Conference Paper

Accounting for route overlap in urban and suburban route choice decisions derived from GPS observations

**Author(s):**
Axhausen, Kay W.; Schüssler, Nadine

**Publication Date:**
2009

**Permanent Link:**
https://doi.org/10.3929/ethz-a-005916981

**Rights / License:**
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.
Accounting for Route Overlap in Urban and Suburban Route Choice Decisions Derived from GPS Observations

Nadine Schuessler, IVT, ETH Zurich, Switzerland; Kay W. Axhausen, IVT, ETH Zurich, Switzerland

Abstract Two of the most prominent factors influencing driver’s route choice behaviour in an urban or suburban environment are the large number of alternatives available to the decision-maker and the similarity between route alternatives, both caused by the high density of urban and suburban street networks. Regarding the modelling of route choice behaviour this high density influences the way the choice set is established as well as the way the overlap between route alternatives is accounted for. This paper addresses both issues for car route choice observations derived from GPS data. Different choice set generation procedures as well as choice set sizes are evaluated regarding their effect on the choice set composition and the resulting route choice models. In addition, the impact of different route attributes is investigated. The focus, however, will be put on the analysis of the similarity factors, which account for route overlap. Different formulations will be tested in order to evaluate which mechanisms are at work in car route choice in an urban or suburban context.

1 Introduction

Driver’s route choice behaviour in an urban or suburban environment is influenced by a wide variety of factors. Two of the most prominent ones are the large number of alternatives available to the decision-maker and the similarity between route alternatives. Both are caused by the high density of urban and suburban street networks and both strongly influence and potentially bias the results of the route choice models. Neither the decision-maker nor the analyst is able to evaluate the full set of alternatives, the universal choice set. Thus, the route choice and the route choice model have to be based on a subset of alternatives. The composition of the true choice set considered by the decision-maker depends on the technique he or she employed to extract routes from the network and the similarity between alternatives. However, how the decision-makers actually derive their choice set and what behavioural implications the similarity between alternatives has, is still an ongoing research issue. Traditional literature often focusses
either on the modelling of the choice set or on the treatment of similarities between alternatives. Only recently, some studies (e.g. Prato and Bekhor 2007; Bekhor et al. 2006; Bliemer and Bovy 2008) systematically investigated the interdependencies between these two aspects.

This paper aims to continue this line of research and examine the influence of different choice set generation approaches, similarity treatments, and their interdependencies for a route choice model based on very high-resolution data. The route observations originate from a person-based GPS study conducted with 2434 participants residing in Zurich. The network employed for the map-matching and the choice set generation is a very high-resolution navigation network. The high level of detail amplifies several of the issues concerned with choice set generation and similarity treatment since the choice set generation algorithms are more likely to produce routes with only slight deviations from each other, i.e. more overlap. Thus, the algorithms need to explore more routes to determine all the routes relevant for the choice decision. Since they also require more computation time than they would on a smaller network, an important question is how many routes have to be explored to derive the relevant choice set and stable modelling results. In addition, it is investigated how a reduction of the choice set size, as suggested by Bovy (2009), following different paradigms influences the stability of the modelling results. Of course both questions cannot be addressed without appropriately accounting for similarities.

The rest of the paper is structured as follows. After an overview about the related work, the modelling approach and the ways to account for similarity between the route alternatives examined in this paper the data set are presented. Then, Section 4 gives a short describes the processing of the GPS data and the employed choice set generation approaches. Moreover, the techniques to reduce the choice set size are introduced. Section 5 discusses the results of the model estimation followed by the conclusions and an outlook in Section 6.

2 Related work

Starting from the universal choice set $U$, the analyst has three different options to derive the individual choice set $C_n$ of decision-maker $n$. The first option is to model membership of each alternative $i$ to $C_n$ explicitly. Prominent examples for this approach are Swait (2001), Swait and Ben-Akiva (1987) or Morikawa (1996). They use the framework by Manski (1977) who stated that the probability that decision maker $n$ chooses $i$ from $U$ depends on the probability $P(i|C_n)$ that he chooses $i$ from $C_n$ and the probability $P(C_n|U)$ that $C_n \subset U$ is his actual choice set. This approach, however, requires the analyst to enumerate all alternatives of $U$ and its computational complexity is already too high for medium sized problems. Therefore, it is not applicable to route choice situation where the $U$
is huge and unknown to the analyst.

The second option is the implicit modelling of the inclusion in the choice set. Based on the assumption that the membership of the an alternative to the individual choice set depends on its attributes, a continuous variable is added to the systematic part of the utility function of each alternative indicating its degree of inclusion in the choice set. This variable is either derived from assumptions about the availability or perception of an alternative (Cascetta and Papola [2001]) or from constraints which act as cut-offs and cannot be compensated by other attributes (Martínez et al. [2009]). These kinds of approaches are often interpreted as heuristic approximations of the explicit choice set models. However, a recent study by Bierlaire et al. (2009) raised the question if the results of these approximations do indeed concur with the outcomes of an explicit choice set model or if they capture different mechanisms. Yet, the main problem for the application of the implicit choice set modelling approach to route choice modelling is again the underlying assumption that the universal choice set is known to the analyst.

Consequently, in route choice modelling mainly the third option of generating the choice set in a step prior to the modelling is used. First, a set of routes is extracted from the network with the aim to derive a master set $M$ as exhaustive as possible in order to ensure that all relevant alternatives are detected. For this, various approaches have been suggested that can be categorised into deterministic repeated least cost path methods (e.g. Ben-Akiva et al. [1984], de la Barra et al. [1993], Azevedo et al. [1993], van der Zijpp and Fiorenzo-Catalano [2005], Schüssler et al. [2010]), stochastic repeated least cost path procedures (e.g. Bliemer et al. [2007], Bovy and Fiorenzo-Catalano [2007], Ramming [2002]), constrained enumeration techniques (e.g. Prato and Bekhor [2006], Hoogendoorn-Lanser et al. [2006]), and random sampling approaches (e.g. Frejinger [2007]). For a more detailed discussion see the recently published papers by Bovy et al. (2009) or Prato (2009).

After having established the master set $M$, the individual choice set $C_n$ can be obtained by reducing $M$ considering attractiveness, plausibility and overlap of the routes (Bovy [2009]).

