Developing advanced route choice models for heavy goods vehicles using GPS data

Stephane Hess* Mohammed Quddus† Nadine Rieser-Schüssler‡ Andrew Daly§

January 20, 2015

Abstract

This paper presents a novel application in route choice modelling using Global Positioning System (GPS) data, focussing on heavy goods vehicles which typically make longer journeys with decisions potentially underpinned by different priorities from those used by car drivers. The scope of the study is larger than many previous ones, using the entire road network of England. Making use of the error components model put forward for route choice by Frejinger & Bierlaire (2007), the work reveals low elasticities in response to changes in travel time, reflecting the limited opportunity for avoiding specific roads on long distance journeys by heavy goods vehicles.

Keywords: *route choice; GPS data; heavy goods vehicles; error components*

1 Introduction

The modelling of route choice behaviour has been a key component of research in transport for several decades. While the number of studies is lower than say mode choice or possibly even destination choice, this is in part a result of the high demands it imposes in terms of data requirements and model complexity. A major problem for route choice modelling has always been the need to capture appropriate data. Despite some work on representing route choices in hypothetical (stated) choice surveys (see e.g. Parkany et al., 2006), the presentation of routes in such surveys requires extensive simplification which can reduce realism. On the other hand, relying on drivers to recall their chosen route in detail after a journey is also problematic as discussed for example by Bierlaire and Frejinger (2008). Since a link by link description of entire routes is extremely burdensome for respondents, usually only path segments (e.g. Ramming, 2002) or intermediate locations (e.g. Ohnmacht and Kowald, 2014) are collected and the missing elements of the chosen routes have to be imputed. Thus, in addition to the errors inherent in self-reporting studies (see e.g. Bricka et al., 2012; Wolf et al., 2003) the imputation approach that is chosen further biases the modelling results. In response to this, there has in recent years been a growing uptake of automatic data capture approaches, primarily relying on GPS data (see e.g. Wolf et al., 2004; Li et al., 2005; Bierlaire and Frejinger, 2008).

The present paper adds to the body of work on modelling route choice behaviour with GPS data, and in particular is one of only a few studies (especially at a national level) looking at route choices

*Institute for Transport Studies, University of Leeds, s.hess@its.leeds.ac.uk
†Civil and Building Engineering, Loughborough University, m.a.quddus@lboro.ac.uk
‡Institute for Transport Planning and Systems (IVT), ETH Zürich, nadine.rieser@ivt.baug.ethz.ch
§Institute for Transport Studies, University of Leeds, and RAND Europe, daly@rand.org

for heavy goods vehicles (as opposed to cars), which typically make longer journeys (up to 500km in the present dataset) and where the decision making is potentially underpinned by different priorities from those used by car drivers. Road based freight transport is an important component of the supply chain in most countries, but also contributes extensively to congestion. A better understanding of the route choices made is thus of great interest for transport planning. In this context, it should already be acknowledged that the actual route choices in the case of heavy goods vehicles (HGVs) may in fact not be made by the driver but by the company. This again potentially leads to very different behaviour from that of car drivers.

The main novelty of our study comes in the scale of the network we use. While many previous route choice studies making use of advanced discrete choice models have been conducted at the level of small networks of metropolitan areas or countries, the present application uses the entire road network of England, which contains some 4.5 million individual links, and a sample of over $20,000$ observed journeys. This compares to previous efforts using discrete choice models for HGV route choice at a national level, where as an example, Quattrone and Vitetta (2011) rely on just 52 chosen routes with only $16,029$ links in their network. Finally, our study makes use of the advanced modelling framework of Frejinger and Bierlaire (2007) which had to date not been applied to a problem of the size studied here.

The remainder of this paper is organised as follows. Section 2 discusses initial data processing and cleaning, before we turn our attention to choice set generation work in Section 3. Section 4 presents our modelling work, with conclusions in Section 5.

2 Data processing, cleaning and conversion

A key advantages of GPS data is the lack of requirement for respondent recall, with choices captured with a high level of precision (individual road links) and the method of capturing the data having little or no impact on driver behaviour. On the other hand, the use of such data leads to other complexities, notably in processing (see e.g. Wolf et al., 2004; Stopher et al., 2005; Schssler, 2010; Moiseeva et al., 2010; Marchal et al., 2011), due to the sheer amount of information available but also in relation to mapping the data onto a road network (map matching) (e.g. Pyo et al., 2001; Marchal et al., 2005; Schssler, 2010; Dalumpines and Scott, 2011) and addressing data errors. This section describes some of these steps for the present study, looking at how the data was prepared and cleaned with a view to making it suitable for choice set generation and the estimation of route choice models.

2.1 Approach for constructing trips

The original data provided by the UK Department for Transport for the present study was a journey database of already processed GPS data which had been matched onto a road network. It represented the movements of 709 HGVs for one month (April 2010), with a total of 8,656,534 observations in the database relating to individual road links. We added additional network data with characteristics for all links as well as the location of service stations and rest locations for HGV drivers.

The observations were grouped into journeys according to when the vehicle's ignition is turned on or off. This suggests that some of the journey ends are for rest breaks or to buy fuel at service stations, rather than real ends of a trip where goods are picked up or dropped off. It is therefore essential to combine journey segments that form part of a single delivery (i.e. a trip) but which have been separated into individual journeys in the data due to breaks or refuelling stops or other reasons (e.g. errors in GPS data). For the purpose of this study, a *trip* is therefore defined as a combination of *journeys* where the ends are the real origin or destination of the goods and/or vehicle, thus grouping subsequent journeys

(i.e. links) that include breaks at fuel stations or truck-stops. This is based on the assumption that few if any of the deliveries are to such service stations or rest breaks. The task of processing the data is made more complicated by the fact that the first and last 500m of each journey are missing for the purpose of preserving customer confidentiality.

There are a total of 68,403 unique journeys (not necessarily by unique HGVs) with an average of 96.5 journeys per vehicle, where on a single journey, a vehicle travels on average through 126 links. The journey data indicates that HGVs normally travel on the major roads, with 90% of the road segments coming from motorways, A roads and B roads[1], with only 5% of the segments for the latter. Our approach for translating journeys into trips involves three distinct processes:

- Identification of a break in a journey
- Calculation of the duration of a break
- Identification of whether a break is a real stop

The first of this steps is rather straightforward. The Trafficmaster dataset was sorted by Vehicle ID, then Journey ID, and then Link Start Time. The resulting order of rows in the dataset should then imply vehicles' trajectories by Vehicle ID. A change in vehicle ID indicates the start of a new trip as this implies the trajectory of another vehicle. A change in journey IDs without a change in vehicle IDs indicates a break within journeys made by the same vehicle.

Once a break is identified in process 1, the next step is to calculate the duration of the break. This is important to make the decision whether a break can be considered to be at a service station (or a truck-stop) or a real break (i.e. a delivery point, a pick up point or the vehicle's depot). The following criteria are utilised:

- A delivery or pick-up task should take more than 2 minutes (a delivery or pick up activity should take at least 1 minute and the other minute is assumed for the travelling time of the missing 500m distance)[2].
- If a break is equal to or more than 2 minutes but less than 15 minutes then it is assumed to be a delivery/pick-up as the minimum time for a break at service stations (or rest breaks) should be 15 minutes (VOSA, 2009).
- If a break is equal to or more than 15 minutes but less than 45 minutes, an algorithm was used to see whether the location of the break is near to a service station (or a truck-stop), using a geo-coded database of service stations/truck-stops in England and the link ID of the road segment on which the vehicle was travelling just before ignition was turned off. If the break has occurred near to a service station (or a rest break) then it is assumed that both journeys should be the part of same trip[3].
- If a break is equal to or more than 45 minutes then a new trip is considered. This is because the daily maximum driving time is 9 hours with a maximum break of 45 minutes (VOSA, 2009).

[1] A roads are primary routes not comprising any motorway sections, while B roads are numbered local routes.

[2] From the TrafficMaster records, slow links due to traffic congestion, incidents or other reasons (i.e. driver stops to check directions or take a phone call, stoppage at major crossings) were identified using link length and speed. Those records were discarded from the analysis as we were not certain whether those were due to a pick up or a delivery or other purposes. This should avoid confounding with deliveries.

[3] Drivers of HGVs do not take a short break that is less than 15 minutes as VOSA (2009) states "Breaks of less than 15 minutes will not contribute towards a qualifying break, but neither will they be counted as duty or driving time." This regulation is being controlled through the tachographs. If a break is calculated to be more 15 minutes and there is no service/petrol stations around then this break indicates a trip end (i.e. a delivery or pick up; see Figure 2).

Applying the approaches discussed above, 46,774 trips were created from the Trafficmaster journey data using Matlab. This is much lower than the total number of journeys (i.e. 68,403) in the Trafficmaster data suggesting that the applied approaches have successfully combined consecutive journeys to form trips. A new dataset describing the details of the trips was then produced, including vehicle and trip IDs, trip distance and duration, as well as the share of the trip using different road types. On average, there are 1.6 journeys per trip, with an average trip time of 58 minutes, an average trip distance of 59 km (with a 95th percentile of 188km), and the average number of links per trip being 200. Validation of the developed algorithm was carried out empirically as there was lack of reference (true) trip data. Using the stratified sampling technique, a total of 245 trips were identified and manually checked using a GIS tool. The results indicate that 92.1% of the trips were found to be correctly matched with the manually constructed trips. Most of the inaccurate trips were 'short distance trips' that were not considered in the subsequent analysis.

