model that we want to estimate, and then our indirect inference based estimator will be introduced. In subsequent sections we will examine the properties of the estimator in Monte Carlo simulation experiments.

2.1 A flexible route choice 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 (vertexes) 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 to a destination node. Let $s(l)$ and $d(l)$ denote the source node and destination node, of a link, respectively. The path between a source node and a destination node could be seen as a sequence of links, where $s(l_1) = v^o$, $d(l_j) = s(l_j + 1)$ for

$j = n - 1$, and $d(l_n) = v^d$. Here, v^o denotes the origin node of the path and the destination node is v^d.

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

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. ϵ_{li} is an individual specific random link cost and β is the vector of coefficients for the links, ought to be estimated, and x_l is the corresponding vector of link characteristics. In this paper, the cost function is assumed to have a linear deterministic component.

$$c(x_l, \epsilon_{li}) = \beta x_l + \epsilon_{li}, \tag{1}$$

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 π 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 π is computed by

$$C_i(\pi) = \sum_{l \in \pi} c(x_l, \epsilon_{li}). \tag{2}$$

Clearly, route costs are naturally correlated for overlapping routes since they share links.

Furthermore, we assume the travelers know both the link characteristics and their idiosyncratic random utility ϵ_{li} regarding their passed links. Since the choice makers tend to maximize their utility, they will choose the path

with the lowest generalized cost in this model

$$\pi_i = arg \min_{\pi \in \Omega(v_i^o, v_i^d)} C_i(\pi). \tag{3}$$

Assuming v_i^o as the origin and v_i^d as the destination, $\Omega(v_i^o, v_i^d)$ represents all the possible paths between the traveler's origin and destination i.e. it represents the choice set.

The random part ϵ_{li} can have an arbitrary distribution like normal, gamma or exponential distributions. When implementing the model, we assumed that the random cost component followed a truncated normal distribution. It is common to introduce a constraint on the values that the cost function can return, in order to avoid negative link costs. Here we assumed that ϵ_{li} follows a truncated standard normal distribution using only the positive values. Given the above assumptions, the prerequisites of the Dijkstra shortest path algorithm (always necessitates a positive link cost on the network) are satisfied. Thus, in order to simulate path choices in accordance with (3), we apply the Dijkstra algorithm.

The final aim of this paper is to estimate β in equation (1). The likelihood function for this model is complicated and hard to estimate; therefore, a simpler method is required. We employ a simulation based strategy built on indirect inference, for this purpose we simulate GPS points based on routes provided by the flexible route choice model.

2.2 Simulating GPS data

For simulating GPS data, we assume that all the devices in our study have similar error characteristics. Furthermore, there is likely some error in the mapped locations of the roads, which contributes to the deviation between the measured location and the location on the map. We model the overall error, in latitude and longitude, as a two-dimensional symmetric Gaussian

with zero mean. This means that we require an estimate of the standard deviation of the distance between the measured location and the actual point on the map.

As explained in the introduction part, we consider low frequency GPS points as observed trip data. Given any parameter β we can simulate routes as described in the previous section. Then, the following algorithm is implemented to create a set of GPS points corresponding to each trip. The used procedure is defined through the following steps:

1. Given some simulated path π_i in concordance with (3).

2. The average speed of cars in the network is assumed to be 32 km/h. Based on this, the location of a virtual car is determined every s seconds on the simulated path.

3. In order to consider GPS errors, The locations of the car are distorted by adding a measurement error. We assume that the distance error follows a symmetric normal distribution with zero mean and a given standard deviation.

4. Add the simulated GPS errors to the simulated locations of vehicles to obtain simulated GPS points.

Now, given GPS data we need a methodology to match GPS points to specific paths. This might seem like a detour, first we generate paths by the route choice model, then simulate GPS-points given those paths, and now we suggest that those GPS-points should be matched back to some paths. Yet, this is an important step to take. Our observations are in GPS format and so are the simulated GPS points. Then, we make sure that the same data manipulations are performed on both real data and simulated data, this will allow the indirect inference estimator to correct for any bias introduced through these data manipulations.

