Route choice modeling with network-free data*

M. Bierlaire E. Frejinger†

February 14, 2007

Report TRANSP-OR 070214
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*This research is supported by the Swiss National Science Foundation grant 200021-107777/1

†École Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland. E-mail: {michel.bierlaire, emma.frejinger}@epfl.ch

Abstract

Route choice models are difficult to design and to estimate for various reasons. In this paper we focus on issues related to data. Indeed, real data in its original format are not related to the network used by the modeler and do therefore not correspond to path definitions. Typical examples are data collected with the Global Positioning System (GPS) or respondents describing chosen itineraries to interviewers. Data manipulation is then necessary in order to obtain network compliant paths. We argue that such manipulations introduce bias and errors and should be avoided. We propose a general modeling framework that reconcile network-free data with a network based model without data manipulations. The concept that bridges the gap between the data and the model is called *Domain of Data Relevance* and corresponds to a physical area in the network where a given piece of data is relevant.

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

1 Introduction

Route choice models play a crucial role in many transport applications, for example traffic assignment and transport planning. Given a transportation network and an origin-destination (OD) pair $s = (s_o, s_d)$ a route choice model predicts the probability that any given path between origin s_o and destination s_d is selected to perform a trip. They are difficult to design and to estimate for various reasons, such as the large size of the choice set and the complex correlation structure (see the discussion by Ben-Akiva and Bierlaire, 2003).

In the paper we focus on the issues associated with data. The concept of path, which is the core of a route choice model, is usually too abstract for a reliable data collection process. Real data, in their original format, do not

correspond to path definitions. A typical example is GPS data, which are more and more available (Murakami and Wagner, 1999, Jan et al., 2000, Schönfelder et al., 2002, Axhausen et al., 2003, Frejinger, 2004, among many). As GPS devices do not explicitly use the transportation network, the coordinates of data points cannot be directly used, and data processing is required in order to reconstruct paths. In the literature, such data processing involves map matching, trip end identification and assumptions on missing data. Recently, Marchal et al. (2005) proposed a map matching algorithm for large choice sets. They evaluate the performance in terms of computation time and underline the difficulty of evaluating accuracy since the "true" chosen routes are unknown (see Quddus et al., 2003, for an overview of map matching algorithms). Du and Aultman-Hall (2007) discuss trip end identification algorithms. They manually identified trip ends in a GPS data stream and evaluate the performance of the algorithms.

Another context is when respondents are asked to describe a path that they have followed during a given trip. They are in general able to identify a sequence of locations that they have traversed, but have difficulties describing a full path in detail. For instance, Ramming (2001) (see also Bekhor et al., 2006) estimated route choice models based on data collected in Boston. The respondents described chosen routes by naming street segments. In case of incomplete or ambiguous descriptions, the routes were reconstructed by taking the shortest path between known street segments.

In this paper, we advocate that the data manipulation required by the underlying network model introduces biases and errors, and should be avoided. We propose a general modeling scheme that reconcile network-free data (such as GPS data or partially reported itineraries) with a network based model without such manipulations.

After a literature review in the next section, we introduce in Section 3 the concept of *domain of data relevance* (DDR) that is designed to be the missing link between the data and the network model. In Section 4, we describe the estimation of a route choice model using the network-free data and the DDRs and in Section 5 we provide simple examples for two different types of data. The framework is then illustrated on a real case study in Section 6.

2 Literature Review

Mail and telephone surveys are conventional methods for collecting trip data. Mahmassani et al. (1993) propose a two-stage data collection, where the second stage involves more detailed trip descriptions. Abdel-Aty et al. (1995) combine computer-aided telephone interviews and GIS capabilities specifically for route choice data. Ramming (2001) also collects route choice data, based on reported path segments. Vrtic et al. (2006) have performed telephone interviews where intermediate locations of long distance trips were reported (see Section 6).

In the past decade many studies presented in the literature compare data obtained with conventional survey methods with GPS data. There is a consensus that passive monitoring have several advantages over conventional surveys. For instance, multiple days of trip data can be collected automatically and are directly available in electronic format. However, GPS data also have issues (see Wolf et al., 1999, and Zito et al., 1995, for detailed discussions). First, constraints of the technology, such as satellite clock errors, receiver noise errors, selective availability (intentional errors inserted by U.S. Department of Defense) and type of receiver limits the accuracy of the data. Second, depending on the number of available satellites, atmospheric conditions, and local environment (high buildings, bridges, tunnels) the GPS receiver can compute an inaccurate position or fail to compute the position which introduces gaps in the data. Wolf et al. (1999) state that an accuracy level of 10 meters is required in order to map match GPS points in urban areas without ambiguity. In their tests, the best performing receiver achieves this level for 63% of the GPS points on average. Nielsen (2004) observed that 90% of the trips collected in the Copenhagen region had missing data. A third issue is that the data are stored in one stream of GPS points and data processing is required in order to reconstruct the trips. Such data processing involves map matching, trip end identification and assumptions on missing data (Marchal et al., 2005, Quddus et al., 2003). Du and Aultman-Hall (2007) found that the best performing algorithm correctly identified 94% of the trip ends. Finally, we note that the data processing is highly dependent on the accuracy of the

geographical information system data base that is used.