2.2 Data conversion and cleaning

Prior to use in choice set generation and subsequent analysis in modelling, both the road network and the data on the 46,774 actual trips had to be processed further.

Most of the issues encountered during the data conversion and cleaning phase related to the network. In the end, a three-step process was necessary to convert and clean the data. The first step was the basic network conversion, primarily concerned with converting bi-directional links into one link per admissible direction in the original network. In the second step, missing node coordinates were imputed (for about 10% of nodes), employing a mass-spring system (Fox and Mahanty, 1970) that takes into account all the available information, such as the network topology, the coordinates of the nodes connected to the node with missing coordinates and the attributes of those links, particularly the length. Finally, in the third step, additional cleaning procedures which are part of the network conversion and cleaning tools of the transport simulation tool MATSim (see MATSim, 2013, for further details) were used to make the final network meet the requirements of the choice set generation. The first of these approaches establishes strong network connectivity, i.e. that every node can be reached from every other node. The second approach ensures that there are no duplicate links in the network. Finally, we also removed pedestrian only roads.

Travel time or speed information per link were required but not present in the data, and speed assumptions per road type were therefore made as summarised in Table 1, using official values provided by the UK Department for Transport.

The final step in preparing the data for choice set generation and route choice modelling is concerned with route conversion. The route conversion first translates the link IDs in the original format into the link IDs used in the new unidirectional network. Next, it checks the trips for topological consistency, i.e. it verifies that the routes are feasible and continuous. This identified a large number of problems, which are unfortunately to be expected in a network of this size. The first issue was that trips had gaps, i.e. links missing, and since for some of the larger gaps there was more than one route alternative available to close the gap, any assumption made by the research team – particularly without access to the original GPS data – could have led to a bias in the modelling results. Closing the gaps automatically was difficult. The second most common issue arose when the order of the links was wrong. Since this was often combined with a few associated links missing completely, an automatic correction was again not possible. The last two issues were network related. On the one hand, the trips contained links that were missing in the network. On the other hand, there were one-directional links where the trips used one direction although the network contained only the other direction. This case was detected and corrected automatically by the route conversion algorithm. A large number of errors in the data meant that overall, the route conversion and topological consistency check was successful for 22,291

Table 1: Speed assumption per road type

Main road type	Subtype	Speed [mph]	Speed [m/sec]
Motorway	Regular motorway lane	55	24.6
	Slip Road	35	15. 6
A road	Dual Carriageway	42	18.8
	Single Carriageway	32	14.3
	Slip Road	25	11.2
	Other†	25	11.2
B road	Dual Carriageway	29	13.0
	Single Carriageway	25	11.2
	Slip Road	20	8. 9
	Other†	20	8. 9
Other		20	8. 9

†: Primarily roundabouts and traffic island links

trips, with the remaining 24,483 trips having to be removed from the data. Nevertheless, this left us with a sufficiently large sample to continue the analysis.

3 Generation of the choice sets of potential routes

Even in modestly size networks, the number of possible routes between a given origin and destination is very large. Many of those routes are substantially longer than the shortest path, and unlikely to be considered by decision makers, or to add much to the estimation of choice models, other than increasing computational cost. Ideally, the set of routes used in model estimation should contain all relevant and no irrelevant alternatives. To achieve this, two different approaches can be employed. If the universal choice set, i.e. all possible routes between an origin and destination pair, is known, the analyst can model the membership of an alternative to the individual choice set (e.g. Swait, 2001; Morikawa, 1996). However, in high-resolution networks it is not possible to enumerate all possible route alternatives. Instead, the analyst has to employ heuristics that extract the route alternatives from the network. The aim is to derive as exhaustive a route set as possible in order to ensure that all relevant alternatives are detected. The main challenge is the high level of spatial detail, which raises the requirements in terms of computation time but also regarding the choice set composition. As several authors (e.g. Prato and Bekhor, 2007; Bekhor et al., 2006; Bliemer and Bovy, 2008) have demonstrated, the size and composition of the choice set strongly influence the outcome of model estimation. Misspecifications of the choice set lead to biased parameter estimates and choice probabilities. As Bliemer and Bovy (2008) showed, this is especially true when there is correlation between alternatives, which is inherently the case due to route overlap.

Several route extraction approaches for car and public transport route choice problems have been proposed in recent years (e.g. Frejinger et al., 2009; Prato and Bekhor, 2006; Hoogendoorn-Lanser et al., 2006; Nielsen, 2000; Ben-Akiva et al., 1984). Rieser-Schssler et al. (2012) tested a number of them and found that only approaches based on repeated least cost path search are suitable for extracting routes from high-resolution networks. Prevalent approaches for route set generation with repeated least cost path search are the (doubly) stochastic choice set generation (e.g. Ramming, 2002; Dugge, 2006;

Bliemer et al., 2007; Bovy and Fiorenzo-Catalano, 2007), link elimination (e.g. Azevedo et al., 1993; Prato and Bekhor, 2007), link penalty (de la Barra et al., 1993) and path labelling (Ben-Akiva et al., 1984). Most of these approaches were developed for lower resolution networks but can be used on high-resolution network. However, the path labelling approach requires a certain variety of link attributes that are preferably uncorrelated with the main criteria of distance and travel time such as the number of traffic lights, information regarding the land use around the link (e.g. commercial area or scenic country side). For the labelling approach to work properly, these attributes have to be available for each link in the network. This is rarely possible for large high-resolution networks in real life applications.

3.1 Methodology

The choice set generation approach used in this paper was developed by Rieser-Schssler et al. (2012) specifically for route generation in high-resolution networks and successfully applied to different bike and private car route choice problems (see e.g. Halldrsdttir et al., 2014; Schssler, 2010; Menghini et al., 2010). The ability to apply this non-behavioural approach easily across different context and countries is a clear advantage, with only an application-specific cost function being needed for each study. The method employs a link elimination approach which means that links of the current least cost path are eliminated before the next least cost path is searched. This is repeated until the required number of routes is found. Some link elimination approaches ensure that the k-least cost paths are found (Lawler, 1976; van der Zijpp and Fiorenzo-Catalano, 2005), while others only accept paths within constraints such as maximum amount of overlap with other paths or a maximum detour time (van der Zijpp and Fiorenzo-Catalano, 2005). The order in which the links are eliminated can be random, duplicating the order of appearance in the route (Azevedo et al., 1993) or controlled by criteria (Prato and Bekhor, 2007).

In the algorithm used in this paper, a *Breadth First Search* approach is employed and combined with a *topologically equivalent network reduction* to ensure high diversity between the routes as well as computational feasibility for large-scale problems. The goal is to find a maximum number of feasible and low cost routes in the shortest amount of time possible. Here, a *feasible* route is continuous, contains no loops and is low in travel cost. *Travel cost*, in this application, is defined by the following cost function:

$$C_i = \sum_{\forall l \epsilon r} \beta_{tt} * tt_l + \beta_{BRoad} * tt_l * \delta_{l,BRoad} + \beta_{otherR} * tt_l * \delta_{l,otherR} \quad (1)$$

where C_i is the cost of route i consisting of links l, tt_l is the free flow travel time on link l, β_{tt} is the cost parameter for travel time, β_{BRoad} is the penalty for travelling on a B road, $\delta_{l,BRoad}$ is a binary variable that equals one if link l is part of a B road and zero otherwise, β_{otherR} is the penalty for travelling on an "other road" and $\delta_{l,otherR}$ is a binary variable that equals one if link l is part of an "other road" and zero otherwise - where *other* refers to non-A, non-B and non-motorway roads. It is important to note that only the ratio between the β-parameters influences the outcome; the absolute values are not decisive. After extensive empirical testing, the following values were assumed: $\beta_{tt} = 1$, $\beta_{BRoad} = 1$ and $\beta_{otherR} = 1.5$. The terms β_{BRoad} and β_{otherR} are penalty terms for non-A and non-motorway links on top of the usual travel time sensitivity. These penalties reflect other sources of inconvenience occurring on minor roads that are not captured by the reduced speeds such as the presence of traffic lights, pedestrian crossings, park search traffic etc. The assumed values resulted in the most realistic route sets based on comparing the share of road types with that observed in the data.

As noted above, the Breadth First Search on Link Elimination (BFS-LE) algorithm calculates repeated least cost paths of a given origin-destination (OD) pair for a given network to find a set of route alternatives for the OD pair. The least cost paths are calculated with the so-called *A-Star Landmarks*

routing algorithm presented in Lefebvre and Balmer (2007). The algorithm follows the same principles as the well-known Dijkstra algorithm (Dijkstra, 1959) but its computational performance is at least one order of magnitude better by using landmarks for estimating the remaining travel time to the destination at each node. The landmarks are network specific and derived in a preprocessing step during the initialisation of the choice set generation. For more information see Lefebvre and Balmer (2007).

3.1.1 Basic algorithm

The basic idea behind the BFS-LE algorithm is illustrated in Figure 1. The algorithm calculates the shortest path between the OD pair in the original network, adds this path to the set of route alternatives and then removes the links of this shortest path step by step and calculates a new shortest path for this new network. If the new shortest path is not yet part of the set of route alternatives, it is added to the set (see Figure 1 where the routes assigned to S are marked with a grey background). To keep track of the eliminated links and the resulting networks and to organise the order of the link elimination, the algorithm uses the tree structure shown in Figure 1. Each node of the tree consists of a network. The original network is the root of the tree and the depth levels d correspond to the number of links that were eliminated, i.e. in depth level $d = 1$, one link has been eliminated, in depth level $d = 2$, two links have been eliminated, etc. Each network is *unique*, i.e. there is no other network containing the exact same set of links, and it contains at least one *valid connection* for the OD pair in question. These conditions are always checked by the algorithm before adding a network to the tree as illustrated in Figure 1 (b) and (c).

