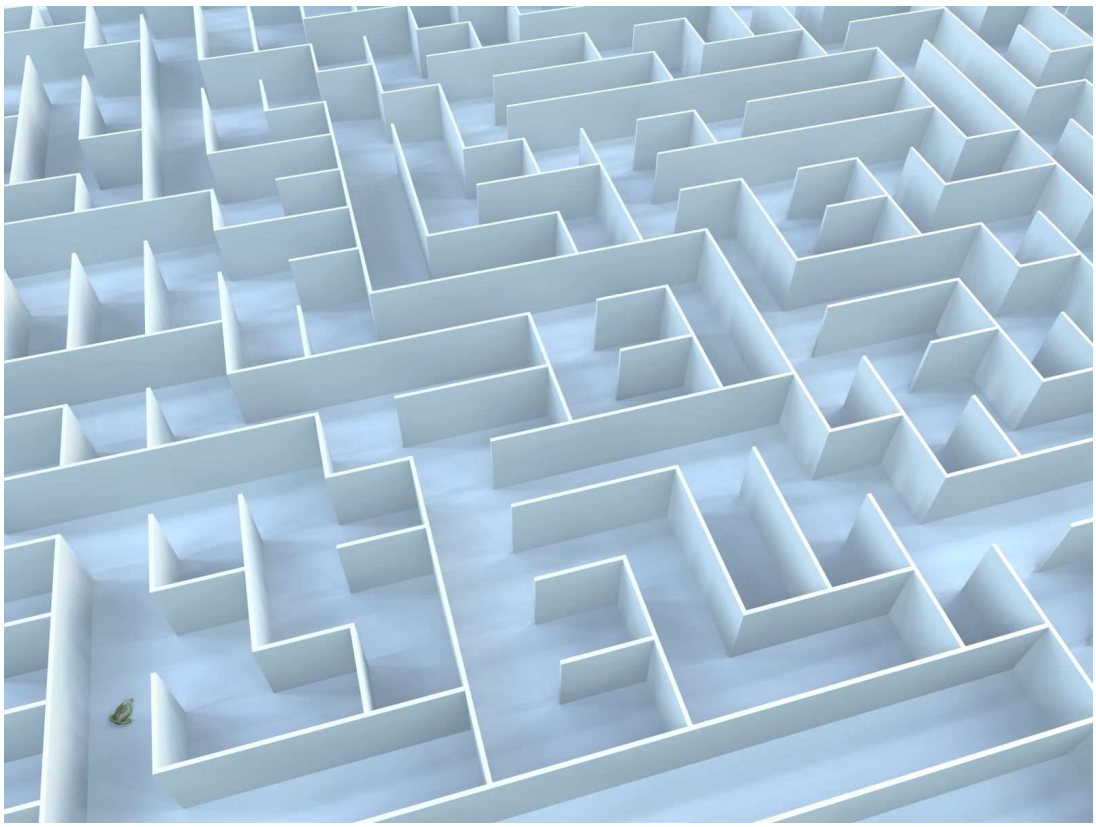

Capturing Correlation in Route Choice Models using Subnetworks

Emma Frejinger, EPFL
Michel Bierlaire, EPFL

Conference paper STRC 2006

STRC | 6th Swiss Transport Research Conference
Monte Verità / Ascona, March 15-17, 2006

Capturing Correlation in Route Choice Models using Subnetworks

Emma Frejinger
Institute of Mathematics, EPFL
Lausanne

Phone: +41 21 693 81 00
Fax: +41 21 693 55 70
Email: emma.frejinger@epfl.ch

Michel Bierlaire
Institute of Mathematics, EPFL
Lausanne

Phone: +41 21 693 25 37
Fax: +41 21 693 55 70
Email: michel.bierlaire@epfl.ch

February 2006

Abstract

When using random utility models for a route choice problem, choice set generation and correlation among alternatives are two issues that make the modelling complex. In this paper we propose a modelling approach where the path overlap is captured with a subnetwork. A subnetwork is a simplification of the road network only containing easy identifiable and behaviourally relevant roads. In practise, 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.

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 and Multinomial Logit models. Moreover, the correlation parameters are significant. We also analyse the performance of the different models regarding prediction of choice probabilities. The results show a better performance of the Error Component model compared to the Path Size Logit and Multinomial Logit models.

Keywords

Large Scale Route Choice Modelling – Subnetworks – Error Components – Path Size Logit – Forecasting – STRC 2006

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 travellers use the shortest (with regard to any arbitrary generalised cost) route among all. Clearly, the poor behavioural 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).

