Capturing Correlation in Large-Scale Route Choice Models*

E. Frejinger M. Bierlaire†

January 6, 2006

Report RO-060106
Operations Research Group ROSO
Institute of Mathematics
Ecole Polytechnique Fédérale de Laussane

Abstract

When using random utility models for a route choice problem, choice set generation and correlation among alternatives are two issues that make the modeling complex. In this paper we discuss different models capturing path overlap. First, we analyze several formulations of the Path Size Logit model proposed in the literature and show that the original formulation should be used. Second, we propose a modeling approach where the path overlap is captured with a subnetwork. A subnetwork is a simplification of the road network only containing easy identifiable and behaviorally relevant roads. In practice, the subnetwork can easily be defined based on the route network hierarchy. We propose a model where the subnetwork is used for defining the correlation structure of the choice model. The motivation is to explicitly capture the most important correlation without considerably increasing the model complexity.

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

†École Polytechnique Fédérale de Lausanne, Institute of Mathematics, CH-1015 Lausanne, Switzerland. E-mail: {emma.frejinger,michel.bierlaire}@epfl.ch

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

1 Introduction

The route choice problem concerns the choice of route between an origin-destination pair on a given transportation mode in a transportation network. The problem is critical in many contexts, for example in intelligent transport systems, GPS navigation and transportation planning. The efficiency of shortest path algorithms has been a strong motivation of many researchers to assume that travelers use the shortest (with regard to any arbitrary generalized cost) route among all. Clearly, the poor behavioral realism of the shortest path assumption motivates the use of more sophisticated models such as discrete choice models.

Designed to forecast how individuals behave in a choice context, discrete choice models (more specifically, random utility models) have motivated a tremendous amount of research in recent years (Ben-Akiva and Lerman, 1985). In the specific context of route choice, the definition of the choice set, and the significant correlation among alternatives are the two main difficulties (Ben-Akiva and Bierlaire, 2003).

In this paper we discuss correlation among alternatives in large choice sets. First, we present in Section 2 a literature review and then analyze the Path Size Logit model (Section 3). In Section 4 we introduce a new modeling approach based on the concept of subnetworks. Finally, we present estimation results for real data of Error Component models based on subnetworks and compare the results with a Path Size Logit model.

2 Literature Review

Several different models have been proposed in the literature. The Multinomial Logit (MNL) model, is simple but restricted by the Independence from Irrelevant Alternatives (IIA) property, which does not hold in the context of route choice due to overlapping paths. Efforts have been made to overcome this restriction by making a deterministic correction of the utility for overlapping paths. Cascetta et al. (1996) were the first to propose such a deterministic correction. They included an attribute, called Commonality Factor (CF), in the deterministic part of the utility obtaining a model called C-Logit. The utility U_{in} associated with path i by individual n is

$$U_{in} = V_{in} - \beta_{CF} CF_{in} + \varepsilon_{in}.$$

The CF_{in} value of a path i is directly proportional to the overlap with other paths in the choice set C_n. Cascetta et al. (1996) present three different formulations of the CF attribute. The first is

$$CF_{in} = \ln \sum_{j \in C_n} \left(\frac{L_{ij}}{\sqrt{L_i}\sqrt{L_j}} \right)^{\gamma}, \tag{1}$$

where L_{ij} is the length of links common to paths i and j, L_i and L_j are the lengths of paths i and j, and γ is a positive parameter (Cascetta et al., 1996 suggest the values 1 or 2). The other two formulations are

$$CF_{in} = \ln \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i} \sum_{j \in C_n} \delta_{aj} \right) \text{ and} \tag{2}$$

$$CF_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i} \ln \sum_{j \in C_n} \delta_{aj} \right), \tag{3}$$

where the fraction $\frac{l_a}{L_i}$ is the proportional weight of link a for path i, here represented by their lengths. Γ_i is the set of all links of path i and δ_{aj} equals 1 if link a is on path j and 0 otherwise. $\sum_{j \in C_n} \delta_{aj}$ is therefore the number of paths in choice set C_n sharing link a. Cascetta et al. (1996) do not provide any guidance for which CF formulation to use. They use

formulation (2) when estimating models for heavy truck path choice on the Italian national network.

Cascetta et al. (2002) present a route perception model. It is a two step model, where the probability that a path belongs to a choice set is modeled with a Binary Logit model, and the choice of path is modeled with a C-Logit model using formulation (1).

Ramming (2001) discusses a fourth CF formulation

$$CF_{in} = \ln\left(1 + \sum_{j \in C_n, j \neq i} \left(\frac{L_{ij}}{\sqrt{L_i L_j}}\right)\left(\frac{L_i - L_{ij}}{L_j - L_{ij}}\right)\right) \tag{4}$$

that is also analyzed by Hoogendoorn-Lanser et al. (2005).

The lack of theoretical guidance for the C-Logit model was the motivation for Ben-Akiva and Bierlaire (1999a) to propose the Path Size Logit (PSL) model. The idea is similar to the C-Logit model. A correction of the utility for overlapping paths is obtained by adding an attribute to the deterministic part of the utility. In this case, the Path Size (PS) attribute. The original PS formulation is derived from discrete choice theory for aggregate alternatives (see chapter 9, Ben-Akiva and Lerman, 1985). The utility is $U_{in} = V_{in} + \beta_{PS} \ln PS_{in} + \varepsilon_{in}$ where the PS attribute is defined as

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}}. \tag{5}$$

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

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_n} \frac{L^*_{C_n}}{L_j} \delta_{aj}}. \tag{6}$$

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

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_n} \left(\frac{L_i}{L_j}\right)^{\gamma} \delta_{aj}}, \tag{7}$$

where γ is a parameter greater or equal to zero. Note that when $\gamma = 0$ the formulation corresponds to the original PS formulation (5). Ramming (2001) proposed this formulation in order to decrease the impact of unrealistically long paths in the choice set. In the original PS formulation (5) the contribution of a link is decreased by the number of paths that share the link. If there are very long paths that no traveler is likely to choose sharing a link, then these long paths have a negative impact on the utility of shorter, more reasonable paths.