The construction and processing of the tree can be done in two ways: Breadth First or Depth First. In a Breadth First approach, the algorithm first finishes a depth level before moving on to the next depth level whereas in the Depth First approach, the algorithm first traverses one branch up to the final level before moving on to the next branch at the first depth level. For the purpose of route choice set generation, the Breadth First approach is more appropriate because it allows quicker exploration of route alternatives with deviations from the shortest path at different sections of the route, resulting in more diverse routes more quickly. The algorithm ends when there are no more valid paths for the OD pair, i.e. the tree is completely processed, or when the number of alternatives n requested by the analyst is reached. However, since the composition of the resulting choice set depends in this case on the processing order, it is necessary to complete the whole tree at the current depth level d before checking if the required n has been reached. If more paths than specified by the analyst have been determined at depth level d, a random subset of these paths is drawn to remove the processing order dependency.

The algorithm searches for new routes until the preset route set size is reached, no further routes exist or the algorithm reaches a time threshold specified by the analyst. In this project, the target choice set size was set to 15 routes and the time threshold was set to 30 seconds - these values were obtained after a number of trial runs aimed at finding settings that minimised impacts on model results. The results of the modelling work remained very consistent when the choice set size was increased beyond 15 - we see this as an indication of a lower number of realistic alternative routes on longer journeys and also for heavy goods vehicles than is perhaps the case in intra-urban car driver route choices, more often studied in the past. We retained a final setting of 15 routes per O-D as this also makes the modelling analysis more computationally tractable. If the chosen route was not reproduced by the choice set generation approach, it was added to the choice set. The minimum number of alternatives in a choice set is thus one route, while the maximum number with these parameter settings is 16 routes if the chosen route had to be added and 15 if it was reproduced. Ideally, the majority of choice sets would have 15 or 16 routes. We will return to this point in Section A. The entire choice set generation process for the whole sample took just under 36 hours.

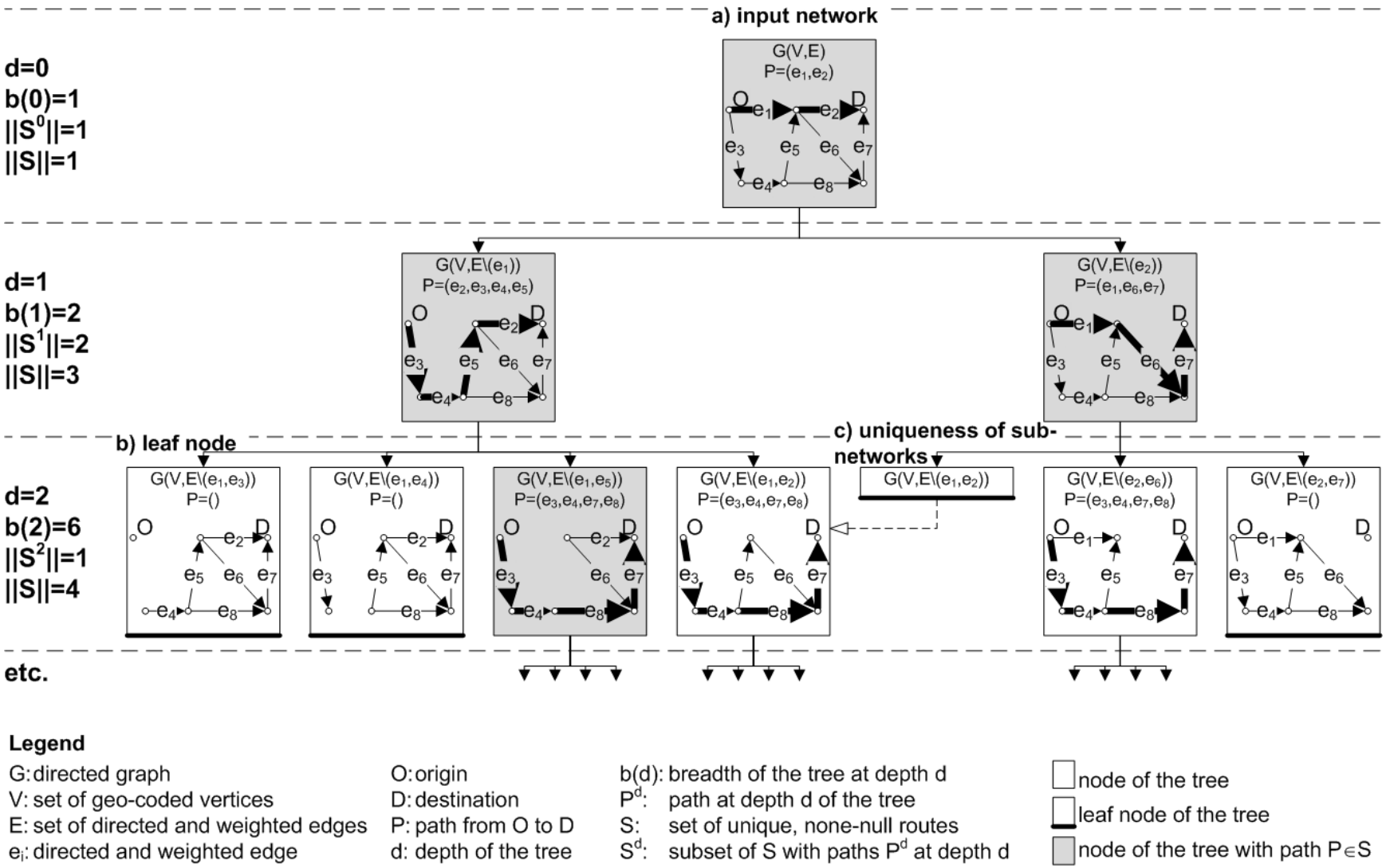


Figure 1: Basic BFS-LE tree

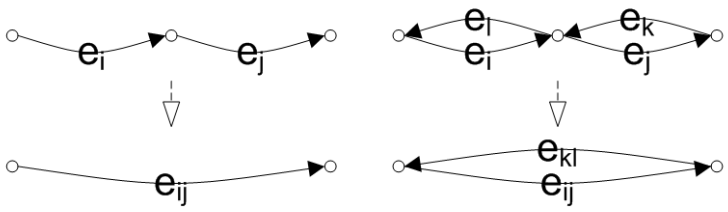


Figure 2: Examples of merging edges for producing a topologically equivalent network

3.1.2 Topologically equivalent network reduction

The topologically equivalent network reduction step addresses the issue that there are nodes in the network that do not model junctions, intersections or dead ends but are still important for the routing because they represent road attribute changes such as speed limits, gradients or the number of lanes. Ignoring them in the link elimination step markedly reduces the complexity of the tree compared to the basic BFS-LE approach. Therefore, a topologically equivalent network reduction, which is illustrated in Figure 2 is performed that creates a reduced network from the original network in which all nodes that do not represent junctions, intersections or dead ends are removed and their incident links are merged in each direction. Then, the link elimination and tree development is performed on the reduced network, allowing the algorithm to remove entire street segments at once that would have resulted in the same new shortest path anyway. In order to ensure that this performance optimisation does not change the resulting route choice set, the subsequent shortest path calculation is still performed on the equivalent non-reduced network.

3.2 Attribute calculation for the derived routes

For use in the subsequent choice models, the routes in each choice set were described by a set of attributes which might have influenced the route choice. These included travel time, cost, and the mix of road types used on a given route. A number of calculations were necessary to obtain the values for these attributes, and these calculations are described in this section.

The travel time for a given route was calculated as the sum of the travel times for each of the links used by that route, using the speed assumptions from Table 1. For the distance and travel time per road type, only the main road types (motorways, A roads, B roads and other roads) as specified in Table 1 were used. Here, it should be noted that calculated journey times were used for all routes, including the chosen route for which observed journey time was available. Mixing observed journey time for the chosen route with calculated times for the unchosen routes could have led to biased results if the journey time for the chosen route was higher than the calculated one due to e.g. accidents en route.

We also added information on the extent that a given route uses roads of the *Strategic Road Network*, made up of motorways and key A-roads, where the overall distance and travel time on each road is calculated. A more detailed evaluation is performed for the M25 ring road around London, where, for a better distinction of which part of the M25 is used, the ring is divided into four quadrants: a north-west, north-east, south-west and south-east one.