This paper is a continuation of the work presented in Frejinger and Bierlaire (2005) and Frejinger and Bierlaire (2006) where we discuss correlation among alternatives in large choice sets. Here, we especially focus on prediction capacities of different route choice models. First, we present a literature review in Section 2. A new modelling approach based on the concept of subnetworks is introduced in Section 3. Finally, we present estimation and prediction results for real data of Error Component models based on subnetworks and compare the results with Path Size Logit and Multinomial Logit models.

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

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

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. They do however not provide any guidance for which CF formulation to use.

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 modelled with a Binary Logit model, and the choice of path is modelled with a C-Logit model.

The lack of theoretical guidance for the C-Logit model was the motivation for Ben-Akiva and Bierlaire (1999) 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 $U_{in} = V_{in} + \beta_{\text{PS}} \ln \text{PS}_{in} + \varepsilon_{in}$ where the PS attribute is defined as

$$\text{PS}_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}}, \tag{1}$$

l_a is the length of link a, Γ_i the set of all links in path i, L_i the length of i and δ_{aj} equals one if path i use link a, and zero otherwise. Other PS formulations are presented in the literature but Frejinger and Bierlaire (2006) show that only the original formulation should be used. 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.

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, to appear, 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 generalised formulation (Ramming, 2001) 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

$$\mathbf{U}_n = \mathbf{V}_n + \varepsilon_n = \mathbf{V}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n, \tag{2}$$

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:

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

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 harbour 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 Subnetworks

We are proposing a modelling approach which is designed to be both behaviourally 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 labelled, and is behavioural 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 hypothesise 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

$$\mathbf{U}_n = \beta^T \mathbf{X}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n \tag{3}$$

where $\mathbf{F}_{n\ (JxQ)}$ is the factor loadings matrix (J is the number of paths and Q is the number of subnetwork components), $\mathbf{T}_{(QxQ)} = \text{diag}\,(\sigma_1, \sigma_2, \dots, \sigma_Q)$ (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 1. 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

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

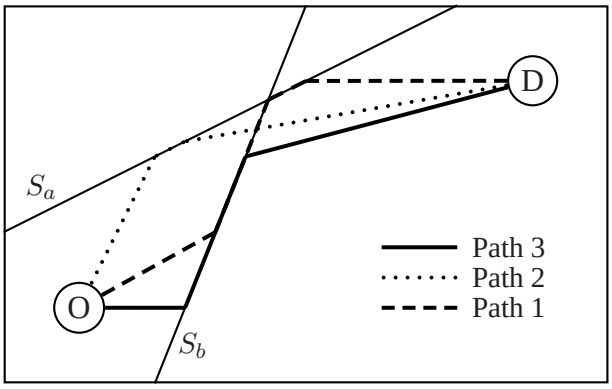

Figure 1: Example of a Subnetwork

4 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 centre. 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 $1\,693$ observations corresponding to $1\,408$ observed simple routes of 24 vehicles and $1\,353$ origin-destination pairs. Note that we

	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	145	157	248	317	226
Weighted Nb. of Observations (N_q)	32	91	68	70	132

Table 1: Statistics on Observations of Subnetwork Components

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 centre. 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 centre where R.50 and R.70 overlap (called R.C.). The Borlänge route network and the subnetwork are shown in Figure 2. 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 of observations 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) minimising free flow travel time. 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 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 centre. 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 (90%) include less than 15 paths.

4.1 Model Specification

We compare MNL and PSL models with two different specifications of an 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 EC model (EC_2) is specified with one covariance parameter per subnetwork component.

All models are specified with the same linear in parameters formulation of the deterministic

Figure 2: Overview of Borlänge Road Network and Subnetwork Definition

part of the utility function. The deterministic part V_i for alternative i is

$$V_i = \beta_{\text{PS}} \ln(\text{PS}_i) + \beta_{\text{EstimatedTime}} \text{EstimatedTimeGr10min}_i + \beta_{\text{Crossing}} \text{CrossingDummy}_i + \beta_{\text{NbLeftTurns}} \text{NbLeftTurns}_i.$$

and is described in detail below.

A PS attribute, defined by the original formulation (1) (Ben-Akiva and Bierlaire, 1999) based on length, is included in all models in order to capture the topological correlation among alternatives. The use of this formulation is motivated by the results presented by Frejinger and Bierlaire (2006). PS based on length and estimated travel time show similar results, length was therefore preferred since it is known with certainty. Figure 3 shows density estimates for the PS values of observed (solid line) and of non observed (dashed line) routes. A large proportion of the routes have a high overlap (low PS values). This is expected since a link elimination algorithm has been used for the choice set generation. Note that the curve for the observed routes is on the right side of the curve for non observed routes. Meaning that the observed routes have less overlap with other routes than non 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.