Ramming (2001) compares the C-Logit and PSL models with the different formulations but does not provide a theoretical comparison. Empirically he finds an inappropriate sign of the estimated parameter β_{CF} for CF formulations (2) and (3). He also finds that the PSL model with the Generalized PS formulation (7) outperforms the C-Logit model with formulations (1) and (4).

Note that the two CF formulations (2) and (3) are quite similar to one another and also to the original PS formulation (5). The difference lies in how the number of paths sharing a same link is taken into account. For the original PS formulation, a link's contribution to a path is reduced proportionally to the number of paths sharing the link. Whereas in CF formulation (2) the links contribution is multiplied with the number of paths sharing it, and in formulation (3) it is multiplied with the natural logarithm of the number of paths.

Hoogendoorn-Lanser et al. (2005) (see also Hoogendoorn-Lanser, 2005) study how to define overlap in multi-modal networks. Based on the conclusions of Ramming (2001), they do not further analyze the C-Logit models but focus on PSL models. They investigate if the β_{PS} parameter should be estimated or set to one, and conclude that it should be estimated since the PS attribute can capture behavioral perceptions regarding overlapping paths. Moreover, they compare different PS formulations in terms of model fit measures and finds that the generalized formulation (7) with $\gamma = 14$ shows best results. They also observe best model fit when overlap is expressed in terms of number of legs[1] compared to time and distance.

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

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

Vovsha and Bekhor (1998) propose the Link-Nested Logit model, which is a Cross-Nested Logit (CNL) formulation (see Bierlaire, forthcoming, for an analysis of the CNL model) where each link of the network corresponds to a nest, and each path to an alternative. Ramming (2001) estimated the Link-Nested Logit model on route choice data collected on the Boston network (34 thousand links). The large number of links makes it impossible to estimate the nest-specific coefficients. He concludes that the PSL model with the generalized formulation (7) outperforms the Link-Nested Logit model.

The Multinomial Probit model (Daganzo, 1977) has a flexible model structure that permits an arbitrary covariance structure specification. But numerical integration techniques must be used which limits the application of the model to large-scale route choice. Yai et al. (1997) propose a Multinomial Probit model with structured covariance matrix in the context of route choice in the Tokyo rail network. The maximum number of alternatives was however limited to four.

An Error Component (EC) model is a Normal mixture of MNL (MMNL) model and was described namely by Bolduc and Ben-Akiva (1991). The utility function for individual n and alternative i is

$$U_{in} = V_{in} + \xi_{in} + \nu_{in}$$

where V_{in} are the deterministic utilities, ξ_{in} are normally distributed and capture correlation between alternatives, and ν_{in} are independent and identically distributed Extreme Value.

The EC model can be combined with a factor analytic specification where some structure is explicitly specified in the model to decrease its complexity. Bekhor et al. (2002) estimate an EC model based on large-scale route choice data collected in Boston. The utility vector $\mathbf{U}_n$ (Jx1, where J is the number of paths) is defined by

$$\mathbf{U}_n = \mathbf{V}_n + \varepsilon_n = \mathbf{V}_n + \mathbf{F}_n\mathbf{T}\zeta_n + \nu_n, \tag{8}$$

where $\mathbf{V}_n$ (Jx1) is the vector of deterministic utilities, $\mathbf{F}_n$ (JxM) is the link-path incidence matrix (M is the number of links), $\mathbf{T}$ (MxM) is the link factors variance matrix, and ζ_n (Mx1) is the vector of i.i.d. normal variables with zero mean and unit variance. Bekhor et al. (2002) assume that link-specific factors are i.i.d. normal and that variance is proportional to link length so that $\mathbf{T} = \sigma \text{ diag}\left(\sqrt{l_1}, \sqrt{l_2}, \ldots, \sqrt{l_M}\right)$ where σ is the only parameter to be estimated. The covariance matrix can then be defined as follows:

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

where $L_{i,j}$ is length by which path i overlaps with path j.

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

The Paired Combinatorial Logit model, developed by Chu (1989), has been adapted to the route choice problem by Prashker and Bekhor (1998). Recently, the Link-Based Path-Multilevel Logit model has specifically been developed for the route choice problem by Marzano and Papola (2004). These models have been used for small-scale route choice analysis on test networks.

3 Deterministic Correction of Correlation

In this section, we discuss the Path Size Logit model in detail. We show that the original PS formulation (5) should be used for correcting utilities of overlapping paths. This is the formulation that both shows intuitive results and has a theoretical motivation. We start by deriving the original

PS formulation from the theory on aggregation of alternatives (Ben-Akiva and Lerman, 1985).

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

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

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

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

where μ is a positive scale parameter.