2.3 Map-Matching

As a natural characteristic, the points reported through the GPS do not match directly to a network on a digital map. Therefore, we have to apply some method to map the reported GPS points on the network. Krumm et al. [16] introduce a simple nearest road matching as a practical solution in order to map collected GPS points to the road data. Through the algorithm, a number of paths, close to the GPS points, which start from the origin and end at the destination are considered. The total distance between the measured points and the nearest point on each considered path is computed. The path with the minimum calculated distance is stated the matched path. In this paper, we use this method for map-matching.

The following algorithm is used to match a set of GPS points corresponding to each trip to the road map.

1. Take the network data and the given set of GPS points.
2. Identify the origin and destination of the trip.
3. Calculate the summed distance between GPS points and potential paths.
4. Find the path with minimum distance to GPS points and return as the matched path

2.4 Indirect inference

In this section, the indirect inference method to estimate the proposed route choice model is introduced. Indirect inference make use of an auxiliary model, which should be easily estimated, yet rich enough to capture relevant variation in the data. In the context of route choice modeling we have a number of natural candidates which could be used as auxiliary models, e.g. the multinomial logit (MNL), path-size logit [1] or C-Logit [4].

In this study, the multinomial logit (MNL) model is used as an auxiliary model, with the objective of showing that even a simplistic auxiliary model allow us to consistently estimate our flexible route choice model. As stated by Smith [22]: " The auxiliary model does not need to be an accurate description of the data generating process. Instead, the auxiliary model serves as a window through which to view both the actual, observed data and the simulated data generated by the economic model: it selects aspects of the data upon which to focus the analysis".

Our MNL-model use the same number of parameters as the true model. From the MNL perspective on route choice, each individual i is choosing a route r from a set of alternative routes $I_i(o,d)$, each route is given a utility of U_{r_i} as follows:

$$U_{r_i} = \theta X_r + \epsilon_{r_i}, \tag{4}$$

where θ is the vector of the auxiliary parameters , X_r represents route characteristics and ϵ_{r_i} are assumed i.i.d Gumbel distributed. $I_i(o,d)$ is the auxiliary individual specific choice set, representing some paths between the origin and the destination.

Considering that our auxiliary model is an MNL model, we need to generate choice sets corresponding to the origin-destination pairs which are found in data, in order to estimate the auxiliary parameter. To generate the auxiliary individual specific choice sets, a pseudo-universal choice set is generated by simulations of shortest paths from the flexible route choice model, for each of the relevant OD-pairs. We use random samples from the pseudo-universal choice sets as the individual specific choice set[i]. Regarding

[i] In our application of th estimator we create pseudo-universal choice sets with a maximum size of 800 paths, out of which 201 are sampled into the auxiliary individual specific choice set.

the proposed auxiliary model, estimations of the auxiliary parameter θ using choice set sampling are rather easy to retrieve, but in our case they are biased due to misspecification. Still, based on the experience of Karlström et al. [14], we can apply this auxiliary model to the real and simulated data sets and arrive at an estimate of the true parameters β.

In principle, there is a correspondence between the parameters of the flexible route choice model and the auxiliary parameters which is manifested through a smooth binding function $\theta(\beta)$. As we shall see, indirect inference exploits the binding function, for this purpose we have to explore this function. This exploration is done by simulations. Consider a set of arbitrary structural parameters $\beta_m, m = 1, \cdots, M$ which are drawn from some specified domain, corresponding to each parameter M different data sets can be simulated by means of our flexible route choice model, and GPS point simulation. Provided the simulated GPS points we can employ the map-matching method in order to convert GPS data into matched path choices. We denote the set of matched paths as $\tilde{y}$, and these matched observations are actually depending on the data generating parameter β_m. Thus, all matched simulated choices $\tilde{y}(\beta_m)$ in a particular data set is generated by the same parameter β_m. For any such parameter we can estimate the corresponding auxiliary parameter by conventional maximum likelihood

$$\tilde{\theta}_m(\beta_m) = arg \max_{\theta} \mathcal{L}(\tilde{y}(\beta_m); x, \theta). \tag{5}$$

The mapping $\tilde{\theta}_m(\beta_m)$ is discontinuous due to the discrete nature of the matched paths $\tilde{y}(\beta_m)$. A smooth binding function is estimated by local regression or OLS, based on M different given values of β_m and their corresponding $\tilde{\theta}_m(\beta_m)$. We denote this smooth binding function by $\tilde{\theta}(\beta)$.