Two cost attributes were calculated: *fuel cost* and *other cost*, with both calculations following official WebTAG guidance (DfT, 2009). For fuel cost, the following equation was used:

$$\begin{aligned} fuelCost_i = \sum_{a \in \Gamma_i} & (0.6(149.948/v_a + 24.929 - 0.36259 v_a + 0.003110 v_a^2) \\ & +0.4(344.145/v_a + 40.028 - 0.47118 v_a + 0.003646 v_a^2)) d_a \end{aligned} \tag{2}$$

where Γ_i is the set of all links of route i, v_a is the free flow speed of link a in km/h and d_a is the length of the link in km. The other costs include all non-fuel related costs and were calculated using the formula:

$$otherCost_i = \sum_{a \in \Gamma_i} (0.6(6.714 + 263.817/v_a) + 0.4(13.061 + 508.525/v_a)) \quad (3)$$

4 Modelling analysis

A key interest in route choice modelling is the representation of the competition and correlation between individual routes in a choice set. This recognises that route overlap affects the probabilities of choosing a given option, and also the substitution patterns between routes. The main issue is overestimation of choice probabilities for routes with a high overlap with other routes, as shown for example by Cascetta et al. (1996), Ben-Akiva and Bierlaire (1999) or Ramming (2002). The typical approach for capturing correlation between alternatives in choice modelling is the specification of a nesting structure (see e.g. Vovsha and Bekhor, 1998), but this is not a viable approach in this context where the number of possible types and extents of route overlaps is very large. Rather, researchers have developed a number of techniques that allow a model to give a deterministic account of the level of overlap or independence of given routes and adjust their probabilities as a result (e.g. Cascetta et al., 1996; Ben-Akiva and Bierlaire, 1999; Hoogendoorn-Lanser and Bovy, 2007). A more recent development by Frejinger and Bierlaire (2007) has looked at the potential for correlation caused by the specific roads used on given routes, rather than the actual sharing of links by different routes. This would for example mean that there could be correlation between two routes both using a section of the M25 for going around London, without that specific section necessarily being the same one for the two routes. The longer distances travelled in many of our observations increase the scope for studying strategic (long distance) route choice and this allowed us to make full use of this addition to the toolkit of route choice models. Our analysis makes use of both types of approaches, starting with a simple overlap measure in our preliminary models below.

4.1 Preliminary models and filtering of choice sets

From the initial set of $22,291$ observations, we removed 776 O-D pairs where only a single route was found by the choice set generation algorithm. Our analysis initially made use of simple Multinomial Logit (MNL) models (see e.g. Train, 2009). In these models, we specified the deterministic utility of route i (out of I, with $I \leq 16$) for observation n (out of $N = 21,115$) as:

$$V_{n,i} = \beta_{l-cost} \ln(C_{n,i}) + \beta_{l-time} \ln(T_{n,i}) + \beta_{pathSize} pathSize_{n,i}, \quad (4)$$

where $C_{n,i}$ and $T_{n,i}$ give the cost and time respectively of alternative i for person n. The use of a log-transform was motivated by early evidence of strong non-linearity in response, manifested through decreasing marginal sensitivity to increases in travel time and travel cost. This specification thus allows for cost damping, as discussed in detail by Daly (2010). Different possibilities arise in this context, including a combination of a linear and log specification to obtain intermediate levels of damping. Results from early models suggested a level of damping sufficiently close to a log transform to allow us to take this specification forward. It should also be noted that the use of damping on both time and cost ruled out the estimation of separate sensitivities for individual elements of a route as this would

have ignored the overall damping[4] in the absence of a more complex specification which would not have been practical in a large scale analysis.

The final term is the linear sensitivity to the path size component. The *path size* gives an indication of the similarity of a route with the other routes in the choice set. It was developed by Ben-Akiva and Bierlaire (1999) and its values range between zero and one[5]. A distinct route, i.e. a route with no overlaps with other routes, has a 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. Mathematically, the path size is calculated as follows:

$$pathSize_{n,i} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i} \right) \frac{1}{\sum_{j \in C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}} \tag{5}$$

where Γ_i is the set of all links of route i, l_a is the length of link a, and L_i the length of route i. The term δ_{aj} is 1 if link a is on route i and 0 otherwise. The formulation additionally accounts for the relative ratio between the length of the shortest route $L^*_{C_n}$ in C_n using link a and the length of each route j using link a. The underlying idea is that, all else being equal, a route with little overlap with other routes (and thus a higher path size value) has a higher probability of being chosen.

Data on reliability and level of congestion were not available, and we acknowledge that this is a potential shortcoming of the models estimated in the work. Future work should attempt to include such information, for example also looking at different sensitivities in peak and off-peak conditions, and rural *vs* urban locations. It is important to acknowledge that specific road characteristics could also influence route choice, with examples including road gradient and characteristics such as roundabouts. Information on gradients was not available but can be expected to play only a minor role in England, especially on major roads. The reduced speed on roundabouts is captured through the assumptions in Table 1, where additional penalty terms associated with roundabouts and other features could not be estimated due to the strong correlation with travel time.

A number of other model specifications were also attempted, notably attributes such as the time and cost by road type in particular. However, this split leads to excessive correlation for example between time on given road types and travel cost (given the correlation between speed and cost), and would have prevented the separate estimation of a cost coefficient, which is required for value of time (VOT, the ratio of time and cost sensitivities) calculations. In addition, the simple model with time and cost attributes gave better fit than specifications excluding cost at the expense of road type specific time components, potentially also again due to allowing for overall damping, and this specification was thus retained.

While initial estimation results with this specification showed the expected negative sign for the time coefficient and a positive sign for the path size component, the estimate for the cost coefficient was positive, and standard errors were large for all parameters. A closer investigation of the data showed very high impacts on model results for a relatively small subset of individuals making what at first hand appear to be irrational choices. In particular, some chosen routes are several times longer than the shortest path. While this is potentially reasonable on very short journeys, the impact on model results is dramatic when it occurs for long journeys. Extreme cases included a situation where the chosen route covered a distance of 172km when the shortest path between the identified origin and destination was

[4] E.g. with $T = T_1 + T_2$, and $T_1 > 0$ and $T_2 > 0$, we have that $\ln(T) < \ln(T_1) + \ln(T_2)$.

[5] We decided to use the path size model as opposed to other possibilities such as C-Logit, basing our decision on the authoritative review conducted by Prato (2009). A comparison between path size and C-Logit (or other approaches) would in our opinion have added little value to our work, which was primarily interested in testing the added benefit of the error components specification in a large scale setting.

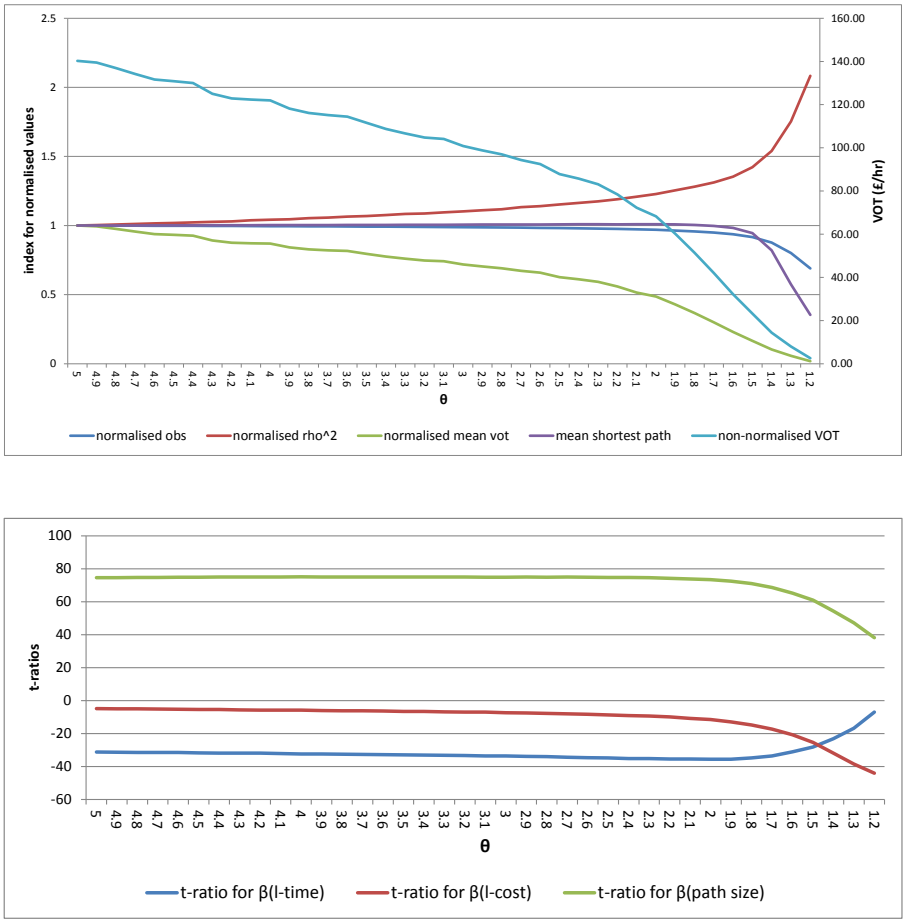


Figure 3: Impact of censoring on model results

a mere 629 metres. It is almost certain that observations of this type can be explained by problems in the data processing approach, with a failure to identify separate trips in case of very short breaks. This process was never expected to be completely foolproof and further data cleaning was thus required.

An observation by observation approach to data cleaning was not possible, and pragmatic approaches are needed with such a large sample. The specific approach used was to apply censoring by removing trips for which the chosen distance was a certain multiple of the minimum distance. To recognise that the permissible ratio between chosen distance and minimum distance should be a function of the minimum distance, i.e. a bigger relative difference is acceptable on shorter trips, we devised the following criterion after extensive testing:

$$\gamma_n = \frac{L_{chosen,n}}{L{*}_{C_n}} \frac{\ln\left(L{*}_{C_n}\right)}{8.95}, \tag{6}$$

where $L_{chosen,n}$ is the length of the chosen route and $L{*}_{C_n}$ is the length of the shortest route for observation n, while 8.95 is the sample mean for $\ln\left(L{*}_{C_n}\right)$.