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

The contribution of a link a is then $\frac{1}{\mu}\ln\frac{1}{\sum_{j\in C_n}\delta_{aj}}$ where δ_{aj} is the link-path incidence variable. Furthermore, we assume that the size of a path is proportional to the length of its links. If l_a denotes the length of link a and L_i the length of path i, we have derived the original PS formulation

$$PS_{in} = \sum_{a\in\Gamma_i}\frac{l_a}{L_i}\frac{1}{\sum_{j\in C_n}\delta_{aj}}.$$

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

$$U_{in} = V_{in} + \beta_{PS}\ln PS_{in} + \varepsilon_{in}, \quad i \in C_n,$$

where $\beta_{PS} = \frac{1}{\mu}$.

Two questions regarding the original PS formulation that are discussed in the literature can be answered based on how the PS formulation is derived. First, whether β_{PS} should be fixed to one or estimated. Second, to which extent the PS attribute can capture correlation.

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

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

Ben-Akiva and Bierlaire (1999b) present an alternative PS formulation (6) including the length of the shortest path in the choice set $L^*_{C_n}$. The

correlation of the utility $\ln PS_{in}$ can be written as follows:

$$\ln PS_{in} = -\ln L_i - \ln L^*_{C_n} + \ln \sum_{a \in \Gamma_i} \frac{l_a}{\sum_{j \in C_n} \frac{1}{L_j} \delta_{aj}}.$$

Note that, including $L^*_{C_n}$ adds a constant $\ln L^*_{C_n}$ to all path utilities in the choice set which does not change their relative utility.

The Generalized PS formulation (7) was introduced by Ramming (2001) in order to decrease the influence of unrealistically long paths on the utility of shorter paths in the choice set. The formulation is however difficult to interpret for $\gamma > 0$. (Note that $\gamma = 0$ corresponds to the original PS formulation.)

In order to analyze the influence of the γ parameter, we write $\ln PS_{in}$ as follows:

$$\ln PS_{in} = -(\gamma + 1) \ln L_i + \ln \sum_{a \in \Gamma_i} l_a \frac{1}{\sum_{j \in C_n} \left(\frac{1}{L_j}\right)^{\gamma} \delta_{aj}}. \tag{9}$$

Independently of the value of the γ parameter, this formulation yields a zero correction when path i has no overlap with any other path in the choice set. However, it is theoretically difficult to give an interpretation as well as a motivation of the γ parameter, especially for large values. Indeed when $\gamma \to +\infty$, if we assume that $L_i > 1 \quad \forall\, i \in C_n$, the limits of the two terms in equation (9) are

$$\lim_{\gamma \to +\infty} -(\gamma + 1) \ln L_i = -\infty \qquad \lim_{\gamma \to +\infty} \ln \sum_{a \in \Gamma_i} l_a \frac{1}{\sum_{j \in C_n} \left(\frac{1}{L_j}\right)^{\gamma} \delta_{aj}} = +\infty.$$

This result can be explained by the fact that the sum in the denominator of formulation (7) is composed of terms $\left(\frac{L_i}{L_j}\right)^{\gamma}$ where $\frac{L_i}{L_j}$ can be greater or equal to one, or less than one depending on the lengths L_i and L_j. Since Ramming (2001) considered an example with only two correlated alternatives this effect was not illustrated in his thesis. Here we consider

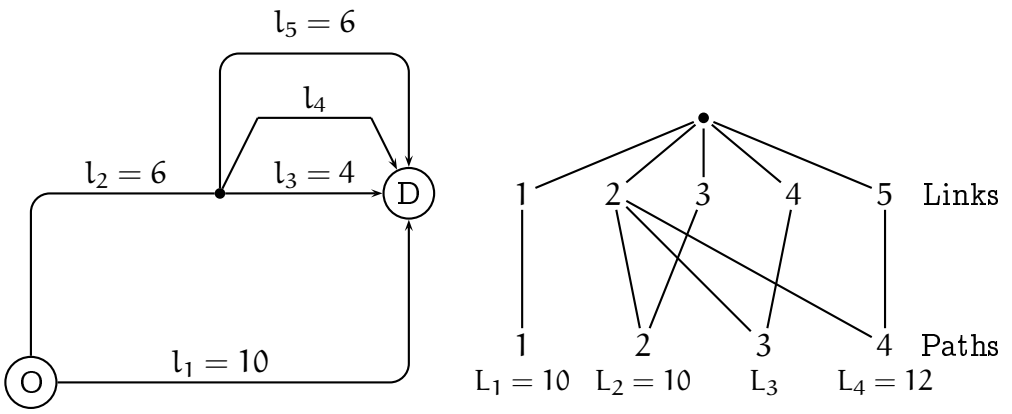

Figure 1: Example for Deterministic Correction Formulation

instead an example with three correlated alternatives (shown in Figure 1) where the length of path 3, L_3, varies with the length of link 4, l_4.

In Figure 2 we compare the values of the original PS formulation (5), $\gamma = 0$ (thin lines), with the generalized formulation (7) using a high value of γ (thick lines) as a function of l_4. Only the PS values for the correlated alternatives are shown.

The original PS formulation penalizes path 2 the most and path 4 the least, which is intuitive since the correlated part (link 2) has a higher proportion of the total length for path 2 than path 4. Moreover, path 3 is penalized proportionally to the length of link 4. For a high value of γ the results are counter intuitive since path 2 is not penalized at all, except when the length of path 3 is close to the length of path 2 (shortest path). In this case, the correction is highly unstable with respect to variations of l_4.

We now consider a choice set where two alternatives have almost the same length and one of those alternatives is the shortest path, that is $L_1 = 10.0, L_2 = 10.0, L_3 = 10.1$ and $L_4 = 12$. A case which is common in practice. In Figure 3 we show the PS values for this case as γ varies. First of all, note that the ordering of the paths changes. Path 4 is more

penalized than path 3 for $\gamma < 170$ and then the order is inverted. Second, even though path 3 is only 1% longer than path 2, its PS value decreases as γ increases.