The aim of the indirect inference approach is to select parameters of the flexible route choice model such that the simulated and observed data look the same from the auxiliary model's point of view. We use the likeli-

hood function based on the auxiliary model, observed GPS data which are matched to paths (y), and network characteristics (x)

$$\hat{\beta} = arg \max_{\beta} \mathcal{L}(y; x, \tilde{\theta}(\beta)). \tag{6}$$

That is, rather than directly stating a criterion function for the model of interest, we use the criterion function of a much simpler, misspecified, MNL model in order to indirectly infer the parameter of the underlying data generating process. Again, we make sure that the same data manipulations are performed on both real data and simulated data, we perform map-matching to retrieve the matched paths both in the case of real data (y) and for simulated data ($\tilde{y}(\beta_m)$). Therefore the map-matching procedure should be viewed as an integral part of our auxiliary model. Proceeding in this manner will allow the indirect inference estimator to correct for potential biases introduced through such data manipulations. The above description of the indirect inference approach follows closely to that of Keane and Smith [15]. The main difference from their approach is that we use OLS regression of the binding function to smooth the objective function.

2.5 Estimation algorithm

In this section, we provide a summary of the proposed estimator in the form of an algorithm. Since a simple MNL model is proposed as our auxiliary model and the input data sets for the estimation of this model preferably should 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 map-matching method in section 2.3 in order to translate the GPS data into paths.

The main characteristic of our study is the fact that we consider low frequency GPS data with measurement errors as observed trips data. In

order to estimate a flexible route choice model given such data we take the following steps:

1. Given GPS points (gp^{obs})
2. Use the map-matching method to find unique matched paths corresponding to the GPS points y.
3. Draw M points β_m from a given domain $\mathcal{D}$.
4. Simulate N path choices.
5. Simulate GPS data.
6. Match simulated GPS to N path observations for each β_m, M different data sets $\tilde{y}(\beta_m)$ are generated.
7. Employ the auxiliary model to compute the corresponding auxiliary parameters $\tilde{\theta}_m(\beta_m)$.
8. Estimate the smooth binding function by the data $\{\tilde{\theta}_m, \beta_m\}_{m=1}^{M}$ using OLS regression.
9. Insert the estimated binding function into equation (6) and find the indirect inference estimate of the true parameter $\hat{\beta}$.

The final algorithm is defined below:

Indirect inference route choice estimator

Require: network $\mathcal{N}$

Require: Individual choice sets $I_i(o,d)$

Require: Observed GPS points (gp^{obs})

for m=1,$\cdots$,M **do**

Draw $\beta_m \in \mathcal{D}$

for $i = 1, \cdots, N$ **do**

Generate simulated route choice π_i

Generate simulated GPS data given route π_i

end for

Do map-matching and state simulated matched paths $\tilde{y}(\beta_m)$

Estimate auxiliary parameters $\tilde{\theta}_m(\beta_m) = arg\max_\theta \mathcal{L}(\tilde{y}(\beta_m); x, \theta)$

end for

Given $\{\tilde{\theta}_m, \beta_m\}_{m=1}^M$, estimate $\tilde{\theta}(\beta)$ using linear OLS

Do map-matching for gp^{obs} to get matched paths y

Estimate $\hat{\beta} = arg\max_\beta \mathcal{L}(y; x, \tilde{\theta}(\beta))$

3 Case study

Using a network representing the real world road network of Borlänge city, Sweden, we simulate a data set using the flexible route choice model, and use our developed estimator to see whether the parameter can be recovered. For the illustrations we use the network of Borlänge city as described by Frejinger and Bierlaire. [8]. The network contains 3077 nodes and 7459 links.

Suppose that link length L_l is the only attribute which is taken into

account. Hence, the cost of traversing link l, for individual i is given by

$$c(L_l, \epsilon_{li}) = \beta L_l + \epsilon_{li}, \tag{7}$$

and the corresponding cost of a path is simply formed by adding link costs. Then, individuals are assumed to choose the shortest path. This summarizes our flexible route choice model. The methodology introduced in this paper can be used for estimating cost functions including multiple attributes, see Karlström et al. [14] for such indirect inference estimation based on path observations.