Once the individual route choice set $C_{in}$ is established, the choice of route $i$ from $C_{in}$ has to be modelled. This is usually done assuming Random Utility Maximisation (RUM). Various factors originating from attributes of the routes, traffic conditions, the environment around the route, the choice situation or the decision-maker can influence the route choice. One of the most prevalent factors, however, is the similarity between alternatives. Especially in urban or suburban networks the routes of a choice set can overlap extensively. In general, there are three ways to account for similarities in route choice: Allowing for non-zero off-diagonal elements in the variance-covariance matrix of the errors, employing factorial error components as replacement of or in addition to i.i.d. Gumbel errors, and adjustment terms in the systematic part of the utility function.
The only representatives of the class of models with non-zero off-diagonal elements in the variance-covariance matrix that have been applied to route choice modelling are the Paired Combinatorial Logit (PCL) model (Chu, 1989; Koppelman and Wen, 2000) and the Link-Nested Logit (LNL) model (Vovsha et al., 2002). Due to their computational complexity they have only been applied to small to medium size choice sets. The most extensive example documented in the literature is the application of the LNL to choice sets of up to 55 alternatives and up to about 900 links (Ramming, 2002; Bekhor et al., 2006). The resulting matrix that describes the nesting structure is at least 1 order of magnitude smaller than it would be for the route choice problem at hand with up to 100 alternatives and up to 3600 links per choice set.

Similar problems occur for the route choice models based on Probit (Yai et al., 1997) or Mixed Multinomial Logit (MMNL) models (Ramming, 2002; Bekhor et al., 2002). The only MMNL model with a feasible computation time for a route choice model is the Subnetwork model by Frejinger and Bierlaire (2007). In this approach a subnetwork is a continuous subsection of the network that is easily identifiable and behaviourally relevant. Routes using the same subnetwork are assumed to be correlated even if they are not physically overlapping. For them a joined error component is estimated. However, applying the Subnetwork model to very high-resolution data is computationally very demanding and thus omitted in this study.

Adjustment terms assign a value to a specific source of similarity between alternatives. They are calculated prior to the model estimation and treated as another attribute in the deterministic part of the utility. Thus, they only lead to a small increase in computation time and are especially suitable for the modelling of route choice situation. The two main approaches discussed in the literature are the C-Logit model by Cascetta et al. (1996) and the Path Size (PS) Logit model first introduced by Ben-Akiva and Bierlaire (1999). For the PS model modified versions have been proposed by Ramming (2002) and Bovy et al. (2008). Moreover, Hoogendoorn-Lanser and Bovy (2007) presented a PS model that allows to determine a separate Path Size factor for each part of the trip. In this paper, several formulations of these adjustment factors will be tested.

3 Modelling Approach

The general model form used in this analysis is a linear in parameters Multinomial Logit (MNL) model employing the utility function proposed by McFadden (1974). In the MNL model, the utility \( U_{in} \) of alternative \( i \) for decision-maker \( n \) consists of two parts: a deterministic component \( V_{in} \), defined by a vector \( \beta \) of taste coefficients and a vector \( x_{in} \) of attributes of the alternative, and a stochastic component \( \varepsilon_{in} \) comprising identically and
independently (i.i.d.) Gumbel distributed error terms:
\[ U_{in} = V_{in} + \varepsilon_{in} = f(\beta, x_{in}) + \varepsilon_{in} \] (1)

The choice probability of alternative \( i \) can then be calculated as:
\[ P(i|C_n) = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}} \] (2)

For the basic model without accounting for similarities several variables and formulations were tested. Models with a travel impedance based on distance, free flow travel time and time-of-day dependent travel time were compared against each other. The time-of-day dependent travel time values were derived from a MATSim \footnote{MATSIm} \footnote{MATSIm-T, 2008} run of the study area. Similar to the setup reported in \cite{Balmer et al., 2009}, the run included a demand relaxation process with routes and times employing the Charypar/Nagel utility function \footnote{Charypar and Nagel, 2005} for a synthetic population based on the Swiss Census 2000 \footnote{Swiss Federal Statistical Office, 2000} on the Swiss Navteq network. As expected, the models employing time-of-day dependent travel times resulted in the best model fit and the most reliable travel time parameters.

Moreover, it was investigated if the sensitivity towards the travel time differs depending on the type of road the traveller is driving on, as previous studies \cite{Bierlaire et al., 2006} have shown. Therefore, each link of the network was assigned one of four major road types (motorway, extra-urban main road, urban main road, and local road) based on the hierarchical road types coded in the Swiss Navteq network. For each route alternative, the amount of travel time spend on the respective road type was determined. A separate travel time parameter was estimated for each road type. These models using the differentiated travel time parameters systematically outperformed models with only one travel time parameter and are thus used in the subsequent analysis.

In order to obtained unbiased travel time parameters, an additional correction element for the perception of different travel times was introduced following the approach presented by \cite{Bierlaire et al., 2006}. Instead of using a fixed value, the constants for each road type were weighted by the proportion of travel time relative to the total travel time of the route spend on the respective road type. Thereby, the road type proportion values add up to one. In the model estimation, the travel time proportion on extra-urban main roads was defined as reference category to facilitate the interpretation of the model results.

Summarising these findings, the deterministic part of the utility function
is formulated as following:

\[
V_{in} = \beta_{ttMW} \cdot tt_{MW} + \beta_{ttEU} \cdot tt_{EU} + \beta_{ttUM} \cdot tt_{UM} + \beta_{ttLR} \cdot tt_{LR} \\
+ \beta_{rtpMW} \cdot rtp_{MW} + \beta_{rtpUM} \cdot rtp_{UM} + \beta_{rtpLR} \cdot rtp_{LR}
\]

(3)

where \(tt_{MW}\) is the time-of-day dependent travel time on motorways, \(tt_{EU}\) is the time-of-day dependent travel time on extra-urban main roads, \(tt_{UM}\) is the time-of-day dependent travel time on urban main roads, and \(tt_{LR}\) is the time-of-day dependent travel time on local roads. \(rtp_{MW}\), \(rtp_{UM}\), and \(rtp_{LR}\) are the proportions of travel time on motorways, urban main roads and local roads, respectively. The proportion of travel time travelled on extra-urban main roads serves as reference categories. The \(\beta\) represent the parameters to be estimated.

Due to the size of the problem, the only way to account for similarities within a reasonable computation time are adjustment terms. Different formulations of the C-Logit model and the Path Size logit model, including an adaptation of the trip part specific Path Size proposed by Hoogendoorn-Lanser and Bovy (2007) were tested. All adjustment terms enter the deterministic part of the utility function as follows:

\[
V_{in} = f(\beta, x_{in}) + \alpha g(A_{in})
\]

(4)

where \(g(A_{in})\) is a transformation of the similarity term \(A_{in}\) and \(\alpha\) a parameter to be estimated. Thereby, \(\alpha\) can take positive as well as negative values because it is unclear a-priori how the similarity between alternatives will affect the choice behaviour [Axhausen and Schüssler (2007)].