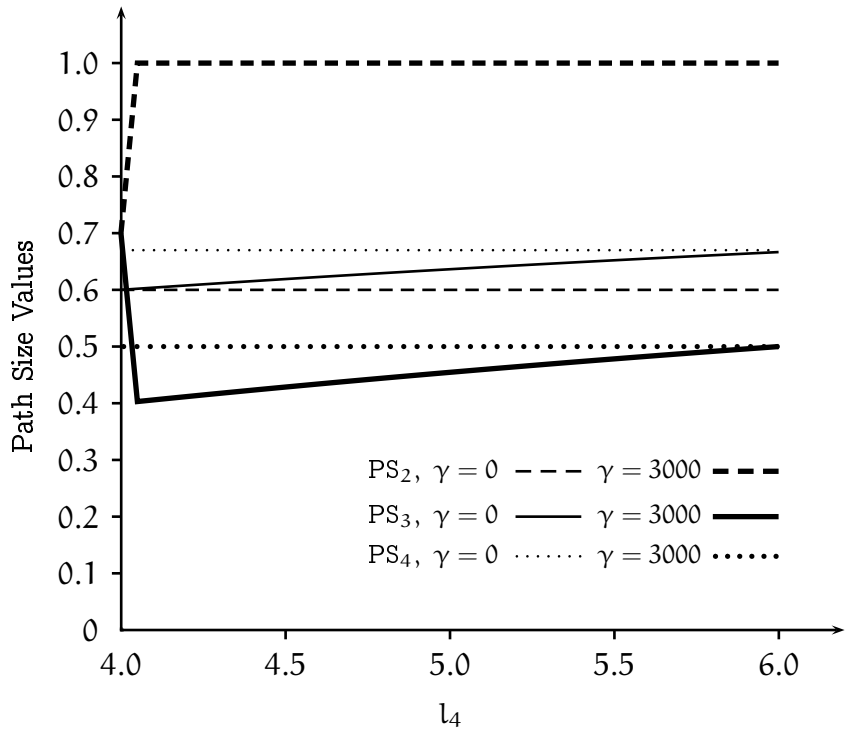

Figure 2: PS values ($\gamma = 0$ and $\gamma = 3000$) for correlated alternatives in example 1 as a function of l_4

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

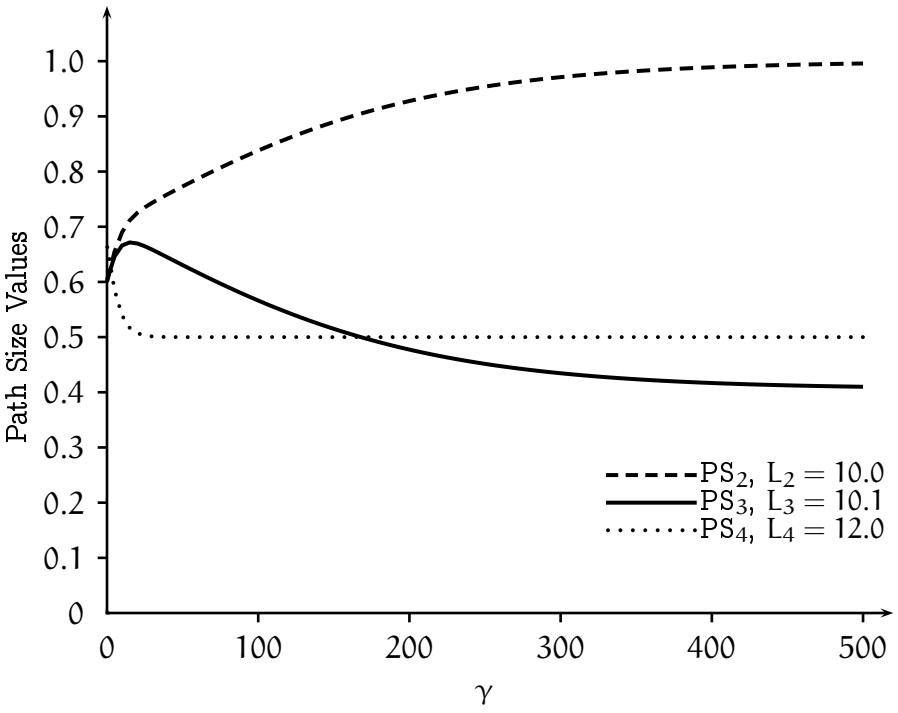

Figure 3: PS values as a function of γ for correlated alternatives

4 Subnetworks

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

We hypothesize that paths sharing a subnetwork component are correlated, even if they are not physically overlapping. We propose to explicitly capture this correlation within a factor analytic specification of a EC model. The model specification is combined with a PS attribute that accounts for the topological correlation on the complete network. The LK model specification builds on the model presented by Bekhor et al. (2002). We define the utility as

$$\mathbf{U}_n = \beta^T \mathbf{X}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n \tag{10}$$

where $\mathbf{F}_{n\ (JxQ)}$ is the factor loadings matrix (J is the number of paths and Q is the number of subnetwork components), $\mathbf{T}_{(QxQ)} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_Q)$ (σ_q is the covariance parameter associated with subnetwork component q, to be estimated), $\zeta_{n\ (Qx1)}$ is a vector of i.i.d. N(0,1) variates, and $\nu_{(Jx1)}$ is a vector of i.i.d. Extreme Value distributed variates. An element $(f_n)_{iq}$ of $\mathbf{F}_n$ equals $\sqrt{l_{niq}}$ where l_{niq} is the length by which path i in choice set C_n overlaps with subnetwork component q.