It is important to note that the PS attribute can be highly correlated with link additive attributes such as length and free-flow time. Indeed, the logarithm of the original PS formulation (1) can be written as follows

$$\ln \text{PS}_{in} = -\ln L_i + \ln \sum_{a \in \Gamma_i} \frac{l_a}{\sum_{j \in C_n} \frac{1}{L_j} \delta_{aj}}.$$

Here, special attention has been given to the specification of the estimated travel time attribute in the deterministic utilities. It is reasonable to assume that for short trips, other attributes than travel time play an important role in the route choice decision process. We tested a piecewise linear specification of the estimated travel time that confirmed this hypothesis since the coefficients associated with low travel times were estimated not significantly different from zero. We therefore include the estimated travel time attribute in the deterministic utilities if it is greater or equal to ten minutes. 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. It is difficult to obtain an accurate estimation of the travel time. In order to capture a preference for main crossings that could explain possible detours compared to shorter alternatives, a main crossing dummy is included in the deterministic utilities. (Actually, there are four main crossings in the centre of Borlänge, but only one crossing dummy was estimated significantly different from zero.)

The number of left turns is also included in the deterministic utilities. Since left turns can be considered more dangerous, and in general take more time than right turns, we expect this attribute to have a negative impact on the utility. Statistics on the attributes included in the model specifications are given in Table 2.

We deal with heteroscedasticity by specifying different scale parameters for different individuals. After systematic testing of various specifications, four individuals have one scale parameter

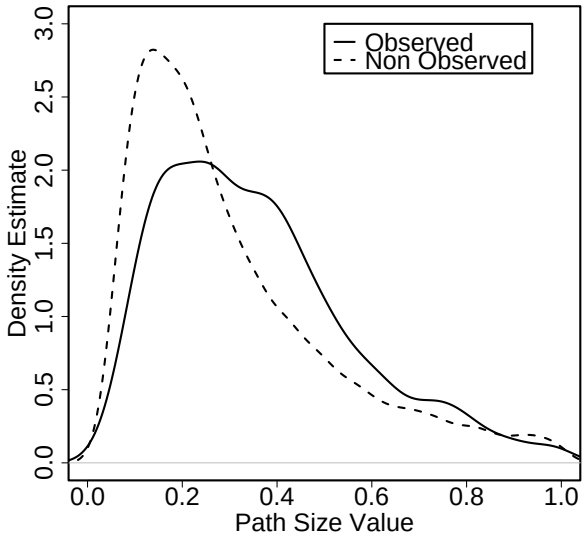

Figure 3: Density Estimate of PS values

Attribute	Min	Average	Max
Estimated Travel Time [min]	0.4	6.5	37.5
Number of Left Turns	0	5.0	27
Main Crossing Dummy	0	0.2	1
$\ln(\text{PS})$	-3.7	-1.2	0.0

Table 2: Statistics on Attributes

each which are estimated significantly different from one. For the remaining individuals the scale parameter is fixed to one.

4.2 Model Estimation

The parameter estimates are given in Table 3. We provide 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 MNL model. The scaled estimate for this parameter is consequently the same for all the models.

The parameter estimates related to estimated travel time, left turns and crossing dummy are all significantly different from zero. Moreover, the parameter values as well as the robust t-test statistics are very stable across models, which is very good news.

The PS parameter (β_{PS}) estimate is positive and significantly different from zero for the PSL model. This is consistent with theory since β_{PS} corresponds to a scale parameter (see Frejinger and Bierlaire, 2006). For the two EC models, the β_{PS} estimate is not significantly different from zero. Hence, when the correlation structure on the subnetwork is explicitly captured by the error terms, the topological correlation captured by the PS attribute is not significant.

These results are different from those presented in Frejinger and Bierlaire (2006) where the

PS parameter was estimated negative and significant. A systematic data cleaning has been performed excluding observations corresponding to short distances (trip duration less than 2 minutes). Indeed, they do not reflect the result of an actual choice process.

From the log-likelihood values reported in Table 4, we observe a large increase in model fit for the EC models compared to the PSL and MNL models. Moreover, the PSL model is significantly better than the MNL model (likelihood ratio test statistic of 6.24 compared to a threshold at 95% of 3.84) and the EC_2 is significantly better than EC_1 (likelihood ratio test statistic of 12.72 compared to a threshold at 95% of 9.49). The hypothesis of equal covariance parameters for all subnetwork components can therefore be rejected. The EC_2 model can however be simplified since the estimate of σ_{R50S} is not significantly different from zero. This can be explained by the limited number of observations using this subnetwork component. As shown in Table 1, there are 145 observations that use R.50 S but since the number of weighted observations is only 32, the length by which they overlap with the subnetwork component is relatively short. The EC_2 can be further simplified since the hypotheses that $\sigma_{R50N}=\sigma_{RC}$, $\sigma_{R50N}=\sigma_{R70N}$ and $\sigma_{R70N}=\sigma_{RC}$ cannot be rejected. For the estimation on this data the EC_2 model could therefore be specified with two different covariance parameters.