Observations with $\gamma_n > \theta$ were removed from the data, where different versions for the limit θ were tested. In addition, we removed trips where the shortest path was shorter than 500 metres given unreliable results on such very short routes. We tested values of θ from 5 down to 1.2 in steps of 0.1,

meaning that 39 models were run. The findings from this censoring process are illustrated in Figure 3, which uses a mixture of normalised and non-normalised values as appropriate. We see that as the censoring increases (i.e. as θ reduces), the significance of the log-cost coefficient increases substantially, with modest increases for the significance of the log-time coefficient up until a censoring of $\theta = 2$ beyond which significance levels drop. For the path size term, the significance levels remain quite stable, up until $\theta = 2$. Relative model fit, measured through the adjusted ρ^2 term, increases throughout. As would be expected with this censoring approach, the shortest path distribution remains fairly stable, up until the point where a very low value for θ is used. The key impact is on the VOT findings. The VOT is obtained by taking the ratio of the partial derivatives of the utility function against time and cost, i.e.:

$$VOT_n = \frac{\partial V_{i*n}}{\partial T_{i*n}} \frac{\partial C_{i*n}}{\partial V_{i*n}} = \frac{\beta_{l-time}}{\beta_{l-cost}} \frac{C_{i*n}}{T_{i*n}}, \tag{7}$$

and thus depends on the values for the chosen alternative. We see very high initial values of time, which are a reflection of the earlier observation of problems with the cost coefficient. As the level of censoring increases, we see reduced values of time. The actual values remain high, and higher than standard WebTAG values, but the latter do not factor in the value of the freight, which in this case is unknown to us.

After extensive investigation, we decided on a censoring value of $\theta = 1.6$. This leads to a final sample size of $18,150$ observations, i.e. a loss of $2,965$ additional observations. The choice of this specific censoring point was motivated by this being the last point before major drops in both sample size and mean shortest path are incurred. It leads to a final mean VOT of $£32.19$/hr, which, as pointed out above is higher than WebTAG values of $£13$/hr, but which is deemed to still be reasonable as the latter does not include the time value of the load. It should also be acknowledged that while our speed assumptions follow official guidance, if these assumed speeds are too high, then this could translate into an overestimation of the time coefficient (given the resulting lower time attribute) and potentially lead to upwards bias in the value of time measures. An analysis of the characteristics of the choice sets in the full sample and after censoring is presented in Appendix A.

4.2 Advanced models

The models presented in Section 4.1 capture the overlap between competing routes through the inclusion of the path size factor. The estimate for $\beta_{pathSize}$ is positive and highly significant, showing that routes with reduced overlap, i.e. more independent routes, have a higher probability of being chosen.

In this section, we develop the models further to capture additional correlation between routes that result not from overlap between routes, i.e. the use of common links, but correlation as a result of making use of specific identified roads. While the incorporation of the path size factor gives an advantage to more *independent* routes, this additional component of the models ensures heightened competition and hence substitutability between routes that make use of the same roads. For this purpose, we make use of the specification of the error components logit (ECL) model put forward by Frejinger and Bierlaire (2007) for route choice modelling. As an illustration, let us use the very simple example where we factor in correlation between routes that use the M1 and also between routes that use the M25. The utility function from Equation 4 would then be rewritten as:

$$\begin{aligned} V_{n,i} = {} & \beta_{l-cost} \ln\left(C_{n,i}\right) + \beta_{l-time} \ln\left(T_{n,i}\right) + \beta_{pathSize} pathSize_{n,i} \\ & + \sigma_{M1} \sqrt{L_{n,i,M1}} \xi_{n1} + \sigma_{M25} \sqrt{L_{n,i,M25}} \xi_{n2}, \end{aligned} \tag{8}$$

where $L_{n,i,M1}$ and $L_{n,i,M25}$ give the distances covered by route i for observation n on the M1 and M25 respectively, and where ξ_{n1} and ξ_{n2} are independent standard Normal random variables, distributed

independently and identically across observations. The inclusion of these additional random components ensures that the model allows for correlation between two routes that both make use of say the M1, where the correlation increases as a function of the distance both routes cover on the M1, and where higher absolute estimates for σ_{M1} indicate higher correlation. Routes that are more correlated with one another are in stronger competition with one another, and also become better substitutes if for example one of the two routes becomes unavailable or becomes less attractive for example as a result of road works. With the example above, the covariance between the error terms for route i and j would be given by $\sigma^2_{M1}\sqrt{L_{n,i,M1}}\sqrt{L_{n,j,M1}} + \sigma^2_{M25}\sqrt{L_{n,i,M25}}\sqrt{L_{n,j,M25}}$.

It is important to note that with this model, the correlation between two routes as a result of both using say the M1 is a result solely of the distance both routes cover on the M1, and not of the specific sections of the M1 travelled on, i.e. whether they use the same subsections. This means that the model captures a phenomenon that is associated with the specific nature of the given routes while correlation as a result of sharing specific links is captured by the path size factor.

The estimation of the ECL structure is computationally very expensive given the need to use simulation to approximate the integral representing the choice probabilities in the model (cf. Train, 2009). For the purposes of the present project, we relied on the highly efficient package ALogit, which was the only feasible solution as any alternatives would, in the face of the size of the data and choiceset, have led to estimation times of several weeks. We made use of $1,000$ random draws per error component in estimation. While the $18,150$ observations come from only 709 HGVs, it is not possible to capture the correlation between journeys for the same HGV in ALogit, meaning that we have to accept some loss of efficiency in our estimates, which should however solely impact the standard errors.

Our initial specifications for the ECL model made use of a large number of error components, namely 14 error components associated with the most heavily used motorways in our data, 4 additional error components associated with the four separate quadrants of the M25, capturing quadrant specific correlation, 20 error components associated with the most heavily used A-roads in our data, and 3 higher level error components, associated with usage of the A-road network (i.e. correlation between all A-roads), the motorway network (i.e. correlation between all motorways) and the strategic road network (correlation between motorways and key A-roads).

The initial estimation process showed a number of insignificant error component terms and we gradually simplified the specification to obtain a final model with 15 error components, namely:

- 6 error components associated with specific A roads, namely the A12, A13, A14, A2, A46, A50;
- 1 general error component associated with all A-roads in the network;
- 6 error components associated with specific motorways, namely the M11, M3, M4, M40, M6, M60; and
- 2 error components associated with the two separate quadrants of the M25, hereafter referred to as M25NW (north-west) and M25SW (south-west)

Our final ECL model obtained a log-likelihood of $-31,360.49$, an improvement by 365.85 units over the MNL log-likelihood of $-31,726.35$, which is highly significant at the cost of 15 additional parameters. The estimates for the final ECL model are presented alongside those for the final MNL model in Table 2. We note that the inclusion of the ECL components leads to an increase in the impact of time and especially cost when compared to the path size formulation. We also observe a drop in significance levels for these three components, possibly as some of the effects are now captured by the error components, where this is especially likely in the case of $\beta_{pathSize}$. As regards the error component terms, it should be noted that the sign of the σ estimates is not relevant as the covariance between routes 1 and 2 for observation n caused by say the use of the $M11$ is given by $\sigma^2_{M11}\sqrt{L_{1n,M11}}\sqrt{L_{2n,M1}}$. The highest

Table 2: Estimation results for final MNL and ECL models

	MNL		ECL	
log-likelihood	$-31,726.35$		$-31,360.49$	
pars.	3		18	
	est.	t-rat.	est.	t-rat.
β_{l-cost}	-4.162	-20.7	-5.167	-18.72
β_{l-time}	-6.412	-31.1	-6.915	-25.20
$\beta_{pathSize}$	4.377	65.2	4.299	49.30
σ_{A12}			0.055	8.97
σ_{A13}			0.017	2.65
σ_{A14}			0.009	2.09
σ_{A2}			0.026	5.24
σ_{A46}			0.014	3.95
σ_{A50}			-0.015	-3.52
σ_{A}			-0.052	-24.68
σ_{M11}			0.023	3.66
σ_{M3}			0.020	2.64
σ_{M4}			0.007	1.03
σ_{M40}			0.017	3.07
σ_{M6}			0.008	2.26
σ_{M60}			-0.036	-4.42
σ_{M25NW}			-0.055	-5.90
σ_{M25SW}			0.020	3.67
VOT	MNL		ECL	
5%	24.18		21.01	
25%	27.45		23.84	
median	31.06		26.98	
mean	32.19		27.96	
75%	35.30		30.66	
95%	45.42		39.45	

significance level for any error component is observed for σ_A which is to be expected given the size of the A-road network. All remaining error components are significant at usual levels of confidence, with the exception of σ_{M4}. Finally, we see a reduction in VOT when moving from MNL to ECL by just over 13%, which, given the values in WebTAG, would suggest that the estimates from the ECL model are more reliable.

4.3 Forecasting analysis

As a final step in the analysis, we used the final MNL and ECL model in a forecasting example, looking at the changes in the total distance travelled on specific roads as a result of a number of hypothetical changes, namely:

- reductions and increases in average A-road speeds by 10%

- reductions and increase in average motorway speeds by 10%
- reductions and increases in average M1 speeds by 10%
- an increase by 20 minutes for the Dartford crossing
- an increase by 20 minutes for the M25 section close to Heathrow
- an increase by 20 minutes for travelling from the M25 to the M1
- a reduction in average A1 speeds by 10%

Table 3 summarises the results from the forecasting exercise. For ease of interpretation, the changes to the directly affected roads are shown in bold font. We first note small differences between the two models in terms of predicted total distances travelled in the base scenario, which are a result of the fact that no road type specific constants were estimated as these would have led to biased time and cost sensitivities given the strong correlations between road types and times and costs.