We illustrate the model specification with a small example presented in Figure 4. We consider one origin-destination pair, three paths and a subnetwork composed of two subnetwork components (S_a and S_b). Path 1 uses both subnetwork components whereas path 2 only uses S_a and path 3 only S_b. Path 1 is assumed to be correlated with both path 2 and path 3 even though path 1 and path 2 do not physically overlap. The path utilities

for this example are consequently

$$
\begin{aligned}
U_1 &= \beta^T X_1 + \sqrt{l_{1a}}\sigma_a\zeta_a + \sqrt{l_{1b}}\sigma_b\zeta_b + \nu_1 \\
U_2 &= \beta^T X_2 + \sqrt{l_{2a}}\sigma_a\zeta_a + \nu_2 \\
U_3 &= \beta^T X_3 + \sqrt{l_{3b}}\sigma_b\zeta_b + \nu_3,
\end{aligned}
$$

where ζ_a and ζ_b are distributed N(0,1), l_{iq} is the length path i uses subnetwork component q. σ_a and σ_b are the covariance parameters to be estimated.

The variance-covariance matrix of ζ for this example is

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

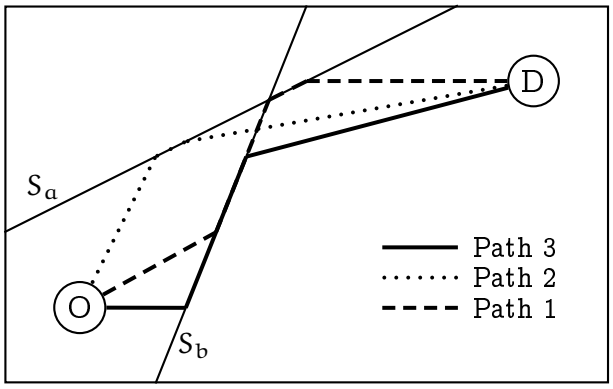

Figure 4: Example of a Subnetwork

4.1 Empirical Results

The estimation results presented in this section are based on a GPS data set collected during a traffic safety study in the Swedish city of Borlänge. Nearly 200 vehicles were equipped with a GPS device and the vehicles were monitored within a radius of about 25 km around the city center.

Since the data set was not originally collected for route choice analysis, an extensive amount of data processing has been performed in order to clean the data and obtain coherent routes. The data processing for obtaining data for route choice analysis was mainly performed by the company GeoStats in Atlanta. Data of 24 vehicles and a total of 16 035 observations are available for route choice analysis. (See Axhausen et al., 2003, Schönfelder and Samaga, 2003 and Schönfelder et al., 2002 for more details on the Borlänge GPS data set.) For the model estimations we consider a total of 2 978 observations corresponding to 2 244 observed simple routes of 24 vehicles and 2 179 origin-destination pairs. Note that we make a distinction between observations and observed routes since a same route can have been observed several times.

Borlänge is situated in the middle of Sweden and has about 47 000 inhabitants. The road network contains 3 077 nodes and 7 459 unidirectional links. We have defined a subnetwork based on the main roads for traversing the city center. Two of the Swedish national roads ("riksväg") traverse Borlänge. The subnetwork is composed of these national roads (referred to as R.50 and R.70) and we have defined two subnetwork components for each national road (north and south directions). In addition, we have defined one subnetwork component for the road segment in the city center where R.50 and R.70 overlap (called R.C.). The Borlänge route network and the subnetwork are shown in Figure 5. In Table 1 we report for each subnetwork component its length and the number of observations that use the component. Table 1 also reports the weighted number N_q, defined by $N_q = \sum_{o \in O} \frac{l_{oq}}{L_q}$, where l_{oq} is the common length between the route corresponding to observation o and subnetwork component q, L_q is the length of q, and O is the set of all observations.

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

The main drawback of the link elimination approach is that it generates similar routes. When one link is removed, there exists often a short