Considering the important improvement in model fit for the EC models compared to the PSL and MNL models, as well as the significant covariance parameter estimates, we conclude that the specification based on subnetwork captures an important correlation structure. In the next section we continue the comparison of these four models, but regarding their forecasting capacities.

4.3 Forecasting Results

Route choice models are often used to predict individual behaviour. It is therefore important to compare models, not only in terms of model fit, but also regarding the performance of predicting choice probabilities. For this purpose, the correct modelling of correlation among alternatives is crucial.

In order to test the different models prediction power, we use only a part of the observations to estimate the models, and apply them to the other part of the observations. The models are estimated on observations corresponding to 80 % of the origin destination pairs, and they are applied on observations of the remaining origin destination pairs. The origin destination pairs are randomly chosen. We have selected five different subsets of data. This test is particularly challenging since the models predict choice probabilities for origin destination pairs whose choice sets have not been used for estimating the models. Information about the five different data sets are given in Table 5. Since, in general, there is only one observation per origin destination pair, all the data sets have more or less the same size.

The MNL, PSL and EC_1 models are estimated with the same utility specifications as in the previous section. The EC_2 model is specified with four σ parameters where σ_{R50S} is not included because it has not been estimated significantly different from zero for any of the data sets. The estimation results are reported in the Appendix (Table 6 to Table 10). With few exceptions, the same interpretation of the estimation results as in the previous section can be made. Namely, the parameter values are stable across models as well as the t-test statistics. Moreover, the PS attribute is significant in the PSL models but not in the EC models. Except for the covariance parameters, a systematic loss in significance can be observed for all parameters compared to

Parameters	**MNL**	**PSL**	**EC$_1$**	**EC$_2$**
Path Size		**0.16**	**0.01**	**-0.01**
Scaled Estimate		*0.14*	*0.01*	*-0.02*
Rob. Std.		0.07	0.07	0.07
Rob. t-test		2.27	0.12	-0.20
Crossing Dummy	**0.55**	**0.53**	**0.60**	**0.62**
Scaled Estimate	*0.55*	*0.48*	*0.67*	*0.68*
Rob. Std.	0.22	0.22	0.24	0.23
Rob. t-test	2.44	2.41	2.50	2.67
Estimated Time $\geq$ 10 min	**-0.07**	**-0.08**	**-0.06**	**-0.06**
Scaled Estimate	*-0.07*	*-0.07*	*-0.07*	*-0.07*
Rob. Std.	0.03	0.03	0.03	0.03
Rob. t-test	-2.50	-2.80	-2.42	-2.50
Left turns	**-0.40**	**-0.40**	**-0.41**	**-0.42**
Scaled Estimate	*-0.40*	*-0.36*	*-0.47*	*-0.45*
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-25.91	-25.97	-25.30	-25.32
σ			**0.04**	
Scaled Estimate			*0.04*	
Rob. Std.			0.01	
Rob. t-test			7.39	
$\sigma_{\mathbf{R50N}}$				**0.05**
Scaled Estimate				*0.05*
Rob. Std.				0.01
Rob. t-test				3.57
$\sigma_{\mathbf{R50S}}$				**0.00**
Scaled Estimate				*0.00*
Rob. Std.				0.00
Rob. t-test				-0.20
$\sigma_{\mathbf{R70N}}$				**0.05**
Scaled Estimate				*0.05*
Rob. Std.				0.01
Rob. t-test				4.10
$\sigma_{\mathbf{R70S}}$				**-0.04**
Scaled Estimate				*-0.04*
Rob. Std.				0.01
Rob. t-test				-5.95
$\sigma_{\mathbf{RC}}$				**0.05**
Scaled Estimate				*0.05*
Rob. Std.				0.01
Rob. t-test				4.64

Table 3: Estimation Results

Model	Nb. σ Estimates	Nb. Estimated Parameters	Final L-L	Adjusted Rho-Square
MNL	-	7	-2455.40	0.318
PSL	-	8	-2452.28	0.318
EC_1	1	9	-2437.12	0.322
EC_2	5	13	-2430.76	0.323
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1693 observations Null Log-Likelihood: -3609.92 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 4: Model Fit Measures

	Estimation		Forecast	
	Sample Size	Null L-L	Sample Size	Null L-L
Data 1	1355	-2885.61	338	-724.31
Data 2	1363	-2906.65	300	-703.26
Data 3	1360	-2907.44	333	-702.47
Data 4	1356	-2877.30	337	-732.62
Data 5	1347	-2872.01	346	-737.90

Table 5: Information on Data Sets used for Forecasting Tests

the estimation results on the complete data set, as a result of the decreased sample size. Most parameters remain however significant, at least at 90%. The exceptions are the PS parameter in the PSL model for data set 1, and the estimated travel time parameter in EC_1 and in EC_2 for data set 5.