The Commonality Factor \(CF_{in}\) of the C-Logit model introduced by Cascetta et al. (1996) indicates the percentage of route length that route \(i\) shares with other routes by comparing the total length of route \(i\) with the length of the overlapping links. Following the work by Ramming (2002), two formulations for \(CF_{in}\) have been examined:

\[
CF1_{in} = \ln \sum_{j \in C_n} \left( \frac{L_{ij}}{\sqrt{L_i \cdot L_j}} \right)^\gamma
\]

(5)

and

\[
CF2_{in} = \ln \left[ 1 + \sum_{j \in C_n, i \neq j} \left( \frac{L_{ij}}{\sqrt{L_i \cdot L_j}} \right) \left( \frac{L_i - L_{ij}}{L_j - L_{ij}} \right) \right]
\]

(6)

where \(L_{ij}\) is the length of the links shared by \(i\) and \(j\), \(\Gamma\), \(L_i\) the set of links of route \(i\), \(l_a\) the length of link \(a\), \(N_{an}\) number of links using link \(a\), and
\( \gamma \) is a parameter to be defined by the analyst or to be calibrated. Due to the lack of suitable calibration data it was set to 1 in this analysis. However, the formulation in Equation 6 sometimes resulted in very large values and disproportionate differences in the \( CF_{in} \) values for alternatives within the same choice set, particularly for large choice sets. Consequently, the parameters for most other attributes were insignificant, suggesting that the similarity between the alternatives was the only decisive attribute. Since this was more founded on the numerical effect than on actual behaviour, the formulation was left out of the subsequent analysis.

In the Path Size logit model, first presented by Ben-Akiva and Bierlaire (1999), the length of each route is corrected by the so-called Path Size \( PS_{in} \). Only a distinct route, i.e. a route with no overlaps with other routes, can get the maximum path size of one. Path Sizes different from one are calculated based on the length of the links within the route \( i \) and the length of the routes that share a link with it relative to the length of the shortest route using the link. In this work, the two original formulations proposed by Ben-Akiva and Bierlaire (1999) are applied:

\[
PS1_{in} = \ln \sum_{a \in \Gamma_i} \left( \frac{l_a}{L_i} \right) \sum_{j \in C_n} \frac{1}{\delta_{aj}} \tag{7}
\]

where \( \Gamma_i \) is the set of all links of path \( i \), \( l_a \) is the length of link \( a \), and \( L_i \) the length of path \( i \). \( \delta_{aj} \) equals one if link \( a \) is on path \( i \) and zero otherwise. The second formulation additionally accounts for the relative ratio between the length of the shortest path \( L_{C_n}^* \) in \( C_n \) using link \( a \) and the length of each path \( j \) using link \( a \).

\[
PS2_{in} = \ln \sum_{a \in \Gamma_i} \left( \frac{l_a}{L_i} \right) \frac{1}{\sum_{j \in C_n} \frac{L_{C_n}^*}{L_j} \delta_{aj}} \tag{8}
\]

In addition, the Path Size Correction (PSC) model recently presented by Bovy et al. (2008) was tested. Bovy et al. (2008) argue that their PSC model has a clear theoretical derivation and outperforms the classic PS model in the empirical application. The PSC factor is defined as follows:

\[
PSC_{in} = -\sum_{a \in \Gamma_i} \left( \frac{l_a}{L_i} \right) \ln \sum_{j \in C_n} \delta_{aj} \tag{9}
\]

The last variant of adjustment terms examined in this study was inspired by the trip part specific Path Size factor developed by Hoogendoorn-Lanser and Bovy (2007) for route choice modelling in multi-modal networks. The experiments that led to the basic model described in Equation 3 have demonstrated that the road types are perceived differently and that the sensitivity to travel time changes depending on the road type. Thus, the hy-
hypothesis is that also the overlap on individual parts of the trip, i.e. different road types, has a varying influence on the choice probability of the alternative. In order to verify this hypothesis, the links belonging to a specific road type were defined as a trip part. Then for each road type $x$ a separate road type specific Path Size factor $PSRT_{ixn}$ was calculated. Models for each of the two following formulations were estimated:

$$PSRT_{1ixn} = \frac{1}{L_i} \sum_{a \in \Gamma_{ix}} l_a \frac{N_{na}}{N}$$

(10)

$$PSRT_{3ixn} = \frac{1}{L_i} \sum_{a \in \Gamma_{ix}} l_a \frac{N_{na}}{N}$$

(11)

where $L_i$ is the length of the of full route $i$, $L_{ix}$ is the travel time on road type $x$ for route $i$, $\Gamma_{ix}$ is the set of all links of road type $x$ of route $i$, $l_a$ is the length of link $a$, and $N_{na}$ is the number of unique full paths using link $a$.

Moreover, a sampling correction term is introduced for the choice sets generated with the Stochastic Choice Set Generation approach. According to Bovy et al. (2009), the Stochastic Choice Set Generation is a case of sampling importance because the probability of a route to be selected depends on its characteristics. Thus, routes with a higher probability to be chosen also have a higher probability to be included in the choice set. In order to account for the unequal selection probabilities Bovy et al. (2009) developed a Sampling Correction term $SC_{in}$ for route choice sets originating from Stochastic Choice Set Generation. The Sampling Correction term also accounts for the influence of spatial overlap on the selection probability. $SC_{in}$ is defined as:

$$SC_{in} = \ln \left( \frac{f_i}{Q_i} \right)$$

(12)

where $f_i$ is the number of times, alternative $i$ was drawn during the stochastic choice set generation and $Q_i$ is the selection probability of $i$ in the choice set generation. $Q_i$ is calculated using Equation 13

$$Q_i = \frac{PS_i \cdot \exp(-C_i/b)}{\sum_{j \in C_n} PS_j \cdot \exp(-C_j/b)}$$

(13)

where $C_i$ is the cost of route $i$ and $b$ is the positive variance parameter. Since a uniform distribution was employed in the Stochastic Choice Set Generation in this study, $b$ is determined differently than in Bovy et al. (2009). Analogously to Bovy et al. (2009) it is assumed that the route error variances are equal for all routes and can be determined from cost of the
shortest path $C_{\text{min}}$. Then, $b$ is defined as:

$$b = \sqrt{\frac{2 \cdot C_{\text{min}}}{\pi}}$$  \hspace{1cm} (14)

The influence of the Sampling Correction term on the modelling results is evaluated in Section 5.

4 Derivation of the modelling data

The dataset used in this paper originates from a study conducted by a private sector company with the aim to determine whether or not participants noticed certain billboards (Pasquier et al., 2008). We obtained the data, but without the socio-economic details of the respondents, from one of the sponsors of the original data collection effort as part of a joint project. 2434 participants residing in Zurich were asked to carry an on-person GPS logger for 6.65 days on average. No additional information, such as modes or trip purposes, was collected. The large amount of data and the lack of respondent-derived information imposed several challenges and required the development of advanced postprocessing methods. As described in Schüssler and Axhausen (forthcoming), the GPS traces have been cleaned and subdivided into trips and activities. For each trip the respective mode-sequence has been established. The result of this procedure are complete trip and activity chains for more than 17’000 person-days. For this analysis, however, only a sub-sample of the the about 37’000 car trips will be employed.