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

Table 1: Statistics on Observations of Subnetwork Components

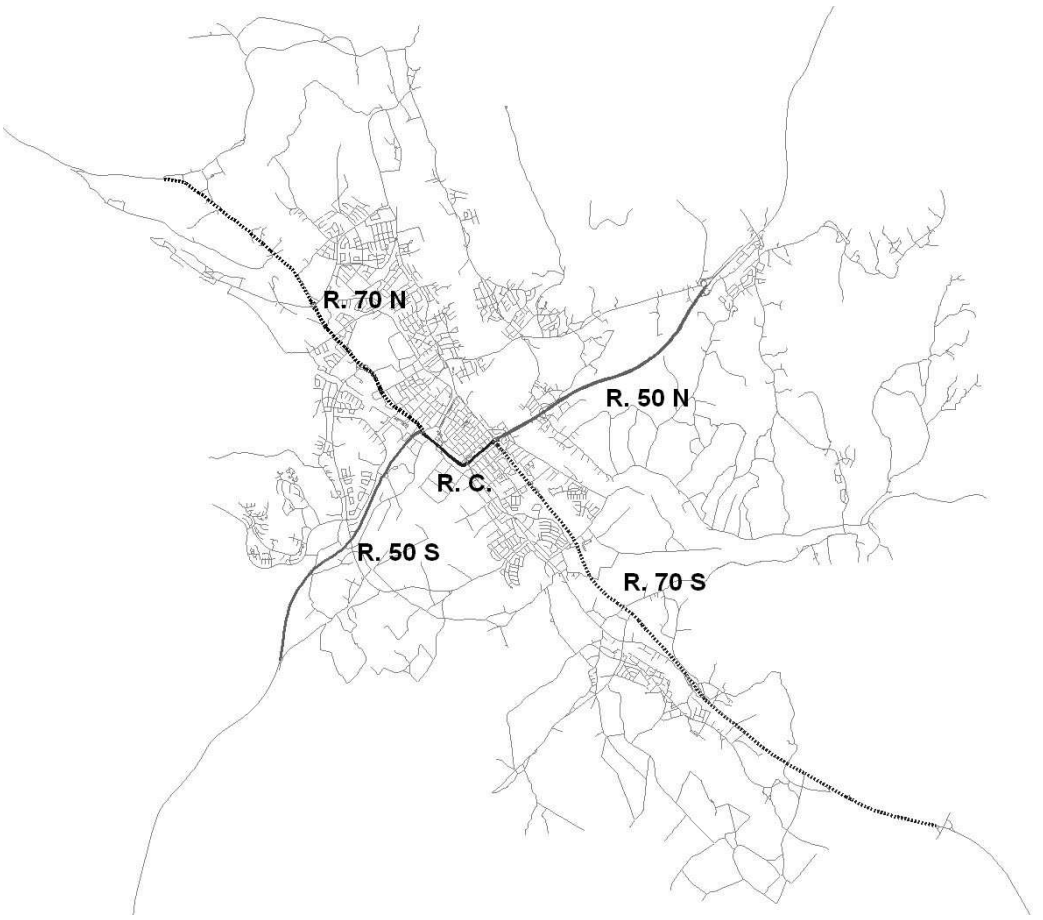

Figure 5: Overview of Borlänge Road Network and Subnetwork Definition

deviation using roads next to the removed link. In order to address this drawback we have used two generalized costs for the shortest path computation. In addition to estimated travel time, we have also used link length divided by the number of lanes. For each origin-destination pair, the link elimination algorithm is therefore applied to two shortest paths.

The observed routes that were not found by the choice set generation algorithm were added afterwards. The algorithm found all the observed routes for 80% of the origin-destinations pairs. However, for 20% of the origin-destination pairs, none of the observed routes were identified, which corresponds to 23% of the observed routes. Typically, this is the case when the observed routes make long detours compared to the shortest path, for example, in order to avoid the city center. These results are consistent with the findings of Ramming (2001) who at best found 84% of the observed routes by combining all the choice set generation algorithms that he had tested. The number of paths in the choice sets varies between 2 and 43 where a majority of the choice sets (93%) include less than 15 paths.

4.1.1 Model Specification

We compare a PSL model with three different specifications of a EC model based on the subnetwork defined previously. One EC model (EC_1) is specified with a simplified correlation structure where the covariance parameters are assumed to be equal. The second and third EC models (EC_2 and EC_3) are specified with one covariance parameter per subnetwork component.

Even though the number of individuals is small, we provide a model (EC_3) where we take into account that we have panel data. We assume that the perception of correlated alternatives on the subnetwork is individual specific and that the taste is constant over choice situations. The random parameters in the correlation structure are therefore specified to be invariant across the observations of a given individual.

All models are specified with the same linear in parameters formulation of the deterministic part of the utility function. The deterministic part V_i

for alternative i is

$$\begin{aligned} V_i = \; & \beta_{PS} \ln(PS_i) + \beta_{EstimatedTime} EstimatedTime_i + \\ & \beta_{NbSpeedBumps} NbSpeedBumps_i + \beta_{NbLeftTurns} NbLeftTurns_i + \\ & \beta_{AvgLinkLength} AvgLinkLength_i. \end{aligned}$$

In addition to classical attributes such as estimated travel time, number of speed bumps and number of left turns in uncontrolled crossings, we have included average link length which is intended to capture an attraction for routes with few crossings. The estimated travel time is computed for each link in the network based on its length and an average speed. We have used one average speed for each speed limit that corresponds to the observed average speed. Statistics on all attributes included in the model specifications are given in Table 2.

A PS attribute, defined by the original formulation (5) based on length, is included in all models in order to capture the topological correlation among alternatives. PS based on length and estimated travel time shows similar results, length was therefore preferred since it is known with certainty. A high correlation among the routes is expected since a link elimination approach has been used for generating the choice sets. In Figure 6 we show the PS values for all routes and all choice sets. The generated routes are shown with black bars and the observed routes with gray bars. A majority of the routes have a high overlap (low PS values). Only 5% of the routes have no overlap (PS value that equals 1). Note however that almost 50% of the routes that have no overlap are observed routes. This can be explained by the poor performance of the choice set generation algorithm discussed in the previous section. Namely, for 20% of the origin-destination pairs, none of the observed routes were found by the algorithm. These observed routes are therefore expected to have a low overlap with the other routes in the choice set.