Looking first at the impact of changes in A road and motorway speeds, we see lower impacts on A roads than on motorways in the MNL models, while in the ECL models, the findings are essentially the same between the two types of roads. Additionally, we observe smaller overall changes in the ECL models than in the MNL models, suggesting that lower cross-road elasticities result from the specific correlation structure of that model. The impacts are more substantial when looking at changes to specific roads and the predicted changes in distance travelled on those roads. For example, a change in the speed on the M1 has a far stronger impact on the M1 distances than a change in general motorway speeds has on motorway distances. This is to be expected as changes affecting just one road make the shift to alternative roads far more likely than changes affecting an entire road type.

We next look at three specific scenarios in which we see increases by 20 minutes on specific key road segments, namely the Dartford crossing, the M25 segment close to Heathrow and the junction between the M25 and the M1. We observe bigger impacts on those specific segment, with e.g. the MNL predicting a 21.67% reduction on distance travelled jointly on the M25 SE and NE quadrants. The overall impact on the individual M25 SE and M25 NE segments is obviously much lower as this also includes journeys using just one of the two roads. Similar findings are observed across the other two scenarios. Once again though, the ECL elasticities are substantially lower than the MNL ones. The final forecasting scenario looks at a reduction in average speed on the A1, where we see a non-trivial shift to the A19; the impacts are once again lower in the ECL model than in the MNL model.

5 Summary and conclusions

This paper has described the various stages in a study using GPS data to model route choices for heavy goods vehicles. The study has a number of novel components in comparison with past work, with a focus on heavy goods vehicles, long distance journeys and a wide geographic area, in this case the entire road network of England. The data provided for the study had already been preprocessed and the way in which this was done was not amenable to direct analysis and a substantial amount of further processing and data cleaning was needed. This leads to the strong recommendation that future studies of this type start with the raw GPS data, avoiding the issues resulting from using pre-processed data.

In the modelling analysis, we then showed that the error components model of Frejinger and Bierlaire (2007) provides important gains in mathematical performance over the simple multinomial logit model, along with more realistic value of time findings. This is a result of capturing correlations between routes making use of the same key roads, where the sections used do not necessarily overlap - such correlation

Table 3: Forecasting results

	base distances (total metres)		count non-zero (out of 18,150)		A roads speed -10%		A roads speed +10%		M-way speeds -10%		M-way speeds +10%		M1 speed -10%	
	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL
dist on MW	165,492,378.48	162,989,352.92	6932	7082	1.66%	0.90%	-1.85%	-1.01%	**-1.73%**	**-0.97%**	**1.88%**	**1.05%**	**-0.22%**	**-0.14%**
dist on A roads	227,631,513.94	233,587,585.82	18000	18054	**-1.46%**	**-0.94%**	**1.64%**	**1.05%**	0.97%	0.51%	-1.03%	-0.55%	0.13%	0.07%
dist on B roads	1,399,633.80	1,408,073.83	15445	15526	2.90%	2.32%	-3.13%	-2.51%	0.24%	0.20%	-0.26%	-0.22%	0.02%	0.02%
dist on other roads	20,842,592.50	21,002,258.29	17860	17906	2.34%	2.08%	-2.44%	-2.19%	0.15%	0.15%	-0.16%	-0.16%	0.03%	0.03%
dist on SRN	278,267,683.08	278,969,144.03	15491	15694	0.32%	0.13%	-0.35%	-0.14%	-0.50%	-0.29%	0.55%	0.32%	-0.08%	-0.06%
dist on M25 SW	3,461,687.73	3,451,737.27	359	373	2.73%	1.03%	-2.99%	-1.14%	-2.78%	-1.10%	3.10%	1.21%	0.28%	0.05%
dist on M25 NW	3,319,681.69	3,360,515.89	336	336	4.27%	1.06%	-4.58%	-1.17%	-4.22%	-1.10%	4.78%	1.21%	-0.53%	-0.07%
dist on M25 SE	3,498,124.28	3,330,760.68	513	527	2.07%	1.28%	-2.47%	-1.47%	-2.23%	-1.35%	2.29%	1.42%	0.05%	0.01%
dist on M25 NE	4,148,326.42	3,912,928.96	482	485	3.77%	1.40%	-4.23%	-1.58%	-3.87%	-1.48%	4.19%	1.59%	0.07%	0.06%
dist on M25	14,427,820.13	14,055,942.80	1170	1196	3.23%	1.20%	-3.59%	-1.35%	-3.29%	-1.27%	3.60%	1.37%	-0.02%	0.01%
dist on M25 SE & NE	2,631,415.12	2,378,929.16	170	170	3.06%	1.29%	-3.50%	-1.45%	-3.09%	-1.29%	3.29%	1.38%	0.23%	0.06%
dist on M25 SW & NW	2,451,699.99	2,338,697.92	78	78	3.13%	0.67%	-3.31%	-0.72%	-2.97%	-0.63%	3.41%	0.72%	-0.39%	-0.10%
dist on M25 & M1	5,134,317.48	4,924,262.40	139	139	4.19%	1.00%	-4.50%	-1.08%	-4.17%	-1.01%	4.72%	1.14%	-3.83%	-0.89%
dist on M1	25,585,592.46	25,275,344.50	1388	1411	1.30%	0.79%	-1.46%	-0.88%	-1.38%	-0.85%	1.50%	0.93%	**-2.90%**	**-1.55%**
dist on A1	13,603,875.95	13,557,514.13	1294	1323	-0.12%	0.11%	0.13%	-0.13%	0.02%	-0.21%	-0.01%	0.22%	0.75%	0.41%
dist on A19	2,948,567.34	3,116,544.30	575	579	-2.11%	-1.35%	2.41%	1.52%	1.85%	1.10%	-1.95%	-1.18%	0.10%	0.07%

	base distances (total metres)		count non-zero (out of 18,150)		M1 speed +10%		Dartford crossing +20 min		Heathrow area +20 min		M25-M1 connection +20 min		A1 speed -10%	
	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL	MNL	ECL
dist on MW	165,492,378.48	162,989,352.92	6932	7082	**0.24%**	**0.15%**	**-0.27%**	**-0.11%**	**-0.18%**	**-0.04%**	**-0.38%**	**-0.10%**	0.01%	-0.02%
dist on A roads	227,631,513.94	233,587,585.82	18000	18054	-0.13%	-0.08%	0.19%	0.09%	0.11%	0.02%	0.23%	0.05%	-0.02%	0.00%
dist on B roads	1,399,633.80	1,408,073.83	15445	15526	-0.02%	-0.02%	-0.02%	-0.01%	0.01%	0.00%	0.02%	0.00%	0.07%	0.06%
dist on other roads	20,842,592.50	21,002,258.29	17860	17906	-0.03%	-0.03%	0.00%	0.01%	-0.01%	0.00%	0.02%	0.02%	0.01%	0.01%
dist on SRN	278,267,683.08	278,969,144.03	15491	15694	0.09%	0.07%	-0.08%	-0.03%	-0.05%	-0.01%	-0.10%	-0.04%	-0.01%	-0.03%
dist on M25 SW	3,461,687.73	3,451,737.27	359	373	-0.28%	-0.05%	-0.60%	-0.18%	**-4.48%**	**-1.32%**	-1.03%	-0.14%	0.01%	0.00%
dist on M25 NW	3,319,681.69	3,360,515.89	336	336	0.59%	0.07%	-1.24%	-0.30%	**-4.84%**	**-1.18%**	**-10.73%**	**-2.91%**	-0.23%	-0.06%
dist on M25 SE	3,498,124.28	3,330,760.68	513	527	-0.05%	-0.01%	**-4.24%**	**-1.93%**	-0.29%	-0.06%	-0.06%	-0.02%	0.01%	0.00%
dist on M25 NE	4,148,326.42	3,912,928.96	482	485	-0.08%	-0.06%	**-7.75%**	**-3.37%**	-0.12%	-0.01%	-5.76%	-1.25%	0.02%	-0.01%
dist on M25	14,427,820.13	14,055,942.80	1170	1196	0.03%	-0.02%	-3.68%	-1.51%	-2.29%	-0.62%	-4.39%	-1.08%	-0.04%	-0.02%
dist on M25 SE & NE	2,631,415.12	2,378,929.16	170	170	-0.25%	-0.06%	**-21.67%**	**-12.25%**	0.02%	0.02%	-2.57%	-0.69%	0.15%	0.04%
dist on M25 SW & NW	2,451,699.99	2,338,697.92	78	78	0.45%	0.11%	0.03%	0.02%	**-18.05%**	**-7.08%**	-5.24%	-0.95%	0.07%	0.01%
dist on M25 & M1	5,134,317.48	4,924,262.40	139	139	4.10%	0.93%	-2.35%	-0.59%	-2.55%	-0.43%	**-26.37%**	**-8.56%**	0.96%	0.27%
dist on M1	25,585,592.46	25,275,344.50	1388	1411	**3.09%**	**1.67%**	0.15%	0.03%	0.06%	0.01%	**-1.30%**	**-0.38%**	0.55%	0.31%
dist on A1	13,603,875.95	13,557,514.13	1294	1323	-0.83%	-0.44%	0.13%	0.03%	0.05%	0.01%	0.62%	0.20%	**-2.95%**	**-2.00%**
dist on A19	2,948,567.34	3,116,544.30	575	579	-0.10%	-0.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**1.20%**	**0.81%**

cannot be captured by the more simplistic approaches used in standard models. Additionally however, the more advanced model produces lower elasticities, which were already low. Low elasticities arise as a result of no obvious alternative route being available, i.e. a substantial deterioration in conditions on the current best route are needed to lead to a shift to another route. This is at least in part a result of the design of the road network, and attempts at using differently sized choice sets led to similar results. This situation is clearly very different from past studies looking mainly at intra-urban journeys where more options for alternative routes exist.