Once the car trips are identified, two further steps have to be undertaken before route choice models can be estimated: identifying the chosen routes and establishing the choice set. The network employed in both steps is the Swiss Navteq network, a high-resolution navigation network covering all regions of Switzerland and containing 408,636 nodes and 882,120 unidirectional links. The link characteristics include free flow speed, length, hierarchical road type, number of lanes, and capacity. Overall, there are 44-times more links in Swiss Navteq network than in the planning network for the same area. This high network resolution is essential for an accurate identification of the chosen routes. However, it substantially increases the computation time for the identification of the chosen routes and the choice set generation and imposes new requirements for the choice set composition.

Regarding the identification of the chosen routes it was considered to use the network-free approach by Bierlaire and Frejinger (2008). However, taking into account the very high-resolution of the network and that each car trip consisted of 620 GPS points on average and 10’000 at a maximum, this idea was abandoned. Instead, the chosen routes were derived with our own
map-matching algorithm (Schüssler and Axhausen, 2009) which builds on the map-matching algorithm presented by Marchal et al. (2005).

In total, four algorithms were evaluated for the generation of the choice sets. The branch & bound algorithm by Prato and Bekhor (2006) and the random walk by Frejinger (2007) were implemented because the respective authors showed that the resulting route sets had attractive properties. Both algorithms, however, proved to be inappropriate for the high-resolution network. The branch & bound algorithm terminated in reasonable computation time only for OD pairs connected by very short paths. Given the exponential increase of computation time with the number of links in the paths (Prato, 2009) an average number of 65.69 links per chosen route was too much for the branch & bound algorithm. The random walk on the other hand proved to be difficult to calibrate. Several parameter setting were tested but non was suitable for the entire data set with its high variation in route lengths. If the parameters were too strict only the shortest path was found. If they were too weak the algorithm wandered around and needed a long time to terminate i.e. find the end node of the route. Then, the resulting route is unreasonably long, i.e. more then 10-times longer than the shortest path. Both aspects lead to impractical computation times even for a small number of alternatives and short routes. Thus, the only algorithms that applicable to choice set generation in the very high-resolution network were the well-known Stochastic Choice Set Generation (SCSG), often also called simulation approach, and a performance optimised Breadth First Search on Link Elimination (BFS-LE) (Schüssler et al., 2010).
Whereas a more detailed performance analysis of the SCSG and the BFS-LE is given in Schüssler et al. (2010). Figure 1 shows a comparison of the run times for all four algorithms for a target choice set size of 20. It has to be noted that in all algorithms a time abort threshold was introduced to capture OD pairs for which the route set generation could not be completed within a time interval predefined by the analyst. In the runs shown in Figure 1, this time abort threshold is set to 90 minutes per OD pair. Since the time criterion is only checked after a route has been completed, an additional abort criterion was imposed on the random walk to prevent it from exploring the network for hours trying to finish just one route and to avoid completely unrealistic routes. If the number of links in the random walk path exceeds ten times the number of links of the shortest path for the given OD pair, the random walk was stopped and restarted. If by then the time abort threshold was violated, the choice set generation for this OD pair was stopped. But even with these measures, the SCSG and the BFS-LE are substantially faster than the random walk and the branch & bound. The BFS-LE by far outperforms the SCSG in terms of computation time.

Bovy (2009) stated, that one recommendable way to derive a choice set for route choice modelling is to first establish a master set that is exhaustive as possible and then to reduce this master set to the individual choice set taking into account the attractiveness, the plausibility and the overlap of the routes. Thus, this paper does not only evaluate the modelling results of choice sets directly derived from the SCSG and BFS-LE choice set generation algorithms but also compares them to the results for choice sets that were established by first generating a 100 alternatives with each of the choice set generation algorithms and then reducing these choice sets following different paradigms. Thereby, special attention is paid to the behavioural mechanisms with respect to route overlap. Four different choice set reduction procedures are investigated in this paper:

- Random reduction
- Similarity distribution-based reduction
- Similarity-based reduction
- Rule-based reduction

The most simple reduction procedure is the random reduction, where route are randomly removed from the master set until only the target number of alternatives is left. The only restriction in this procedure is that the chosen route has to remain in the choice set.

The next two reduction procedures use the overlap between the routes as criterion for the reduction. The objective of the similarity-distribution based reduction is to obtain a choice set with many different levels of overlap whereas the similarity-based reduction aims to derive a choice set that
is as heterogeneous as possible. Both procedures use the Path Size factor defined in Equation \textit{X} as their overlap measure. For the \textit{similarity distribution-based reduction}, first the Path Size factor for each route in the complete choice set is calculated. Then the routes are assigned to previously defined similarity classes according to this Path Size factor. In this case the interval of possible Path Size values between 0 and 1 was subdivided into 10 classes of equal width. Subsequently, the quota of routes for each class is calculated based on the number target routes, the number of classes and the assumption that each class should be equally represented. Then, from each class routes are randomly drawn until the class quota is met. If the number of routes in the choice set is lower than the target choice set size because some classes contain less routes than their quota, additional routes from the classes in which routes remained are drawn in random order until the target choice set size is reached.

The approach employed in the \textit{similarity-based reduction} is based on the work by Kuby \textit{et al.} (1997). The basic idea is to successively add routes from the master set to the choice set in a way that each new route is least similar to the routes already contained in the choice set. First the chosen route is added to the choice set and removed from the master set. Second, the Path Size of each remaining route in the master set with respect to the routes of the choice set, and only those, is calculated. Third, the route with the highest Path Size is added to the choice set and removed from the master set. Consequently, step two and three are repeated until the choice set has the target size.

The \textit{rule-based reduction} evaluates certain characteristics of the alternatives and deletes all route that violate at least one previously defined threshold. The thresholds can be defined in absolute terms or relative to the characteristics of other routes. In this paper, four route characteristics are employed: length, travel time, plausibility, i.e. number of consecutive motorway links and the route overlap. Moreover, two different levels of strictness are applied: a weak and a strong restriction. For the length and the travel time the thresholds are defined relative to the best path according to the respective criteria. The thresholds for the weak and strong restriction level for length are 3.5 and 1.9 respectively. These values correspond to the 85 and 70 percentiles of the distribution of the ratios between the length of a route alternative and the length of the shortest path in the choice set. If a route alternative is longer than the threshold times the length of the shortest route in the choice set it is discarded. The filtering with regard to travel time is carried out analogously. The respective thresholds are 3.5 and 2 times the travel time of the quickest route. The thresholds for the plausibility and the overlap criterion are defined in absolute numbers. The only reasonable assumption concerning the plausibility of a route that could be made in this analysis is that people do rarely enter the motorway for only a small number of links, especially if there are other, shorter and quicker alternatives,
available. Thus, all routes that contained motorway segments of less than three links were removed from the choice set. Finally, all routes that overlapped with at least one route for at least a certain percentage of their own length, were discarded. Two different levels of overlap were tested: 90% at the weak restriction level and 70% at the strong restriction level.