Regarding the model fit, the general conclusions are the same for all data sets. There is an important increase in model fit when comparing the two EC models with the PSL and MNL models. The PSL is significantly better than the MNL (see likelihood ratio tests in the Appendix, Table 11) except for data set 1 where the PS parameter estimate is not significantly different from zero. Moreover, the EC_2 model is always significantly better than the EC_1 except for data set 3.

We compare the log-likelihood of the data not used for estimation to compare the performance of the different models. The log-likelihood values for all models and data sets are reported in Figure 4, where the superiority of the EC models compared to the MNL and PCL clearly appears. Of course, the differences between the performance of the models regarding prediction results are less prominent than for the estimation results. Interestingly, the prediction performance of the PSL and MNL are very similar, while the fit of estimated data is better for the PSL. These results are very satisfactory given the simplicity of the models. Indeed, there are only three explanatory variables included in the deterministic part of the utility, and no characteristic of the decision-maker is involved.

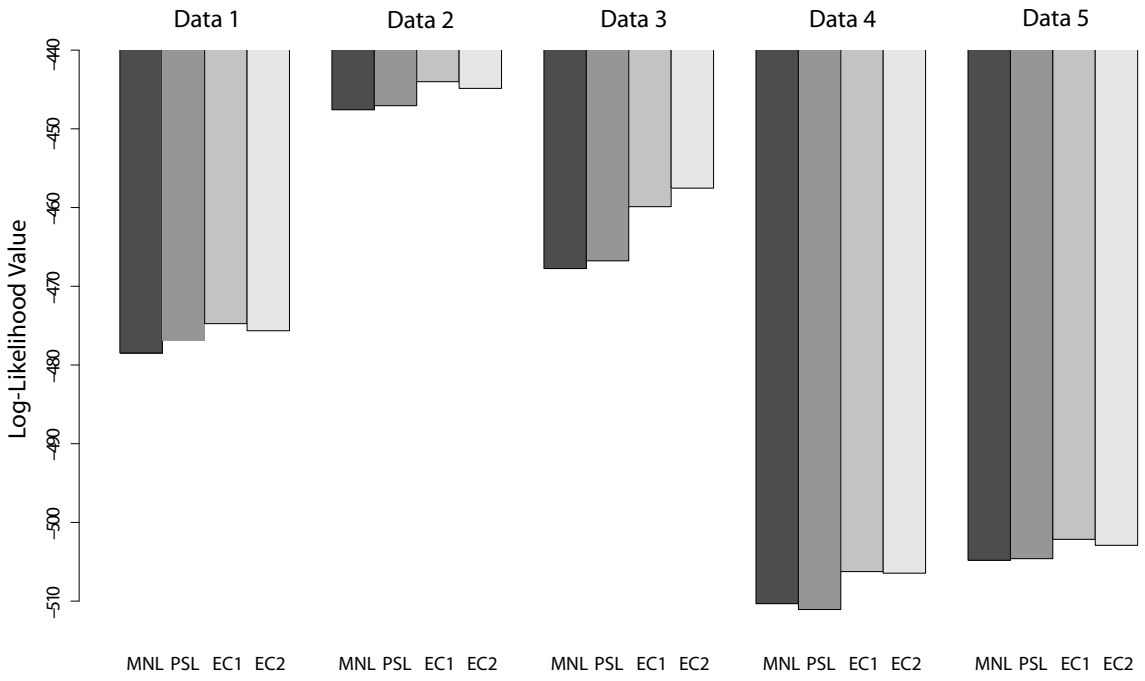

Figure 4: Forecast Test: Final L-L Values for All Models and Data Sets

5 Conclusion

In this paper we propose a novel modelling 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 Path Size parameter estimate is however not significantly different from zero. The estimation results are promising and the estimates of the covariance parameters suggest that the specification captures an important correlation structure.

Preliminary tests of the prediction power of the Error Component, Multinomial Logit and Path Size Logit models are presented. The Error Component model performs better than the Path Size Logit and the Multinomial Logit models. The difference between the Path Size Logit and the Multinomial Logit models are however less clear.