Frejinger and Bierlaire (2007) estimate route choice models based on a GPS dataset collected in the Swedish city of Borlänge (see Schönfelder et al., 2002, for more details on the data). The data processing was performed by the Atlanta based company GeoStats. Nielsen (2004) study route choice behavior based on a large GPS dataset collected in Copenhagen.

Based on the previous discussion, we conclude that network compliant route choice data are never available. This motivates the approach proposed in this paper, where we acknowledge this nature of the data, and model it explicitly instead of trying to fix it through various manipulations.

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

3 Domain of Data Relevance

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

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

We define an *observation* as a sequence of individual pieces of data related to an itinerary, such as a sequence of GPS points, or of reported locations. For a given piece of data, the *domain of data relevance* is defined as the physical area where the piece of data is relevant. Its exact definition depends on the context. For example, consider a GPS reporting coordinates (x, y). Due to the intrinsic technological limitations of the device, we can identify a 95% confidence interval, say, around the point (x, y). This would be the DDR of this piece of data. An example of GPS data is shown in Figure 1 where the GPS points are represented by small circles and their corresponding DDR with dashed lines. The size of the DDR areas vary depending on the accuracy (e.g. quality of satellite signals) of each piece of data.

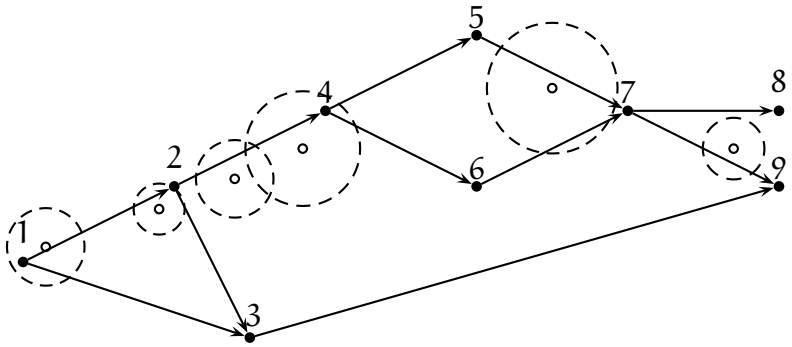

Figure 1: Example of GPS data

In the context of reported paths, notions such as "downtown", "next to the Eiffel Tower" or "intersection of Massachusetts Avenue and Newbury Street" can easily be associated with a DDR. The size of the DDR is inversely proportional to the fuzziness of the concept. It may be unambiguous (such as the area corresponding to "downtown"), or ambiguous and left to the modeler's judgment (such as "next to the Eiffel Tower"). An example is shown in Figure 2 where the reported locations are "home", "intersection Main St and Cross St", "city center" and "mall". The home and intersection correspond to exact locations in the network and the areas of the associated DDRs (dashed lines) are therefore small, they contain

only one node. The two other reported locations are more fuzzy and the areas of the associated DDRs are therefore larger, in this case the DDRs contain two nodes.

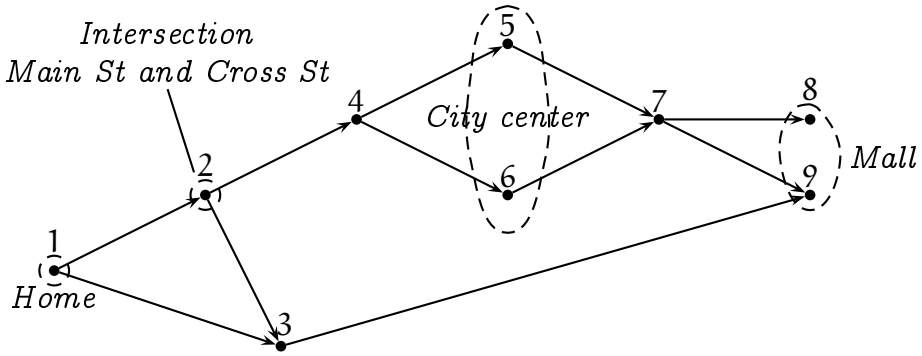

Figure 2: Example of a reported trip

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

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

4 Model Estimation

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

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

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

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

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

$$P_n(i|\mathcal{S}_i) = \sum_{s\in\mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p\in\mathcal{C}_n(s)} P_n(i|p)P_n(p|\mathcal{C}_n(s);\beta), \tag{1}$$