### Table 1: Choice sets used in the model estimation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Choice set size</th>
<th>Reduction procedure</th>
<th>Identification code</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFS-LE</td>
<td>20, 60, 100</td>
<td>–</td>
<td>B20, B60, B100</td>
</tr>
<tr>
<td></td>
<td>20, 60</td>
<td>random</td>
<td>RandB20, RandB60</td>
</tr>
<tr>
<td></td>
<td>20, 60</td>
<td>similarity distribution-based</td>
<td>SimDistB20, SimDistB60</td>
</tr>
<tr>
<td></td>
<td>34, 87</td>
<td>similarity-based</td>
<td>SimB20, SimB60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rule-based</td>
<td>RuleB1, RuleB2</td>
</tr>
<tr>
<td>SCSG</td>
<td>20, 60, 100</td>
<td>–</td>
<td>S20, S60, S100</td>
</tr>
<tr>
<td></td>
<td>20, 60</td>
<td>random</td>
<td>RandS20, RandS60</td>
</tr>
<tr>
<td></td>
<td>20, 60</td>
<td>similarity distribution-based</td>
<td>SimDistS20, SimDistS60</td>
</tr>
<tr>
<td></td>
<td>43, 95</td>
<td>similarity-based</td>
<td>SimS20, SimS60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rule-based</td>
<td>RuleS1, RuleS2</td>
</tr>
</tbody>
</table>

The different reduction procedures lead to different choice set compositions. The subsequent figures give a first insight into these differences for the choice sets used in the model estimations. The overall 22 different choice sets are summarised in Table 1. Each choice set is assigned a code that is used to identify the choice set in the subsequent analyses. With each choice set generation procedure choice sets containing 20, 60 and 100 alternatives were produced. Then, the choice sets containing 100 alternatives were reduced to 20 or 60 alternatives using one of the reduction procedures. An exception from this rule is the rule-based reduction. Here, the number of alternatives in the resulting choice sets varies and depends on the composition of the individual choice set. Table 1 indicates the maximum number of alternatives over all choice sets, while Figure 2 shows the distribution of choice set sizes for the different levels of strictness and the different choice set generation procedures used to produce the master set. It can be seen that especially the strong restriction result in relatively small choice sets containing less than 20 alternatives whereas the weaker restrictions lead to more variety regarding the choice set sizes.

Other indicators of the choice set structure are the distributions of the travel times and the path overlap in the choice sets. The travel time distributions for each choice set is presented in Figure 3. Overall, the mean travel time varies between 8.4 and 12.9 minutes and the median between 6.33 and 10.4 minutes. For both choice set generation algorithms the mean and average travel time increase with increasing choice set sizes. The choice sets originating from the BFS-LE algorithm contain longer trips and more outliers than SCSG choice sets. There is, however, no systematic change
Figure 2: Cumulative distribution of choice set sizes resulting from the rule-based reduction

Figure 3: Travel time distributions for the different choice sets
in the travel time distributions caused by the reduction procedures. Usually, the median travel time decreases slightly after the reduction, but there are also several cases where the median travel time increases and the travel time distribution becomes wider. Only the rule-based reduction lead, as expected, to a distinct decrease in the mean travel time and a narrower travel time distribution.

In order to evaluate the route overlap, Figure 4 shows the distribution of the Path Size factor defined in Equation 8 for each choice set. It can be seen that the level of route overlap differs considerably for the different choice set generation algorithms, reduction procedures and choice set sizes. Unsurprisingly, smaller choice set sizes lead to higher Path Size factors, i.e. smaller overlap, regardless of the procedure used to derive them. The choice set originating from the BFS-LE algorithm contain less overlap than those from the SCSG algorithm, even after the different reduction procedures were applied. Moreover, all choice set reduction procedures, except the random reduction, substantially increase the Path Size. The reduced choice sets contain even less overlap than the choice sets of the same size derived directly from the choice set generation algorithms. This effect is strongest for the rules-based reduction and weakest for the similarity distribution-based reduction.

Figure 4: Distribution of the similarity factor for the different choice sets
5 Modelling results

The aim of this paper is to investigate the influence of different choice set generation approaches, similarity treatments, and their interdependencies on a route choice model for very high-resolution data. Therefore, several choice models have been estimated for different choice sets and different formulations of the similarity factors. While the choice sets are outlined in Table 1, Table 2 gives an overview of the 17 models tested for each choice set. First, a basic model without similarity treatment was estimated. Then, the 6 similarity factors established in Section 3 were added with a logarithmic transformation as suggested by the literature. An exception is the PSC factor by Bovy et al. (2008), where the logarithm is already part of the similarity factor. Afterwards, it was examined whether the natural logarithm is indeed the best transformation. Therefore, additional models were estimated with no transformation and a BoxCox transformation for all similarity factors except the PSC factor. This section shows a selection of these models which were all estimated using BIOGEME (Bierlaire 2009).

Table 2: Models tested for each choice set

<table>
<thead>
<tr>
<th>Name</th>
<th>Similarity factor</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic model</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ln(PS1)</td>
<td>Path Size as in Equation 7</td>
<td>ln</td>
</tr>
<tr>
<td>ln(PS2)</td>
<td>Path Size as in Equation 8</td>
<td>ln</td>
</tr>
<tr>
<td>ln(CF1)</td>
<td>Commonality Factor as in Equation 5</td>
<td>ln</td>
</tr>
<tr>
<td>ln(PSRT1)</td>
<td>Road type specific Path Size as in Equation 10</td>
<td>ln</td>
</tr>
<tr>
<td>ln(PSRT3)</td>
<td>Road type specific Path Size as in Equation 11</td>
<td>ln</td>
</tr>
<tr>
<td>PS1</td>
<td>Path Size as in Equation 7</td>
<td>–</td>
</tr>
<tr>
<td>PS2</td>
<td>Path Size as in Equation 8</td>
<td>–</td>
</tr>
<tr>
<td>CF1</td>
<td>Commonality Factor as in Equation 5</td>
<td>–</td>
</tr>
<tr>
<td>PSRT1</td>
<td>Road type specific Path Size as in Equation 10</td>
<td>–</td>
</tr>
<tr>
<td>PSRT3</td>
<td>Road type specific Path Size as in Equation 11</td>
<td>–</td>
</tr>
<tr>
<td>BoxCox(PS1)</td>
<td>Path Size as in Equation 7</td>
<td>BoxCox</td>
</tr>
<tr>
<td>BoxCox(PS2)</td>
<td>Path Size as in Equation 8</td>
<td>BoxCox</td>
</tr>
<tr>
<td>BoxCox(CF1)</td>
<td>Commonality Factor as in Equation 5</td>
<td>BoxCox</td>
</tr>
<tr>
<td>BoxCox(PSRT1)</td>
<td>Road type specific Path Size as in Equation 10</td>
<td>BoxCox</td>
</tr>
<tr>
<td>BoxCox(PSRT3)</td>
<td>Road type specific Path Size as in Equation 11</td>
<td>BoxCox</td>
</tr>
</tbody>
</table>