In closing, it should be noted that the sample used in this study was relatively small in comparison with the network size, and no information is available on how representative of the overall UK fleet of freight vehicles it is. The size of the sample also prevented us from looking in detail at choices between key competing routes, or paying specific attention to tolls (e.g. M6 Toll or various river crossings). It should also be noted that the error components for which significant estimates were obtained do not necessarily relate to those roads for which a priori knowledge would suggest this to be the case - this could be in large part due to the small size of the sample as well as the lack of representativeness. Future work, making use of larger samples, could also investigate the use of error components for specific combinations of roads. Data limitations also prevented us from looking at other potentially important factors, such as congestion, weather and safety, while, with larger samples, future work could also look at time-of-day effects. These should be incorporated in future studies. Finally, the extensive data cleaning effort required potentially also has some impact on the generalisability of our findings. Nevertheless, the work has clearly shown the potential for making use of advanced discrete choice models in the analysis of route choices with GPS data for very large networks.

Acknowledgements

This paper is based on a project commissioned by the UK Department for Transport. The opinions in this paper are those of the authors and do not necessarily reflect those of the UK Department for Transport. The authors remain solely responsible for any errors or omissions.

A Characteristics of final choice sets

Since the composition of the choice set strongly influences the modelling results, the structure of choice sets produced in Section 3 has to be taken into account as well. To evaluate the structure of the choice sets derived in this project, the following aspects are considered:

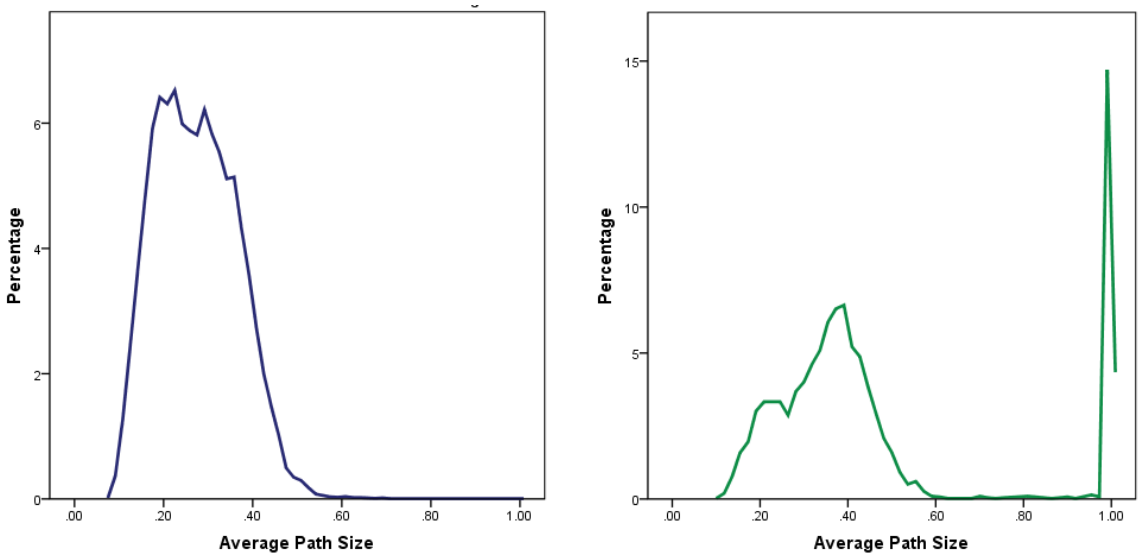
1. Is the size of the route set sufficient?
2. How often/well is the chosen route reproduced?
3. How diverse are the routes?
4. How does the distribution of road types compare to the chosen routes?

Since we had to filter out some of the OD pairs as discused in Section 4.1, the analyses were performed separately for the OD pairs that were used in the subsequent modelling and those that were filtered out. It should first be noted that this had only limited impact on the mean travel times for the chosen routes, with a decrease from 23.22 minutes to 22.87 minutes.

Table 4 shows the distribution of choice set sizes both for the final modelling data set and the OD pairs that were filtered out. It can be seen that in the modelling data set, over 98% of the choice sets are complete, so there should be no issue regarding the choice set size. However, about 20% of the OD pairs that were filtered out have 5 or fewer alternatives. The size for most of these choice sets is

Table 4: Distribution of choice set sizes

Choice set size [n of routes]	Share in final data set [%]	Share in OD pairs filtered out [%]
1 – 5	0.3	20.2
6 – 10	0.5	0.8
11 – 14	0.9	1.3
15 – 16	98.3	77.7



(a) Final data set

(b) OD pairs filtered out

Figure 4: Distribution of the average path sizes for resulting route sets

in fact one, which means that the algorithm was not able to find any alternatives to the chosen route. This only occurs when there is just one road between the origin and the destination. These cases are comparatively short trips in remote areas of the network such as in valleys, rural dead-end roads or at remote coastal locations.

Another important aspect on which choice set generation approaches are evaluated is their ability to reproduce the chosen route. While the choice set generation approach was able to reproduce the chosen route for only about 50% of the OD pairs that were filtered out, a reproduction rate of 74% was achieved for the final data set. This is in line with the figures generally reported in the literature (e.g. Rieser-Schssler et al., 2012; Prato and Bekhor, 2007; Ramming, 2002)

To investigate the diversity of the choice sets, Figure 4 depicts the distribution of the average path sizes. For the final data set, the path size distribution follows a similar pattern to what we have seen in past studies (e.g. Rieser-Schssler et al., 2012). Some choice sets have rather low average path sizes of around 0.2 but there is also a good number of choice sets with average path sizes of around 0.5 which is rather high. The second peak in the path size distribution for those OD pairs that were filtered out stems from the choice sets with just one alternative, which by definition obtain a path size of one.

The last criterion for the evaluation of the choice set structure is the plausibility of the road type distributions. For each of road type, the share of the overall route distance travelled on that type of road was calculated. In Figure 5, the distribution for each of road type proportions for the chosen routes is compared to the distribution of the generated alternatives. Again, this is done separately for final

choice set and those OD pairs that were filtered out. Over all road types, the road type proportions of the generated alternatives follow a similar pattern to those of the chosen routes. For the chosen routes as well as the alternatives, A roads dominate, followed by *other* roads. Motorways and B roads take comparatively small shares. The median road type proportions are very similar for the chosen routes and their alternatives for all road types, however, the width of the distribution differs. For A roads and other roads, the distributions of road type proportions in the chosen routes are wider than those for the generated alternatives. While motorways are more rarely used by the chosen routes, a substantial number of alternative routes used motorways. This is to be expected since motorways in general offer shorter travel times but aspects such as costs have not been taken into account in the choice set generation. In contrast to the previous evaluation steps, no systematic differences appear when comparing the road type proportion distributions of the filtered OD pairs with those of the final data set. Before moving on, it should also be noted that while the mean/median across O-D pairs in the proportion of the route using motorways is low, this is heavily influenced by the presence of O-D pairs with small shortest paths. Indeed, as can be seen later in Table 3, the total distance travelled on motorways is second only to that travelled on A-roads.

References

Azevedo, J. A., Santos Costa, M. E. O., Silvestre Madeira, J. J. E. R., Vieira Martins, E. Q., 1993. An algorithm for the ranking of shortest paths. European Journal of Operational Research 69 (1), 97–106.

Bekhor, S., Ben-Akiva, M. E., Ramming, M. S., 2006. Evaluation of choice set generation algorithms for route choice models. Annals of Operations Research 144 (1), 235–247.

Ben-Akiva, M. E., Bergman, M. J., Daly, A. J., Ramaswamy, R., 1984. Modelling inter-urban route choice behaviour. In: Volmuller, J., Hamerslag, R. (Eds.), Proceedings of the 9th International Symposium on Transportation and Traffic Theory. VNU Science Press, Utrecht, pp. 299–330.

Ben-Akiva, M. E., Bierlaire, M., 1999. Discrete choice methods and their applications to short-term travel decisions. In: Hall, R. (Ed.), Handbook of Transportation Science. Kluwer, Dordrecht, pp. 5–34.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transportation Research Part C 16 (2), 187–198.

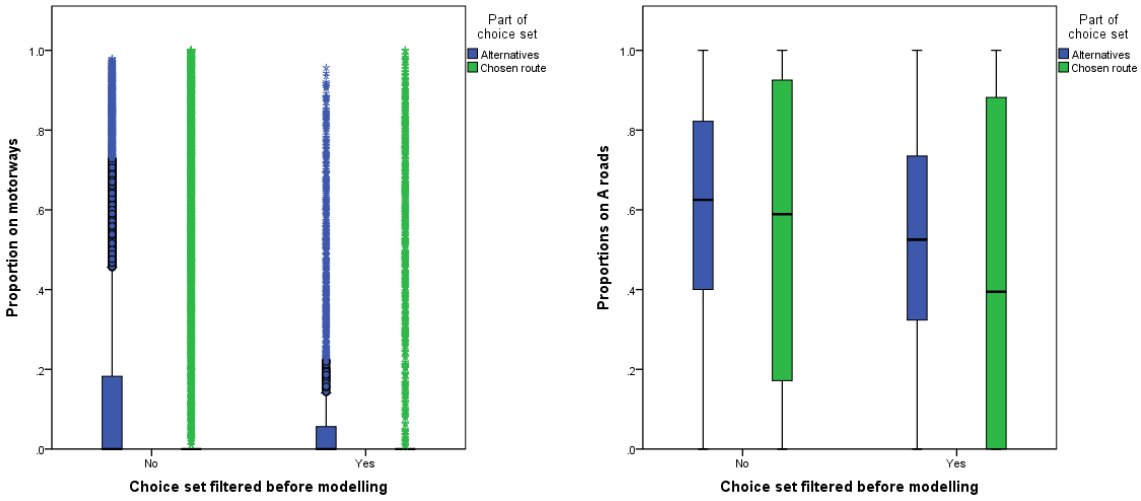
Bliemer, M. C. J., Bovy, P. H. L., 2008. Impact of route choice set on route choice probabilities. Transportation Research Record 2076, 10–19.