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

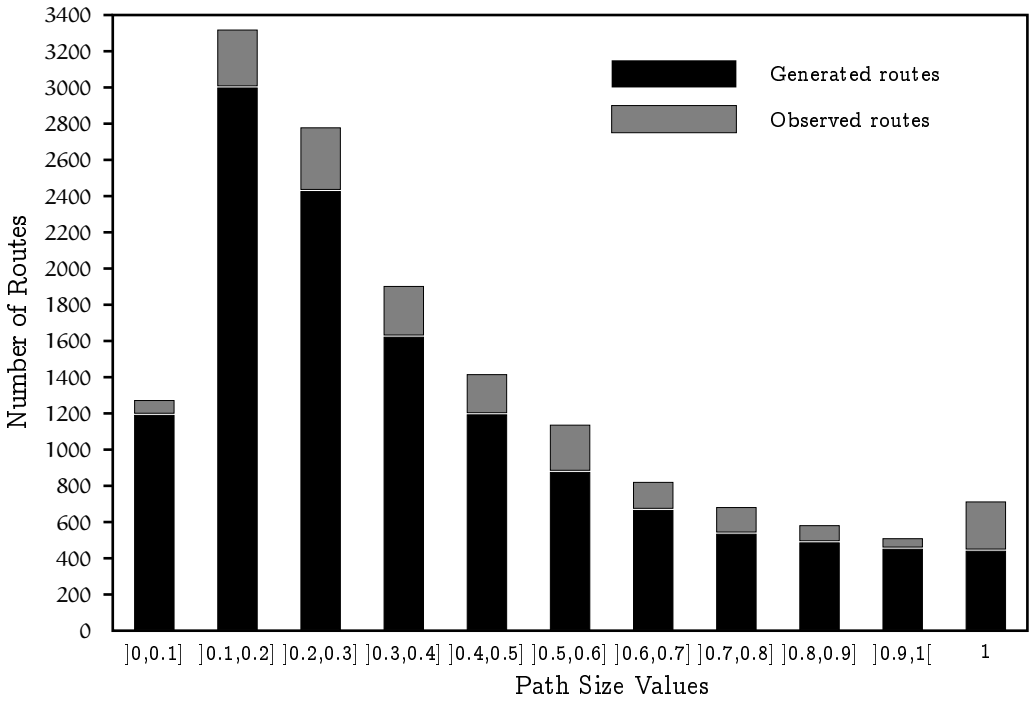

Figure 6: Number of routes for PS values

is fixed to one.

4.1.2 Model Estimation

The parameter estimates are given in Table 3. We have provided a scaled parameter estimate in order to facilitate the comparison of different models. The scaling is based on the estimated travel time parameter in the PSL model. The magnitude of the scaled estimate for this parameter is consequently the same for all the models.

We start by comparing the models PSL, EC_1 and EC_2. The parameter estimates shown in Table 3 related to average link length, estimated travel time, number of left turns and number of speed bumps are all significantly different from zero. Moreover, the parameter values as well as the robust t-test statistics are very stable when comparing the different models.

The PS parameter estimate, β_{PS}, is negative and significantly different from zero and from minus one in models PSL, EC_1 and EC_2. As discussed in

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

Table 2: Statistics on Attributes

Section 3 a negative value of β_{PS} is not consistent with choice theory since it corresponds to a scale parameter and consequently should be positive. The negative estimate suggests that the PS attribute captures an attractiveness for overlapping paths. An increase in magnitude and significance of the scaled β_{PS} estimates can be noted when comparing EC_1 with PSL and EC_2 with EC_1. More precisely, when the correlation structure on the subnetwork is explicitly captured by the error terms, the value of β_{PS} increases in magnitude and significance. Based on these results, we draw the conclusion that the PS attribute as an ambiguous interpretation. On the one hand, it negatively corrects the utility for the independence assumption on the random terms. On the other hand, it has a behavioral interpretation. Namely, it captures an attractiveness for overlapping paths, for example, because they provide de possibility of route switching (this has also been suggested by Hoogendoorn-Lanser et al., 2005 in the context of multi-modal route choice). Another possible explanation for the negative β_{PS} estimate is based on the choice set definition. A majority of the observed paths have a high overlap with other paths in the choice set (see Figure 6). Hence, the utility is increased for overlapping paths.

Based on the log-likelihood values reported in Table 4, and the χ^2-tests shown in Table 5, the PSL model can be rejected when compared with EC_1 and EC_2. Moreover, EC_2 is significantly better than EC_1. The hypothesis of equal covariance parameters for all subnetwork components can therefore be rejected although not as strongly as the PSL model.

The estimate of σ_{R50S} in model EC_2 (see Table 3) is not significantly different from zero. This can be explained by the limited number of obser-

vations using this subnetwork component. As shown in Table 1, there are 173 observations that use R.50 S but since the number of weighted observations is only 36, the length by which they overlap with the subnetwork component is relatively short.

Considering the significant improvement in model fit for the EC_1 and EC_2 models compared to the PSL model, as well as the significant covariance parameter estimates, we conclude that the specification based on subnetwork captures an important correlation structure.

Finally, we compare EC_2 with EC_3 where EC_3 explores the panel data structure of the observations. Referring to the scaled parameter estimates in Table 3 for average link length, estimated travel time, number of left turns and number of speed bumps, the value of the estimates are very stable. On the contrary, the value β_{PS} decreases in magnitude, breaking a trend where it has been increasing in magnitude for the models EC_1 and EC_2 compared to the PSL model. It is possible that the EC_3 model better captures individuals' perception of overlapping paths than EC_1 and EC_2. The behavioral aspect that the PS attribute captures in models EC_1 and EC_2 is therefore captured within the model structure of EC_3. This would explain the decreased magnitude of the β_{PS} value.

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

Due to the small number of individuals there is a systematic loss in significance for all parameters in EC_3 compared to EC_2. In spite of this, there is a remarkable increase in model fit (see Table 4) compared to EC_2.