For the choice sets derived from the Stochastic Choice Set Generation (SCSG) additional models were tested employing the Sampling Correction proposed by Bovy et al. (2009). Table 3 compares the results with and without the SC factor for the basic model without correction for route overlap and for the choice sets directly derived from the SCSG. In these choice sets, the routes are highly correlated, especially those using mainly urban main roads. Thus, without similarity treatment, the travel time parameter for urban main roads can be positive. If, however, the SC factor is added,
Table 3: Parameters basic model for unreduced SCSG choice sets without similarity treatment and with and without Sampling Correction

<table>
<thead>
<tr>
<th>Parameter</th>
<th>S20</th>
<th>S60</th>
<th>S100</th>
<th>S20</th>
<th>S60</th>
<th>S100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel time on MW</td>
<td>-0.13</td>
<td>-0.37</td>
<td>-0.48</td>
<td>0.14 *</td>
<td>0.11 *</td>
<td>0.13 *</td>
</tr>
<tr>
<td>Travel time on EU</td>
<td>-0.09</td>
<td>-0.34</td>
<td>-0.46</td>
<td>0.13 *</td>
<td>0.10 *</td>
<td>0.10 *</td>
</tr>
<tr>
<td>Travel time on UM</td>
<td>0.14</td>
<td>0.00 *</td>
<td>-0.06</td>
<td>0.38</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>Travel time on LR</td>
<td>-0.49</td>
<td>-0.95</td>
<td>-1.16</td>
<td>-0.26</td>
<td>-0.29</td>
<td>-0.32</td>
</tr>
<tr>
<td>Road type perc. MW</td>
<td>4.68</td>
<td>4.86</td>
<td>5.01</td>
<td>4.31</td>
<td>4.62</td>
<td>4.75</td>
</tr>
<tr>
<td>Road type perc. UM</td>
<td>3.85</td>
<td>2.62</td>
<td>2.19</td>
<td>4.25</td>
<td>3.82</td>
<td>3.74</td>
</tr>
<tr>
<td>Road type perc. LR</td>
<td>0.71 *</td>
<td>-0.25 *</td>
<td>-0.35 *</td>
<td>3.78</td>
<td>3.34</td>
<td>3.34</td>
</tr>
<tr>
<td>Sampling Correction</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Initial LL</td>
<td>-4065</td>
<td>-5218</td>
<td>-5668</td>
<td>-4065</td>
<td>-5218</td>
<td>-5668</td>
</tr>
<tr>
<td>Final LL</td>
<td>-3660</td>
<td>-4629</td>
<td>-5002</td>
<td>-3049</td>
<td>-3788</td>
<td>-4078</td>
</tr>
<tr>
<td>Adjusted Rho square</td>
<td>0.10</td>
<td>0.11</td>
<td>0.12</td>
<td>0.25</td>
<td>0.27</td>
<td>0.28</td>
</tr>
</tbody>
</table>

(*) parameter not significant at 95% confidence level

MW = motorway, EU = extra-urban road, UM = urban main road, LR = local road

Road type percentage are estimated with extra-urban main roads as reference category

also the travel time parameters for motorways and extra-urban roads turn positive. This effect can only be partially corrected by adding similarity factors to the utility function and also occurs, though a little weaker, for the reduced choice sets with less correlation. Thus, the Sampling Correction factor is left out of the subsequent analysis.

With regard to the treatment of similarities between route alternatives, two different aspects have been investigated: the formulation of the similarity factor itself and the transformation used to integrate it in the utility function. Figure 5 shows the impact of the different similarity factors formulations and transformations on the values of the travel time parameters for the individual road types for the unreduced choice choice sets. It can be seen that the absolute values of the travel time parameters increase with increasing choice set size and also with the inclusion of similarity factor. While no clear trend could be detected concerning the transformation of the similarity factor, the influence of the travel time for all road types tends to be lower for models employing road type specific Path Size factor formulations.

In order to determine the best transformation for each similarity factor, the adjusted rho square values for the same models are presented in Table 4. Please note that the adjusted rho squares are only comparable for models estimated from the same choice set and cannot be used to evaluate the quality of a choice set itself. The models applying a BoxCox transformation performed for all choice sets and all similarity factors at least equally well as, if not better than, the logarithmic transformation or no transformation at all. Moreover, the BoxCox transformed models, unlike the other models, delivered stable and reasonable results with respect to the travel time parameters for the individual road types over all choice sets and with re-
(a) Motorways

(b) Extra-urban roads

(c) Urban main roads

(d) Local roads

Figure 5: Travel time parameters for models estimated on unreduced choice sets

gard to the overlap punishment for the different road types in the road type specific Path Size factors. Only for the models employing BoxCox transformed similarity factors, all travel time parameters were negative for all choice sets and the relative order of the overlap punishment remained the same. Thus, only the BoxCox similarity factors are considered in the subsequent analysis. As explained above, the PSC factor is excluded from this rule because here the transformation happens within the calculation of the similarity factor itself.

Before the impact of the different choice sets on the modelling results is scrutinised, the relative order of the overlap punishment for the different
Figure 6: Utility correction for route overlap per road type depending on the degree of overlap
### Table 4: Adjusted rho squares for the unreduced choice sets

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B20</th>
<th>B60</th>
<th>B100</th>
<th>S20</th>
<th>S60</th>
<th>S100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>no transformation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic model</td>
<td>0.12</td>
<td>0.19</td>
<td>0.21</td>
<td>0.10</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>PS1</td>
<td>0.13</td>
<td>0.20</td>
<td>0.23</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>PS2</td>
<td>0.13</td>
<td>0.20</td>
<td>0.24</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>PS4</td>
<td>0.13</td>
<td>0.20</td>
<td>0.22</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>CF1</td>
<td>0.15</td>
<td>0.21</td>
<td>0.23</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>PSRT1</td>
<td>0.15</td>
<td>0.22</td>
<td>0.26</td>
<td>0.13</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>PSRT3</td>
<td>0.16</td>
<td>0.22</td>
<td>0.25</td>
<td>0.13</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>ln transformation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(PS1)</td>
<td>0.12</td>
<td>0.19</td>
<td>0.22</td>
<td>0.10</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>ln(PS2)</td>
<td>0.13</td>
<td>0.19</td>
<td>0.22</td>
<td>0.10</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>ln(CF1)</td>
<td>0.15</td>
<td>0.21</td>
<td>0.24</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>ln(PSRT1)</td>
<td>0.15</td>
<td>0.21</td>
<td>0.24</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>ln(PSRT3)</td>
<td>0.14</td>
<td>0.20</td>
<td>0.23</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>BoxCox transformation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BoxCox(PS1)</td>
<td>0.14</td>
<td>0.21</td>
<td>0.24</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>BoxCox(PS2)</td>
<td>0.14</td>
<td>0.21</td>
<td>0.24</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>BoxCox(CF1)</td>
<td>0.15</td>
<td>0.21</td>
<td>0.24</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>BoxCox(PSRT1)</td>
<td>0.16</td>
<td>0.23</td>
<td>0.26</td>
<td>0.14</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>BoxCox(PSRT3)</td>
<td>0.16</td>
<td>0.22</td>
<td>0.25</td>
<td>0.13</td>
<td>0.15</td>
<td>0.16</td>
</tr>
</tbody>
</table>