where

- $P_n(s|\mathcal{S}_i)$ is the probability that the actual OD pair is s given the set of relevant OD pairs $\mathcal{S}_i$,
- $P_n(i|p)$ is the measurement equation, giving the probability of observing i if the actual path is p, and
- $P_n(p|\mathcal{C}_n(s);\beta)$ is the route choice model.

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

- there is at least a link in the path related to each DDR, that is, for any $m = 1, \ldots, k$, there exists q, $1 \leq q \leq P$, such that $\delta(d_m, \ell_q) = 1$,

- the sequence of reported locations is consistent with the order of the links in the path, that is, for any $m_1 \leq m_2$, if $\delta(d_{m_1}, \ell_{q_1}) = 1$ and $\delta(d_{m_2}, \ell_{q_2}) = 1$, then $q_1 \leq q_2$.

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

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

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

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

$$P_n(s|\mathcal{S}_i) = \frac{1}{|\mathcal{S}_i|} \ \forall s \in \mathcal{S}_i. \tag{2}$$

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

As discussed in the previous section, the role of the DDR is to link the network-free data to the network. A problem may occur that need

to be addressed in order to estimate the model. Namely, the DDR of a data d can be empty, that is $\delta(d, e) = 0\ \forall e$, meaning that no network element correspond to this piece of data. In this case, the DDR is not properly defined and a new specification is necessary. A possible solution is to increase the size of the DDR so that at least one link crosses the DDR.

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

Models of type (1) can be estimated with BIOGEME (Bierlaire, 2003).

5 Illustrative Examples

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

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

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

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

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

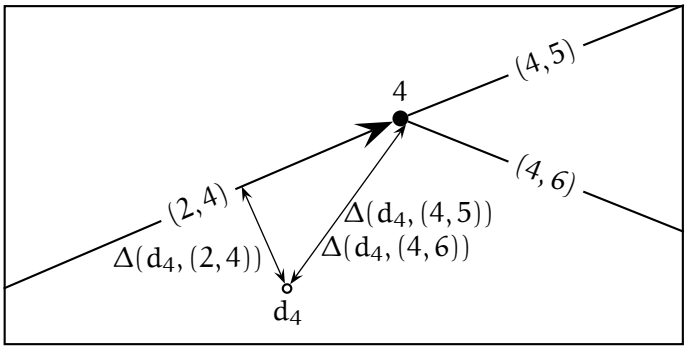

Figure 3: Example of GPS data (continued)

6 Case Study

In this section we illustrate the modeling framework on a dataset collected in Switzerland. The data concern long-distance route choice behavior and

Figure 4: Example of an observation

were collected via telephone interviews (Vrtic et al., 2006). The respondents were asked to describe their last long-distance trip with the names of the origin and destination cities as well as maximum three intermediate cities or locations that they passed through. An example is shown in Figure 4 where a traveler went from Bellemont-sur-Lausanne to Vandoeuvres passing through Morges, Aubonne and Nyon. 940 reported trips are available for route choice analysis.

In this context, the DDR of each reported location is defined by the corresponding zip code. When linking the network-free data with the network through the DDRs it is important to make sure that the precision level of the observations correspond to the precision level of the network. We therefore use a simplified transportation network (Swiss national model, Vrtic et al., 2005). This network covers all regions in Switzerland and contains 39411 unidirectional links and 14841 nodes (to be compared with the Swiss TeleAtlas network that contains approximately 1 million unidirectional links and half a million nodes). To our knowledge, this is the largest network used for estimation of route choice models based on revealed preferences data presented in the literature.

In order to estimate a route choice model we need to specify $P(s|\mathcal{S}_i)$ and choice sets $\mathcal{C}_n(s)$ $\forall s \in \mathcal{S}$. The observations contain no information on relevant OD pairs. Due to the computationally complex choice set generation we do not consider all possible OD pairs for each observation but randomly choose two OD pairs (if more than one is available) and use the probability model given by Equation (2). For each OD pair we generate a choice set of 45 routes using a stochastic choice set generation approach (Bierlaire and Frejinger, 2007). After the choice set generation there are 780 observations available for model estimation. 160 observations are not considered because either all or none of the generated routes correspond to the observation.