5 Conclusion

In this paper we justify the use of the original PS formulation among the deterministic corrections of the IIA assumption on the random terms in a MNL model. This is the formulation that both has a theoretical support and shows intuitive results for the correction of the independence assumption on the random terms. Moreover, we have presented estimation results

Parameters	**PSL**	**EC_1**	**EC_2**	**EC_3**
Path Size	**-0.28**	**-0.49**	**-0.53**	**-0.32**
Scaled estimate	*-0.28*	*-0.45*	*-0.48*	*-0.31*
(Std. Err.) Rob. t-test	(0.07) -4.05	(0.09) -5.61	(0.09) -5.91	(0.19) -1.65
Avg. Link Length	**4.15**	**4.98**	**5.06**	**4.75**
	4.15	*4.58*	*4.61*	*4.53*
	(0.55) 7.58	(0.60) 8.32	(0.61) 8.28	(1.21) 3.92
Estimated Time	**-0.40**	**-0.43**	**-0.44**	**-0.42**
	-0.40	*-0.40*	*-0.40*	*-0.40*
	(0.05) -7.85	(0.06) -7.47	(0.06) -7.51	(0.10) -4.37
Nb. Left Turns	**-0.32**	**-0.33**	**-0.33**	**-0.33**
	-0.32	*-0.30*	*-0.30*	*-0.31*
	(0.02) -15.73	(0.02) -15.62	(0.02) -15.59	(0.04) -9.16
Nb. Speed Bumps	**-0.23**	**-0.22**	**-0.23**	**-0.22**
	-0.23	*-0.20*	*-0.21*	*-0.21*
	(0.07) -3.52	(0.07) -3.14	(0.07) -3.14	(0.19) -1.11
σ		**1.44**		
		1.32		
		(0.19) 7.57		
$\sigma_{\mathbf{R50N}}$			**1.07**	**1.78**
			0.97	*1.70*
			(0.32) 3.28	(0.67) 2.66
$\sigma_{\mathbf{R50S}}$			**-0.27**	**-0.69**
			-0.24	*-0.66*
			(0.69) -0.39	(0.60) -1.16
$\sigma_{\mathbf{R70N}}$			**-2.04**	**0.65**
			-1.85	*0.62*
			(0.39) -5.16	(0.26) 2.55
$\sigma_{\mathbf{R70S}}$			**-1.52**	**-0.83**
			-1.39	*-0.79*
			(0.22) -7.08	(0.20) -4.07
$\sigma_{\mathbf{RC}}$			**2.02**	**1.19**
			1.83	*1.14*
			(0.66) 3.05	(0.32) 3.75

Table 3: Estimation Results

Model	Nb. σ Estimates	Nb. Estimated Parameters	Final L-L	Adjusted Rho-Square
PSL	-	13	-4174.72	0.154
EC_1	1	14	-4142.40	0.161
EC_2	5	18	-4136.92	0.161
EC_3	5	18	-4109.73	0.166
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 2978 observations Null Log-Likelihood: -4951.11 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003, Bierlaire, 2005).				

Table 4: Model Fit Measures

Model 1	Model 2	Test	Threshold (95%)
PSL	EC_1	64.64	3.84
PSL	EC_2	75.60	11.07
EC_1	EC_2	10.96	9.49

Table 5: χ^2-test

that suggest a behavioral interpretation of the Path Size attribute. Namely, overlap can be attractive for travelers since it provides the possibility of switching between different routes.

We propose a novel modeling approach based on subnetworks designed to enhance the performance of simple models, such as the Path Size Logit model. Estimation results show that this approach is significantly better than a simple Path Size Logit model. A subnetwork is a set of subnetwork components. Alternatives are assumed to be correlated if they use the same subnetwork component even if they do not physically overlap. This correlation is captured within a factor analytic specification of an Error Component model combined with a Path Size attribute. The estimation results are promising and the estimates of the covariance parameters suggest that the specification captures an important correlation structure.

We believe that this approach will open new perspectives for large-scale route choice modeling. It is a flexible approach where the trade-off between complexity and behavioral realism can be controlled by the analyst with the definition of the subnetwork. Clearly, more analysis is required to assess

the sensitivity of the results with regard to the definition of the subnetwork. Moreover, additional validity tests on other datasets would be desirable.

References

Axhausen, K., Schönfelder, S., Wolf, J., Oliveira, M. and Samaga, U. (2003). 80 weeks of GPS traces: Approaches to enriching the trip information, *Arbeitsbericht Verkehrs- und Raumplanung 178*, Institut für Verkehrsplanung and Transportsysteme, ETH Zürich.

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

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

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

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

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice, *in* R. Hall (ed.), *Handbook of Transportation Science, Second Edition*, Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 7–37.

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

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

Bierlaire, M. (2005). An introduction to BIOGEME version 1.4. roso.epfl.ch/mbi/biogeme/doc/tutorial.pdf.

Bierlaire, M. (forthcoming). A theoretical analysis of the cross-nested logit model, *Annals of operations research* . Accepted for publication.

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

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 the Theory of Road Traffic Flow (Lyon, France)*.

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

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

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

Han, B. (2001). *Analyzing car ownership and route choices using discrete choice models*, PhD thesis, Department of infrastructure and planning, KTH, Stockholm.

Han, B., Algers, S. and Engelson, L. (2001). Accommodating drivers' taste variation and repeated choice correlation in route choice modeling by

using the mixed logit model, *80th Annual Meeting of the Transportation Research Board*.

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

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

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

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

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

Prashker, J. and Bekhor, S. (1998). Investigation of stochastic network loading procedures, *77th Annual Meeting of the Transportation Research Board* .

Ramming, M. S. (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.

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

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

Yai, T., Iwakura, S. and Morichi, S. (1997). Multinomial probit with structured covariance for route choice behavior, *Transportation Research Part B* **31**(3): 195–208.