road types is examined by calculating magnitude of the utility correction for each road type depending on the degree of overlap measured by the road type specific Path Size. In Figure 6 the results of this calculation are exemplarily depicted for the first formulation of the road type specific Path Size (PSRT1) and the unreduced choice sets. For all these choice sets, and all other choice sets, overlap on motorways is punished the least and overlap on urban main roads the most. In between, overlap on extra-urban roads is punished considerably less than overlap on local roads. This is reasonable since the motorway network in Switzerland is not very dense. Thus, it is difficult to find alternative routes that use a different motorway (segment) for the same OD relation. A similar logic applies to extra-urban roads that connect built-up areas and where the avoiding these routes would lead to large detours. In built-up areas, on the other hand, small detours are more easily available but also easily overlooked. Therefore, the difference between the large number of objectively available alternatives and the number of subjectively perceived alternatives is higher, especially concerning urban-main roads. People usually navigate using main roads as major decision-criterion and take always the same minor roads to access the main roads. Thus, alternatives using the same part of the main road network often collapse to one alternative in the perception of the traveller.

The impact of the random choice set size reduction on the travel time parameters can be seen in Figure 7 where the parameter values are compared to those for the unreduced choice sets with 100 alternatives. For all
randomly reduced choice sets the absolute values of the travel time parameters decrease with decreasing choice set size though this decrease is less strong than for the unreduced choice sets of the same size. The same applies for the parameters of the similarity factors. Apparently, the generation of the full choice set of 100 alternatives and the subsequent reduction causes an approximation of the parameters to those of the full choice set. Another effect is a smoothing of the parameter values with respect to the similarity factor. The travel time parameters vary less depending on the different similarity factors than they do it for the unreduced choice sets in Figure 5.

For the choice sets obtained from the similarity distribution-based re-
Figure 8: Travel time parameters for models estimated on similarity distribution-based reduced choice sets

In the reduced choice sets, the travel time parameters approximate the values for the unreduced choice sets with 100 alternatives even closer as presented in Figure 8. Also, the smoothing effect is more distinct. Sometimes the absolute value of the travel time parameter for the reduced choice set is even higher than the one for the full choice set, for example for the SimDistS20 choice set and the models employing BoxCox transformed CF1, PSRT1 or PSRT3 similarity factors. One exception to this trend is the SimDisB20 choice set. In this choice set, the route overlap has been reduced substantially as shown in Figure 4. This leads to negative beta parameters for the similarity factors.
for PS1, PS2, PSC implying that route overlap would have a positive influence on the utility of an alternative. This in turn leads to a decrease in the absolute values of the travel time parameters on the individual road types for these models. The PSRT1 model, on the other hand, indicates that the positive influence of the route overlap originates from the local roads. In this model, only the parameter for the local road similarity factor is negative. Accordingly, the travel time parameter for travel on local roads is also higher than the one estimated for the unreduced choice set while the travel time parameters for the other road types are close to the unreduced ones.
Figure 9 depicts the travel time parameters for the similarity-based reduced choice sets. The parameters for the SimB60, SimS20 and SimS60 choice sets show a similar behaviour as the corresponding choice sets derived through similarity distribution-based reduction. Though the parameters for the similarity-based choice sets with 60 alternatives are a little bit lower than those for the similarity-distribution based choice sets, the overall approximation and smoothing trend can be confirmed. Exceptions to this are the local road travel time parameter for the SimS20 choice set and all parameters for the SimB20 choice set. The local road parameter for the SimS20 choice set reveals a higher sensitivity to travel time. It is closer to the values for the choice sets derived with the BFS-LE than to those derived with the SCSG algorithm except for the model employing the BoxCox transformed PSRT1 similarity factors, where the parameter for the local road similarity factor is again negative. The patterns of the travel time parameters estimated from the SimB20 model resemble those estimated from the SimDistB20 model. They do, however, jump around even more with respect to the different forms of similarity correction, especially the local road parameter. The most stable parameters, and the ones closest to the full choice set, deliver again the models with road type specific Path Size Correction. In the PSRT1 model not only the local road similarity factor parameter is negative but also the urban main road similarity factor parameter. This leads to the conclusion that the similarity-based reduction procedure, and to a lesser extent also the similarity distribution-based reduction procedure, do not remove the correlation between the routes uniformly over all route types but stronger for road types higher in the hierarchy, probably because they usually make up the longer part of a route. Moreover, correlation on local roads at the start and end of a trip is sometimes unavoidable because there are only a few local roads leading from the origin or to the destination.