We estimate two different types of route choice models $P_n(p|\mathcal{C}_n(s);\beta)$, one Path Size Logit (PSL) model (Ben-Akiva and Ramming, 1998) and one Subnetwork model (Frejinger and Bierlaire, 2007). With the latter, we explicitly model the correlation among paths on a Subnetwork using an Error Component model. Here we create a subnetwork composed of all main freeways. We estimate one covariance parameter which is assumed proportional to the length by which the paths overlap with the subnetwork. The transportation network is shown in Figure 5 where Subnetwork is marked with bold lines.

Finally, we need to specify the deterministic utility functions. We use the attributes reported in Table 1. Namely, Path Size, free-flow travel time and road type attributes. The type of road is defined according to an existing hierarchy of the links. We define four road types; freeway (FW), cantonal/national (CN), main and small roads. The cantonal/national roads connect different regions in Switzerland but have a lower capacity and speed limit than freeways. Main roads refer to fast local roads in urban or rural areas and small roads are the remaining ones.

Both models have the same linear-in-parameters specifications. More precisely, a piecewise linear specification for the free-flow travel time (measured in hours) is used in order to capture travelers' sensitivity to changes in travel time in different ranges of the variable. After systematic testing of different endpoints for the ranges we have defined a specific piecewise linear approximation of the free-flow travel time for each of the four road

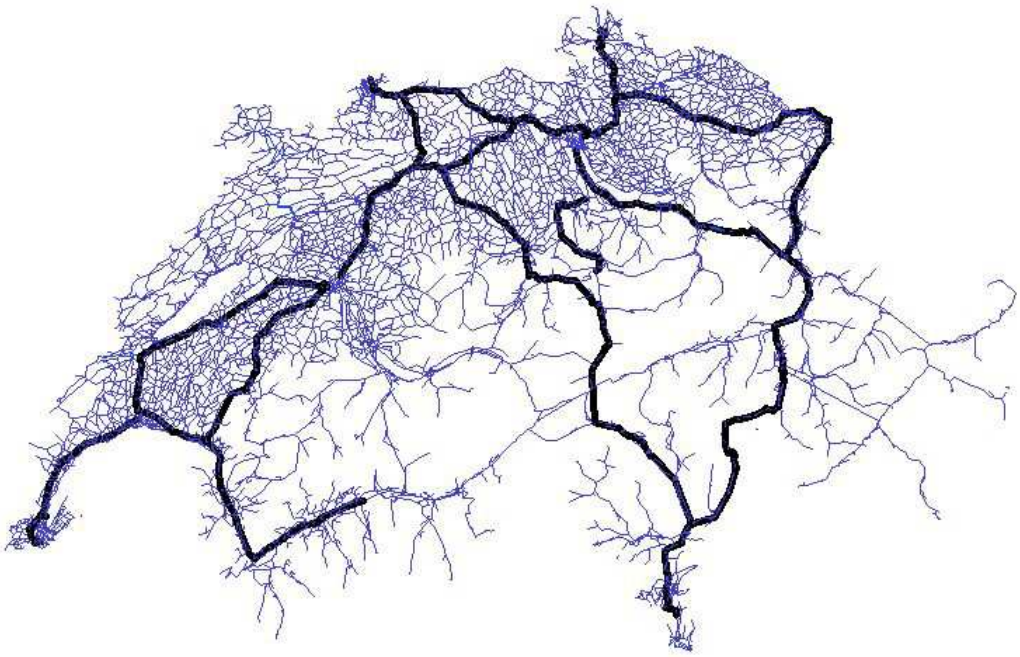

Figure 5: Swiss national network

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

Table 1: Statistics on routes corresponding to observations

types. The utility functions also include a Path Size attribute and the four variables representing the proportion of the total travel time on each type of road.