For the purpose of indirect inference, an auxiliary model is introduced. This model is assumed to take the standard MNL form, where the utility of a path r is given by

$$U_r = \sum_{l \in r} \theta L_l + \epsilon_{li} \tag{8}$$

Given the network and the route choice model, we may assign a value to β, (say $\beta = 1$) and simulate data sets. First we simulate $N = 3000$ paths, then, given paths we simulate GPS observations along those paths. Thereafter, the proposed estimator based on indirect inference including map-matching is applied. We investigate if the estimated parameter is consistent with the assigned parameter of the data generating process (i.e. 1). Briefly, in the next section, consistency of the proposed estimator is evaluated, does it retrieve the "unknown" value of β?

4 Results

In this section we provide Monte Carlo evidence to show the feasibility and accuracy of the proposed estimator. In a real world application of the estimator the initial guess of the parameter domain $\mathcal{D}$ would have to be chosen arbitrarily. It may be the case that the chosen domain 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, thus our local estimate of the binding function may provide a poor estimate. As a remedy, in practice, Karlström et al. [14] implement their estimator in a stage-based format. For each stage the interval of the domain shrinks, and the domain is recentered around the estimate provided by the previous stage. Thus, the first stage may provide crude estimates, but the estimates are improved through the stages.

In this paper, since we know the "true" parameter, we skip stages and assume that we are in a last stage and that an appropriate interval for generating the binding function is known, it is $\mathcal{D} = (0.9, 1.1)$. The valuation of the link length attribute, β, is estimated based on a binding function which itself is estimated using $M = 5$ sample points β_m drawn from $\mathcal{D}$. For each such β_m, $\tilde{\theta}_m$ is estimated based on N=3000 simulated paths. The "observed" paths also contains $N = 3000$ routes which are created based on $\beta = 1$, thereafter GPS related spatial error is introduced following a normal distribution $\epsilon_{GPS} \sim N(0, \sigma)$ for various levels of σ. The estimator applies a pseudo-universal choice set of size 800, and the size of the auxiliary individual specific choice sets is 201.

All Monte Carlo statistics are calculated based on 10 independent estimations of the parameter β. That is, we create ten independent sets of "observed" GPS data and then we apply the estimator once to each of these data sets.

In Table 1 we report the Monte Carlo evidence with a GPS sampling interval of 30 seconds. The true parameter is assumed to equal 1. GPS errors are assumed to have a Gaussian distribution with standard deviation of 0, 10, 25, or 50 meters. As is evident, the estimator is quite precise also with rather large GPS errors. For each of the estimates, the true data gen-

Table 1: Monte Carlo evidence: Estimates of β for different levels of GPS related errors, with GPS sampling every 30 seconds.

GPS error(m)	0	10	25	50
Mean	0.99	0.96	1.02	1.01
Std	0.02	0.02	0.04	0.01
RMSE	0.02	0.05	0.04	0.45
ZETA	-0.73	-2.07	0.50	0.50

erating parameter falls within the 95% confidence interval of the estimated parameter. Thus the true value $\beta = 1$ cannot be rejected.

Table 2: Monte Carlo evidence: Estimates of β for different levels of GPS related errors, with GPS sampling every 120 seconds.

GPS error(m)	0	10	25	50
Mean	1.02	1.01	1.02	0.99
Std	0.02	0.02	0.02	0.02
RMSE	0.03	0.02	0.03	0.02
ZETA	0.87	0.62	1.04	-0.68

In Table 2 we report corresponding results where we assume that the GPS points are sampled every 120 seconds. The precision is good, with somewhat weaker results in the case where the GPS error is also large. Still, the true parameter is not rejected at the 95% level, this holds for all the estimates.

In Table 3 the results are compared for different sample sizes, $N = \{1000, 2000, 3000\}$, for the cases where GPS points are sampled every 120

Table 3: Monte Carlo evidence: Estimates of β for different observed sample size, with GPS sampling every 120 seconds and 10m and 25m GPS error.

GPS error	10 meters		25 meters	
Observed sample size	1000	3000	1000	3000
Mean	1.00	1.01	1.09	1.02
Std	0.03	0.02	0.05	0.02
RMSE	0.03	0.02	0.10	0.03
ZETA	0.14	0.62	1.76	1.04

seconds and the GPS error has a standard deviation of either 10 or 25 meters. The accuracy of statistical parameters improves as the sample size grows. Overall, the results in Table III show that the estimated parameter vector converges to the true value of the data as observed sample size increases.