As shown in Figure 4, the rule based reduction procedures, especially the one employing the stricter parameters, were most efficient in the reduction of route overlap in the choice sets. Moreover, unlike the other choice set reduction procedures, they reduced the travel times considerably. This is also reflected in the travel time parameters estimated for these choice sets that are presented in Figure 10. Especially the travel time parameters for the stronger reduced choice sets RulesB2 and RulesS2 are substantially lower than the travel time parameters for the other choice sets because the travel time plays a less decisive role if the participant has to choose from a choice set in which the travel times are rather similar. Thus, the travel time parameters are not really comparable to the ones estimated from other choice sets. It can, however, be seen that the travel time parameters are more stable with respect to the similarity treatment than they are for other choice sets with comparably low correlation. They all follow a similar pattern that is different from the patterns observable for other choice sets. More-
over, the distance between the travel time parameters for these two choice sets is considerably smaller than it is between other choice sets with corresponding choice set size and reduction procedure but produced with different choice set generation algorithms. This might imply that the choice sets somehow converge in their characteristics. An approximation in the distribution of travel times and route overlap can already be observed in Figures 4 and 3. It has to be noted, though, that the BoxCox(PSRT1) model did not converge for the RulesB2 choice set. The values for the logarithmic transformed PSRT1 model (LN(PSRT1)) are depicted here instead because it outperformed the model without transformation.
Summing up, Table 5 shows the estimation results for the Box-Cox(PSRT1) model for the choice sets that turned out to be most appropriate in this study. All the travel time parameters have the right sign and lie within a similar range with travel time on motorways being punished least and travel time on local roads punished most. Of the parameters indicating a preference for a particular road type only the local road parameter is in some models significant implying that the road type specific aspects are to a large extent covered in the road type specific travel times and the road type specific Path Sizes. In contrast to the results for the unreduced choice sets, the parameters for the motorway and extra-urban road Path Size factors are rarely ever significantly. This confirms that in the reduction procedure the overlap on these road types has been removed to some extent while there is still substantial overlap on urban main roads and local roads and only by correcting for overlap individually for each road type, the travel time parameters remained reasonable.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SimDist</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel time on MW</td>
<td>-0.88</td>
<td>-0.87</td>
</tr>
<tr>
<td>Travel time on EU</td>
<td>-1.02</td>
<td>-1.16</td>
</tr>
<tr>
<td>Travel time on UM</td>
<td>-0.88</td>
<td>-0.97</td>
</tr>
<tr>
<td>Travel time on LR</td>
<td>-1.13</td>
<td>-1.21</td>
</tr>
<tr>
<td>Road type perc. MW</td>
<td>0.64 *</td>
<td>0.02 *</td>
</tr>
<tr>
<td>Road type perc. UM</td>
<td>0.94 *</td>
<td>0.15 *</td>
</tr>
<tr>
<td>Road type perc. LR</td>
<td>-1.45</td>
<td>-3.14</td>
</tr>
<tr>
<td>Beta PSRT1 MW</td>
<td>0.13 *</td>
<td>0.85 *</td>
</tr>
<tr>
<td>Lambda PSRT1 MW</td>
<td>-0.86 *</td>
<td>0.58 *</td>
</tr>
<tr>
<td>Beta PSRT1 EU</td>
<td>6.27 *</td>
<td>2.57 *</td>
</tr>
<tr>
<td>Lambda PSRT1 EU</td>
<td>6.27 *</td>
<td>2.56 *</td>
</tr>
<tr>
<td>Beta PSRT1 UM</td>
<td>10.07</td>
<td>6.10</td>
</tr>
<tr>
<td>Lambda PSRT1 UM</td>
<td>4.44</td>
<td>2.69</td>
</tr>
<tr>
<td>Beta PSRT1 LR</td>
<td>-0.03</td>
<td>3.77</td>
</tr>
<tr>
<td>Lambda PSRT1 LR</td>
<td>-1.22</td>
<td>2.78</td>
</tr>
<tr>
<td>Initial LL</td>
<td>-4486</td>
<td>-4853</td>
</tr>
<tr>
<td>Final LL</td>
<td>-2765</td>
<td>-3490</td>
</tr>
<tr>
<td>Adjusted Rho square</td>
<td>0.38</td>
<td>0.28</td>
</tr>
</tbody>
</table>

(*) parameter not significant at 95% confidence level

MW = motorway, EU = extra-urban road, UM = urban main road, LR = local road

Road type percentage are estimated with extra-urban main roads as reference category

6 Conclusion and Outlook

The aim of the research effort described in this paper is two answer two questions: What is the most suitable choice set for and what is the best way to account for similarities in a car route choice model derived from
GPS observations and a very high resolution network? Concerning the choice set it can be concluded that there is a substantial difference between model parameters estimated for a small choice set directly derived from the choice set generation algorithm and those estimated for a small choice set for which first a large set of alternatives has been generated and then reduced to a smaller choice set. Since this effect even occurs if the reduction was purely random, this might be caused by missing relevant routes in the small choice set directly derived from the choice set generation. Thus, it is advisable to first generate a large set of alternatives to ensure that all relevant routes are found and then reduce the choice set appropriately. This has, however, to be weighted against the computation costs for the generation of a large set of alternatives, especially on a high resolution network. As an example, the generation of the 1500 choice sets for this study with the BFS-LE algorithm lasted 6.6 hours for 20 alternatives, 52 hours for 60 alternatives, and 8.17 days for 100 alternatives whereas the corresponding run times for the SCSG algorithm were 23.98 days, 51.87 days, and 61.83 days. Though the actual computation time was considerably reduced by splitting up the list of OD pairs and running several jobs in parallel, it is still questionable if the SCSG algorithm is appropriate for the choice set generation in very high resolution data given that the resulting choice sets were also characterised by more overlap and a much narrower overlap distribution, even after employing reduction procedures. For the BFS-LE algorithm, on the other hand, the run times are still in an area where the generation of a large set of alternatives can be achieved in an acceptable time.

Regarding the reduction procedure, an important lesson to be learned from these experiments is that less overlap is not necessarily always better, especially if the reduction procedure does not remove the overlap uniformly over all route types. This can lead to unstable estimation results that are difficult to interpret. Thus, the analyst should either aim for a wide distribution of overlap levels or try to reduce the overlap in a meaningful way. The best way to do this is either employing the similarity distribution-based or the rule-based reduction procedure. However, for both procedures additional parameter testing is required. The parameters to be tested for the similarity distribution-based reduction are the number of similarity classes and the target choice set size. The rule-based reduction has no predefined target choice set size. Instead, a systematic testing of the travel time, distance and overlap thresholds is necessary with particular attention to the effects of the individual thresholds.

The best similarity treatment in this study was the road type specific Path Size with the formulation established in Equation 10, the PSRT1. It outperformed the other models consistently throughout all choice sets and transformations. Moreover, it allows to account for the route overlap on each road types individually. This is a benefit because, apparently, the impact of the route overlap on the utility of an alternative does vary with the
road type. While overlap on motorways is rather negligible, also overlap on extra-urban roads has a small impact compared to overlap on local roads or urban main roads. Concerning the transformation, the BoxCox transformation consistently outperformed the other transformations, even accounting for the additional parameter.

7 Acknowledgements

The authors would like to thank the Swiss National Science Foundation for the funding of this work and Stefan Muff (at that time Endoxon AG) for providing the data. Moreover, we would like to thank Michael Balmer for providing the MATSim run that enabled the estimation of models based on time-of-day dependent travel times.

References


Schüssler, N. and K. W. Axhausen (forthcoming) Processing GPS raw data without additional information, *Transportation Research Record*.

Swait, J. (2001) Choice set generation within the generalized extreme
value family of discrete choice models, *Transportation Research Part
B: Methodological*, 35 (7) 643–666.

Swait, J. and M. E. Ben-Akiva (1987) Incorporating random constraints
in discrete models of choice set generation, *Transportation Research Part
B: Methodological*, 21 (2) 91–102.

http://www.bfs.admin.ch/bfs/portal/de/index/
infothek/erhebungen___quellen/blank/blank/vz/
uebersicht.html

van der Zijpp, N. J. and S. Fiorenzo-Catalano (2005) Path enumeration by
finding the constrained k-shortest paths, *Transportation Research Part

demand modeling: Lessons learned from the New York best practice

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