The Path Size Logit model should be used with caution. First of all, the Path Size attribute can be highly correlated with link additive attributes such as free-flow travel time or length. Second, the Path Size values are highly dependent on the definition of the choice set. On the contrary, the subnetwork is defined independently of the choice set. We observe very robust covariance parameter estimates even when the sample of observations used for the estimation varies.

We believe that the subnetwork approach will open new perspectives for large-scale route choice

modelling. It is a flexible approach where the trade-off between complexity and behavioural 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.

6 Acknowledgement

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

References

Axhausen, K., Schönfelder, S., Wolf, J., Oliveira, M. and Samaga, U. (2003). 80 weeks of GPS traces: Approaches to enriching the trip information, *Technical report*, 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 Operational Research* **69**: 97–106.

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

Ben-Akiva, M. and Bierlaire, M. (1999). 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. (2003). *Discrete choice models with applications to departure time and route choice*, Handbook of Transportation Science, 2 edn, Kluwer, chapter 2.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete choice analysis: Theory and application to travel demand*, MIT Press, Cambridge, Massachusetts.

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

Bierlaire, M. (to appear). 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 Transportation and Traffic Theory, 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.

Frejinger, E. and Bierlaire, M. (2005). Route choice models with subpath components, 5th Swiss Transport Research Conference, Monte-Verità, Switzerland.

Frejinger, E. and Bierlaire, M. (2006). Capturing correlation in large-scale route choice models, *Technical report RO-060106*, Operations Research Group ROSO, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

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 repetead choice correlation in route choice modeling by using the mixed logit model, *80th Annual Meeting of the Transportation Research Board*.

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. (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, 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). Link-nested logit model of route choice, Overcoming 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): 195–207.

7 Appendix

	MNL	**PSL**	**EC$_1$**	**EC$_2$**
Nb. Parameters	7	8	9	12
Null L-L	-2885.61	-2885.61	-2885.61	-2885.61
Final L-L	-1977.76	-1976.53	-1964.7	-1956.28
Adj Rho-Square	0.312	0.312	0.316	0.318
Path Size		**0.11**	**-0.04**	**-0.07**
Scaled Estimate		*0.10*	*-0.05*	*-0.08*
Rob. Std.		0.08	0.08	0.08
Rob. t-test		1.41	-0.54	-0.88
Crossing Dummy	**0.42**	**0.40**	**0.48**	**0.48**
Scaled Estimate	*0.42*	*0.38*	*0.57*	*0.55*
Rob. Std.	0.25	0.24	0.27	0.26
Rob. t-test	1.69	1.66	1.77	1.83
Estimated Time $\geq$ 10 min	**-0.06**	**-0.07**	**-0.05**	**-0.05**
Scaled Estimate	*-0.06*	*-0.06*	*-0.06*	*-0.06*
Rob. Std.	0.03	0.03	0.03	0.03
Rob. t-test	-2.30	-2.51	-1.93	-2.02
Left Turns	**-0.39**	**-0.39**	**-0.41**	**-0.41**
Scaled Estimate	*-0.39*	*-0.37*	*-0.48*	*-0.47*
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-23.81	-23.76	-22.73	-22.82
σ			**0.04**	
Scaled Estimate			*0.05*	
Rob. Std.			0.01	
Rob. t-test			6.33	
$\sigma_{\mathbf{R50N}}$				**-0.05**
Scaled Estimate				*-0.06*
Rob. Std.				0.02
Rob. t-test				-3.59
$\sigma_{\mathbf{R70N}}$				**-0.04**
Scaled Estimate				*-0.05*
Rob. Std.				0.01
Rob. t-test				-3.63
$\sigma_{\mathbf{R70S}}$				**-0.04**
Scaled Estimate				*-0.04*
Rob. Std.				0.01
Rob. t-test				-5.78
$\sigma_{\mathbf{RC}}$				**0.06**
Scaled Estimate				*0.06*
Rob. Std.				0.01
Rob. t-test				4.57
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1355 observations Null Log-Likelihood: -2885.61 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 6: Estimation Results Data 1