Bliemer, M. C. J., Bovy, P. H. L., Li, H., June 2007. Some properties and implications of stochastically generated route choice sets. In: 6th Triennial Symposium on Transportation Analysis (TRISTAN). Phuket Island.

Bovy, P. H. L., Fiorenzo-Catalano, S., 2007. Stochastic route choice set generation: Behavioral and probabilistic foundations. Transportmetrica 3 (3), 173–189.

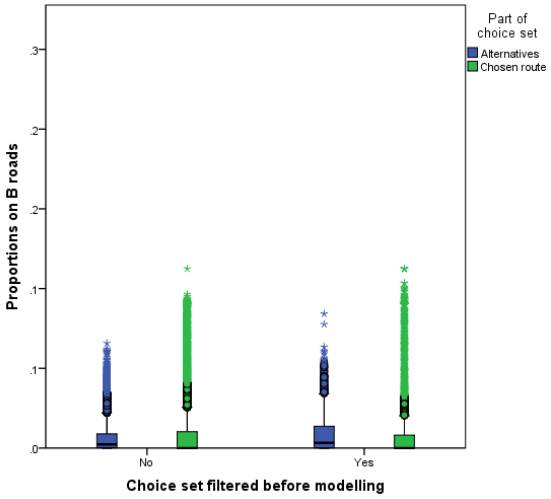
Bricka, S., Sen, S., Paleti, R., Bhat, C. R., 2012. A comparative analysis of GPS-based and travel survey-based data. Transportation Research Part C: Emerging Technologies 21 (1), 67–88.

Cascetta, E., Nuzzola, A., Russo, F., Vitetta, A., 1996. A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks. In:

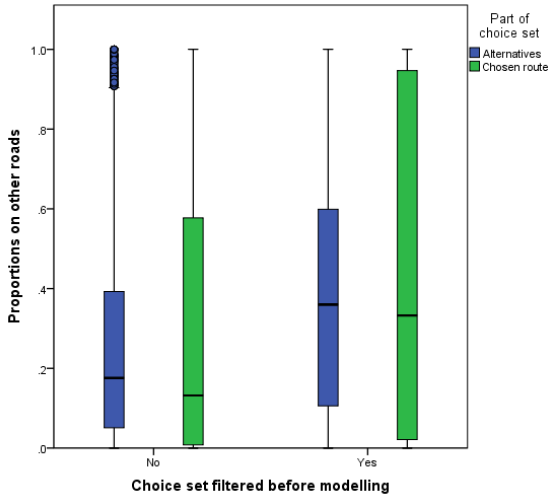


(a) Motorways

(b) A roads



(c) B roads



(d) Other roads

Figure 5: Distributions of road type proportions

Lesort, J. B. (Ed.), Proceedings of the 13th International Symposium on Transportation and Traffic Theory. Pergamon, Oxford, pp. 697–711.

Dalumpines, R., Scott, D. M., January 2011. GIS-based map matching: Development and demonstration of postprocessing map-matching algorithm for transportation research. In: 90th Annual Meeting of the Transportation Research Board. Washington.

Daly, A., 2010. Cost damping in travel demand models: Report of a study for the department for transport. RAND TR-717-DFT.

de la Barra, T., Perz, B., Aez, J., September 1993. Multidimensional path search and assignment. In: 21st Planning and Transport, Research and Computation (PTRC) Summer Meeting. Manchester.

DfT, 2009. Values of time and operating costs, tag unit 3.5.6. http://www.dft.gov.uk/webtag/documents/expert/unit3.5.6.php.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271.

Dugge, B., 2006. Ein simultanes Erzeugungs-, Verteilungs-, Aufteilungs- und Routenwahlmodell. Ph.D. thesis, Technical University Dresden, Dresden.

Fox, J. G., Mahanty, J., 1970. The effective mass of an oscillating spring. American Journal of Physics 38 (1), 98–100.

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

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

Halldrsdttir, K., Rieser-Schssler, N., Axhausen, K. W., Nielsen, O. A., Prato, C. G., 2014. Efficiency of choice set generation methods for bicycle routes. European Journal of Transport and Infrastructure Research 14 (4), 332–348.

Hoogendoorn-Lanser, S., Bovy, P. H. L., 2007. Modeling overlap in multi-modal route choice by inclusion of trip part specific path size factors. Transportation Research Record 2003, 74–83.

Hoogendoorn-Lanser, S., van Nes, R., Bovy, P. H. L., August 2006. A rule-based approach to multi-modal choice set generation. In: 11th International Conference on Travel Behaviour Research (IATBR). Kyoto.

Lawler, E. L., 1976. Combinatorial Optimization: Networks and Matroids. Holt, Rinehart & Winston, New York.

Lefebvre, N., Balmer, M., September 2007. Fast shortest path computation in time-dependent traffic networks. In: 7th Swiss Transport Research Conference. Ascona.

Li, H., Guensler, R., Ogle, J., 2005. Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data. Transportation Research Record 1926, 162–170.

Marchal, F., Hackney, J. K., Axhausen, K. W., 2005. Efficient map matching of large Global Positioning System data sets: Tests on speed-monitoring experiment in Zrich. Transportation Research Record 1935, 93–100.

Marchal, P., Madre, J.-L., Yuan, S., January 2011. Post-processing procedures for person-based GPS data collected in the French National Travel Survey 2007-2008. In: 90th Annual Meeting of the Transportation Research Board. Washington.

MATSim, 2013. Multi Agent Transportation Simulation.
URL http://www.matsim.org

Menghini, G., Carrasco, N., Schssler, N., Axhausen, K. W., 2010. Route choice of cyclists in zurich. Transportation Research Part A: Policy and Practice 44 (9), 754–765.

Moiseeva, A., Jessurun, J., Timmermans, H. J. P., 2010. Semi-automatic imputation of activity travel diaries using GPS-traces, prompted recall and context-sensitive learning algorithms. Transportation Research Record 2183, 60–68.

Morikawa, T., 1996. A hybrid probabilistic choice set model with compensatory and noncompensatory choice rules. In: Hensher, D. A., King, J., Oum, T. (Eds.), World Transport Research: Proceedings of the 7th World Conference on Transport Research. Pergamon, Oxford, pp. 317–325.

Nielsen, O., 2000. A stochastic transit assignment model considering differences in passengers utility functions. Transportation Research Part B 34 (5), 377–402.

Ohnmacht, T., Kowald, M., January 2014. Route-recording on high resolution transportation network databases for national transport surveys: An option for valid and reliable distance measures? In: 93rd Annual Meeting of the Transportation Research Board. Washington.

Parkany, E., Du, J., Aultman-Hall, L., Gallagher, R., 2006. Modeling stated and revealed route choice: consideration of consistency, diversion, and attitudinal variables. Transportation Research Record 1985, 29–39.

Prato, C., 2009. Route choice modeling: past, present and future research directions. Journal of Choice Modelling 2 (1), 65–100.

Prato, C. G., Bekhor, S., 2006. Applying branch-and-bound technique to route choice set generation. Transportation Research Record 1985, 19–28.

Prato, C. G., Bekhor, S., 2007. Modeling route choice behavior: How relevant is the composition of choice set? Transportation Research Record 2003, 64–73.

Pyo, J.-S., Shin, D.-H., Sung, T.-K., August 2001. Development of a map matching method using the multiple hypothesis technique. In: Intelligent Transportation Systems Conference (ITSC). Oakland.

Quattrone, A., Vitetta, A., 2011. Random and fuzzy utility models for road route choice. Transportation Research Part E: Logistics and Transportation Review 47 (6), 1126 – 1139.
URL http://www.sciencedirect.com/science/article/pii/S1366554511000573

Ramming, M. S., 2002. Network knowledge and route choice. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Rieser-Schssler, N., Balmer, M., Axhausen, K. W., 2012. Route choice sets for very high-resolution data. Transportmetrica.

Schssler, N., 2010. Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour. Ph.D. thesis, ETH Zurich, Zurich.

Stopher, P. R., Jiang, Q., FitzGerald, C., June 2005. Processing GPS data from travel surveys. In: 2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications. Toronto.

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

Train, K., 2009. Discrete Choice Methods with Simulation, second edition Edition. Cambridge University Press, Cambridge, MA.

van der Zijpp, N. J., Fiorenzo-Catalano, S., 2005. Path enumeration by finding the constrained k-shortest paths. Transportation Research Part B: Methodological 39 (6), 545–563.

VOSA, 2009. Rules on drivers hours and tachographs: goods vehicles in the uk and europe. Vehicle and Operator Services Agency, Revised 2009 GV262 – 02, Department for Transport.

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

Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of underreporting on mileage and travel time estimates - results from Global Positioning System-enhanced household travel survey. Transportation Research Record 1854, 189–198.

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., Axhausen, K., 2004. Eighty weeks of global positioning system traces: approaches to enriching trip information. Transportation Research Record 1870, 46–54.