5 CONCLUSIONS

In this paper we have proposed an estimator, for a flexible route choice model, using GPS data sampled with a low frequency. Route choice is a central concept both with regard to the analysis of travelers' behavior and the effect of such behavior upon transport systems. 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 governing route choice behavior.

We have designed an algorithm to consistently estimate a flexible route choice model in the presence of sparse GPS data and measurement errors. The indirect inference method is applied as a structured procedure to esti-

mate a model with random link costs, where the likelihood function is difficult to evaluate. Rather, we make use of the much simpler likelihood function related to the auxiliary model. Estimation is performed through simulation. By ensuring that the same data transformations (map-matching) are performed on both real and simulated data, the proposed estimator is able to correct the estimates for bias otherwise caused by such transformations.

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 with measurement errors. The Monte Carlo evidence show that, applying the indirect inference approach to route choice estimation is a worthwhile solution.

In the approach used in this paper, we do not utilize time stamps of GPS observations. This provides crucial information if we want to estimate link travel time or link speed profiles. While our results show that we are able to estimate a consistent route choice model without detail assumption about speed profiles, time stamps will provide information about speed (or travel time). The assumption of known distribution of GPS-errors is another crucial feature of the current approach. Both these issues will be addressed in future work, with the objective of estimating link specific speeds or travel times.

References

[1] Ben-Akiva, M. and Bierlaire, M., (1999). Discrete choice methods and their applications to short term travel decisions, in R. Hall (ed.). *Handbook of Transportation Science*, Kluwer, Dordrecht, The Netherlands, pp. 534.

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

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

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

[5] Chen, W., Yu, M., Li, Z., Chen, Y., Chao, J., (2003). Tight integration of digital map and in-vehicle positioning unit for car navigation in urban areas. *Wuhan University Journal of Natural Sciences*, 8(2):551–556.

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

[7] Fosgerau, M., Frejinger, E. and Karlström, A., (2011). A logit model for the choice among infinitely many routes in a network. *Technical re-*

port, Center for Transport Studies, Royal Institute of Technology KTH, Stockholm, Sweden.

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

[9] Frejinger, E. and Bierlaire, M., (2008). Route choice modeling with network-free data. *Transportation Research Part C*, 16:187-198.

[10] Frejinger, E., Bierlaire, M., Ben-Akiva, M. (2009), Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological*, 43(10):984–994.

[11] Gourieroux, C., Monfort, A. and Renault, E., (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118.

[12] Guevara, C.A., (2010). Endogeneity and sampling of alternatives in spatial choice models. PhD Dissertation, MIT, USA.

[13] Jan, O., Horowitz, A.J., Peng, Z.R., (2000). Using Global Positioning System Data to Understand Variations in Path Choice. *Transportation Research Record: Journal of the Transportation Research Board*, 1725(1):37–44.

[14] Karlström. A., Sundberg, M., Wang, Q., (2011). Consistently estimating flexible route choice models using an MNL lens. *International Choice Modelling Conference*, Leeds, UK.

[15] Keane, M., Smith, A.A., (2003). Generalized indirect inference for discrete choice models. *Manuscript* (Yale University).

[16] Krumm J., Letchner J., Horvitz E., (2007). Map matching with travel time constraints (Paper 2007-01-1102). *Society of automotive engineers (SAE) 2007 world congress*, Detroit, MI, USA.

[17] Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., (2009). Map-matching for low-sampling-rate GPS trajectories. *In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, New York, NY, USA, pp. 352–361.

[18] McFadden, D., (1978). Modeling the choice of residential location. *Modelling the Choice of Residential Location*, pp 75–96.

[19] Miller, H. J. (2005), A Measurement Theory for Time Geography. Geographical Analysis, 37: 1745. doi: 10.1111/j.1538-4632.2005.00575.x

[20] Murakami, E., Wagner, D.P., (1999). Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies*, 7:149–165.

[21] Newman, J., Chen, J., Bierlaire, M., (2009). Generating probabilistic path observation from GPS data from route choice modeling. *Proceedings of the European Transport Conference ETC*.

[22] Smith, A.A., Jr., (2008). Indirect inference. *The New Palgrave Dictionary of Economics Online, Palgrav Macmillan*, DOI:10.1057/9780230226203.0778.