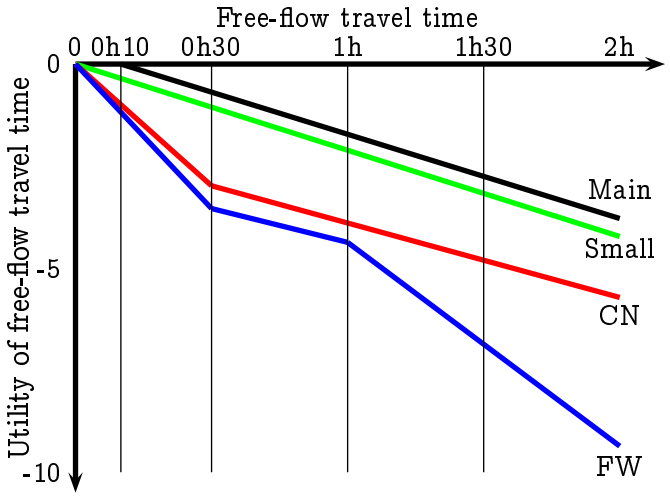

Figure 6: Piecewise linear specification - PSL model

In Figure 6 we illustrate the piecewise linear specification of the free-flow travel time by graphically visualizing the estimates for the PSL model. The coefficient estimates for all the explanatory variables are reported in Table 2. The coefficients have their expected signs and are significantly different from zero. We have provided scaled coefficient estimates in order to facilitate the comparison of the two models. The scaling is based on the "freeway free-flow time 0-30 min" coefficient. The magnitude of the scaled estimate for this coefficient is hence the same for both models. The scaled estimates have comparable magnitudes for the two models. This is also the case for the robust standard errors and the t-test statistics are therefore similar. We conclude that the estimation results are stable for the different model structures.

The model fit measures and the coefficients related to the correlation structure are reported in Table 3. The Path Size coefficient estimates are positive which is consistent with theory (Frejinger and Bierlaire, 2007). Indeed, this results in a negative correction of the utility for overlapping paths.

The covariance estimate is significantly different from zero which can be interpreted as there is a significant correlation among paths using freeways. Furthermore, the Subnetwork model has a significantly better model fit than the Path Size Logit model (the likelihood ratio test statistic is 6.756 to be compared with $\chi^2_{0.05,1} = 3.84$) which is consistent with the findings in Frejinger and Bierlaire (2007).

7 Conclusion

Link-by-link descriptions of chosen routes are never directly available and data manipulation is necessary in order to obtain network compliant paths for the estimation of route choice models. We argue that data manipulation introduces biases and errors and should be avoided. We propose a general modeling framework that reconcile network-free data (for example partially reported trips and GPS data) with a network based model without such manipulations. The concept that bridges the gap between the data and the model is called *Domain of Data Relevance* and corresponds to a physical area in the network where a given piece of data is relevant.

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

We believe that this approach makes the route choice modeling results more accurate. Moreover, it makes the estimation of the models easier since the complex and time consuming data manipulation can be avoided. We provide the methodology for estimating models based on GPS data. Since no GPS dataset in its original form (sequences of GPS points) is at our disposal, the estimation based on this type of data is left for future research.

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

Table 2: Estimation results

Coefficient	PSL	Subnetwork
ln(Path Size) based on free-flow time	**1.04**	**1.10**
Scaled Estimate	*1.04*	*1.05*
(Rob. Std. Error) Rob. T-test	(0.134) 7.81	(0.141) 7.78
Covariance		**0.217**
Scaled Estimate		*0.205*
(Rob. Std. Error) Rob. T-test		(0.0543) 4.00
Number of simulation draws	-	1000
Number of parameters	11	12
Final log-likelihood	-1164.850	-1161.472
Adjusted rho square	0.145	0.147
Sample size: 780, Null log-likelihood: -1375.851		
BIOGEME (Bierlaire, 2003, Bierlaire, 2005) has been used for all model estimations		

Table 3: Estimation results (continued)

Acknowledgments

The dataset was collected by the Swiss Railways as part of the research program on mobility pricing of the Swiss Federal Department of the Environment, Transport, Energy and Communications and the Swiss Federal Roads Authority. We would like to thank the collaborators of the Institute for Transport Planning and Systems (Swiss Federal Institute of Technology, Zurich) and of the Institute of Economics (University of Lugano), directed by Kay Axhausen and Rico Maggi respectively, for their valuable comments. We are grateful to Jelena Stojanovic who verified the coherence of the data. We have also benefited from discussions with Moshe Ben-Akiva.

References

Abdel-Aty, M. A., Kitamura, R., Jovanis, P. P., Reddy, P. and Vaughin, K. M. (1995). New approach to route choice data collection: Multiphase, computer-aided telephone inverview panel surveys using ge-

ographic information systems data base, *Transportation Research Record* **1493**: 159–169.

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

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

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

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

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

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

Ben-Akiva, M. and 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.

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

Bierlaire, M. (2005). An introduction to biogeme version 1.4. http://biogeme.epfl.ch.

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

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

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

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

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

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

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

Murakami, E. and Wagner, D. (1999). Can global positioning system (GPS) improve trip reporting?, *Transportation Research Part C* **7**(2-3): 149–165.

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

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

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

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

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

Vrtic, M., Schüssler, N., Erath, A., Axhausen, K., Frejinger, E., Bierlaire, M., Stojanovic, S., Rudel, R. and Maggi, R. (2006). Einbezug von reisekosten bei der modellierung des mobilitätsverhalten. Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Communications, IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.

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

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