	MNL	**PSL**	**EC$_1$**	**EC$_2$**
Nb parameters	7	8	9	12
Null L-L	-2906.65	-2906.65	-2906.65	-2906.65
Final L-L	-2009.30	-2006.94	-1995.90	-1989.86
Adj Rho-Square	0.309	0.307	0.310	0.311
Path Size		**0.15**	**0.02**	**-0.01**
Scaled Estimate		*0.14*	*0.02*	*-0.01*
Rob. Std.		0.08	0.08	0.08
Rob. t-test		1.92	0.27	-0.13
Crossing Dummy	**0.71**	**0.69**	**0.72**	**0.77**
Scaled Estimate	*0.71*	*0.65*	*0.73*	*0.76*
Rob. Std.	0.24	0.24	0.26	0.26
Rob. t-test	2.94	2.88	2.72	3.00
Estimated Time $\geq$ 10 min	**-0.10**	**-0.10**	**-0.09**	**-0.10**
Scaled Estimate	*-0.10*	*-0.10*	*-0.10*	*-0.10*
Rob. Std.	0.03	0.03	0.03	0.03
Rob. t-test	-2.89	-3.12	-3.07	-3.09
Left Turns	**-0.38**	**-0.38**	**-0.39**	**-0.40**
Scaled Estimate	-0.38	-0.36	-0.40	-0.40
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-23.55	-23.56	-23.00	-22.93
σ			**0.03**	
Scaled Estimate			*0.04*	
Rob. Std.			0.01	
Rob. t-test			6.24	
$\sigma_{\mathbf{R50N}}$				**-0.04**
Scaled Estimate				*-0.04*
Rob. Std.				0.02
Rob. t-test				-2.60
$\sigma_{\mathbf{R70N}}$				**-0.05**
Scaled Estimate				-0.05
Rob. Std.				0.01
Rob. t-test				-4.16
$\sigma_{\mathbf{R70S}}$				**-0.03**
Scaled Estimate				*-0.03*
Rob. Std.				0.01
Rob. t-test				-4.99
$\sigma_{\mathbf{RC}}$				**0.05**
Scaled Estimate				*0.05*
Rob. Std.				0.01
Rob. t-test				4.28
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1363 observations Null Log-Likelihood: -2906.65 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 7: Estimation Results Data 2

	MNL	**PSL**	**EC_1**	**EC_2**
Nb parameters	7	8	9	12
Final L-L	-1990.77	-1988.16	-1979.87	-1977.09
Adj Rho-Square	0.313	0.316	0.316	0.316
Path Size		**0.16**	**0.05**	**0.03**
Scaled Estimate		*0.15*	*0.06*	*0.03*
Rob. Std.		0.07	0.08	0.08
Rob. t-test		2.12	0.70	0.36
Crossing Dummy	**0.84**	**0.80**	**0.84**	**0.86**
Scaled Estimate	*0.84*	*0.76*	*0.95*	*0.96*
Rob. Std.	0.25	0.24	0.26	0.25
Rob. t-test	3.41	3.34	3.26	3.41
Estimated Time $\geq$ 10 min	**-0.09**	**-0.09**	**-0.08**	**-0.08**
Scaled Estimate	*-0.09*	*-0.09*	*-0.09*	*-0.09*
Rob. Std.	0.03	0.03	0.03	0.03
Rob. t-test	-2.73	-3.05	-2.76	-2.77
Left Turns	**-0.38**	**-0.38**	**-0.40**	**-0.40**
Scaled Estimate	*-0.38*	*-0.36*	*-0.45*	*-0.45*
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-22.56	-22.85	-22.13	-22.05
σ			**0.03**	
Scaled Estimate			*0.04*	
Rob. Std.			0.01	
Rob. t-test			5.66	
$\sigma_{\mathbf{R50N}}$				**-0.03**
Scaled Estimate				*-0.04*
Rob. Std.				0.02
Rob. t-test				-2.08
$\sigma_{\mathbf{R70N}}$				**-0.04**
Scaled Estimate				*-0.05*
Rob. Std.				0.01
Rob. t-test				-3.22
$\sigma_{\mathbf{R70S}}$				**-0.04**
Scaled Estimate				*-0.04*
Rob. Std.				0.01
Rob. t-test				-5.06
$\sigma_{\mathbf{RC}}$				**0.04**
Scaled Estimate				*0.04*
Rob. Std.				0.01
Rob. t-test				3.69
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1360 observations Null Log-Likelihood: -2907.44 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 8: Estimation Results Data 3

	MNL	**PSL**	**EC$_1$**	**EC$_2$**
Nb parameters	7	8	9	12
Final L-L	-1945.24	-1942.03	-1934.48	-1928.3
Adj Rho-Square	0.324	0.322	0.325	0.326
Path Size		**0.18**	**0.07**	**0.03**
Scaled Estimate		*0.15*	*0.06*	*0.03*
Rob. Std.		0.08	0.08	0.08
Rob. t-test		2.35	0.85	0.44
Crossing Dummy	**0.56**	**0.52**	**0.62**	**0.64**
Scaled Estimate	*0.56*	*0.43*	*0.58*	*0.56*
Rob. Std.	0.25	0.24	0.27	0.26
Rob. t-test	2.27	2.18	2.32	2.47
Estimated Time $\geq$ 10 min	**-0.04**	**-0.05**	**-0.04**	**-0.05**
Scaled Estimate	*-0.04*	*-0.04*	*-0.04*	*-0.04*
Rob. Std.	0.02	0.02	0.03	0.03
Rob. t-test	-1.71	-2.07	-1.70	-1.74
Left Turns	**-0.40**	**-0.40**	**-0.41**	**-0.42**
Scaled Estimate	*-0.40*	*-0.33*	*-0.38*	*-0.36*
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-23.90	-23.88	-23.09	-23.16
σ			**0.03**	
Scaled Estimate			*0.03*	
Rob. Std.			0.01	
Rob. t-test			5.78	
$\sigma_{\mathbf{R50N}}$				**-0.04**
Scaled Estimate				*-0.04*
Rob. Std.				0.02
Rob. t-test				-2.51
$\sigma_{\mathbf{R70N}}$				**-0.05**
Scaled Estimate				*-0.05*
Rob. Std.				0.01
Rob. t-test				-3.75
$\sigma_{\mathbf{R70S}}$				**-0.04**
Scaled Estimate				*-0.03*
Rob. Std.				0.01
Rob. t-test				-5.51
$\sigma_{\mathbf{RC}}$				**0.04**
Scaled Estimate				*0.03*
Rob. Std.				0.01
Rob. t-test				3.15
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1356 observations Null Log-Likelihood: -2877.3 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 9: Estimation Results Data 4

	MNL	**PSL**	**EC$_1$**	**EC$_2$**
Nb parameters	7	8	9	12
Final L-L	-1943.66	-1940.38	-1928.45	-1923.42
Adj Rho-Square	0.321	0.322	0.325	0.326
Path Size		**0.18**	**0.04**	**0.01**
Scaled Estimate		*0.15*	*0.05*	*0.02*
Rob. Std.		0.08	0.08	0.08
Rob. t-test		2.32	0.48	0.16
Crossing Dummy	**0.64**	**0.63**	**0.66**	**0.69**
Scaled Estimate	*0.64*	*0.53*	*0.84*	*0.85*
Rob. Std.	0.25	0.24	0.26	0.25
Rob. t-test	2.60	2.63	2.51	2.70
Estimated Time $\geq$ 10 min	**-0.05**	**-0.06**	**-0.04**	**-0.04**
Scaled Estimate	*-0.05*	*-0.05*	*-0.05*	*-0.05*
Rob. Std.	0.03	0.03	0.03	0.03
Rob. t-test	-1.74	-2.06	-1.51	-1.57
Left Turns	**-0.39**	**-0.39**	**-0.41**	**-0.41**
Scaled Estimate	*-0.39*	*-0.33*	*-0.52*	*-0.51*
Rob. Std.	0.02	0.02	0.02	0.02
Rob. t-test	-23.55	-23.52	-22.74	-22.81
σ			**0.04**	
Scaled Estimate			*0.05*	
Rob. Std.			0.01	
Rob. t-test			6.57	
$\sigma_{\mathbf{R50N}}$				**-0.05**
Scaled Estimate				*-0.06*
Rob. Std.				0.01
Rob. t-test				-3.59
$\sigma_{\mathbf{R70N}}$				**-0.05**
Scaled Estimate				*-0.07*
Rob. Std.				0.01
Rob. t-test				-4.04
$\sigma_{\mathbf{R70S}}$				**-0.04**
Scaled Estimate				*-0.05*
Rob. Std.				0.01
Rob. t-test				-5.00
$\sigma_{\mathbf{RC}}$				**0.04**
Scaled Estimate				*0.05*
Rob. Std.				0.01
Rob. t-test				3.92
1000 pseudo-random draws for Maximum Simulated Likelihood estimation 1347 observations Null Log-Likelihood: -2872.01 BIOGEME (roso.epfl.ch/biogeme) has been used for all model estimations (Bierlaire, 2003).				

Table 10: Estimation Results Data 5

Data Set	Model 1	Model 2	Test	Threshold (95 %)	Threshold (90%)
1	PSL	MNL	**2.46**	3.84	2.71
1	EC_2	EC_1	16.84	7.81	6.25
2	PSL	MNL	4.72	3.84	2.71
2	EC_2	EC_1	12.08	7.81	6.25
3	PSL	MNL	5.22	3.84	2.71
3	EC_2	EC_1	**5.56**	7.81	6.25
4	PSL	MNL	6.42	3.84	2.71
4	EC_2	EC_1	12.36	7.81	6.25
5	PSL	MNL	6.56	3.84	2.71
5	EC_2	EC_1	10.06	7.81	6.25

Table 11: Likelihood Ratio Tests