Downloaded from orbit.dtu.dk on: Nov 08, 2024

DTU Library

Efficiency of choice set generation methods for bicycle routes

Halldórsdóttir, Katrín; Rieser-Schussler, Nadine; Axhausen, Kay W.; Nielsen, Otto Anker; Prato, Carlo Giacomo

Published in:
European Journal of Transport and Infrastructure Research

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Halldórsdóttir, K., Rieser-Schussler, 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.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

EJTIR

Issue 14(4), 2014
pp. 332-348
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Efficiency of choice set generation methods for bicycle routes

Katrín Halldórsdóttir[1]
Department of Transport, Technical University of Denmark, Kgs. Lyngby, Denmark.

Nadine Rieser-Schüssler[2], Kay W. Axhausen[3]
Institute for Transport Planning and Systems, Swiss Federal Institute of Technology, Zurich, Switzerland.

Otto A. Nielsen[4], Carlo G. Prato[5]
Department of Transport, Technical University of Denmark, Kgs. Lyngby, Denmark.

The current study analyses the efficiency of choice set generation methods for bicycle routes and proposes the extension of cost functions to bicycle-oriented factors not limited to distance and time. Three choice set generation methods for route choice were examined in their ability to generate relevant and heterogeneous routes: doubly stochastic generation function, breadth first search on link elimination, and branch & bound algorithm. Efficiency of the methods was evaluated for a high-resolution network by comparing the performances with four multi-attribute cost functions accounting for scenic routes, dedicated cycle lanes, and road type. Data consisted of 778 bicycle trips traced by GPS and carried out by 139 persons living in the Greater Copenhagen Area, in Denmark. Results suggest that both the breadth first search on link elimination and the doubly stochastic generation function generated realistic routes, while the former outperformed in computation cost and the latter produced more heterogeneous routes.

Keywords: *bicycle route choice, bicycle route generation, branch and bound, breadth first search, choice set generation, stochastic generation function.*

1. Introduction

The growing interest in the transition towards the use of sustainable transport modes motivates the search for factors determining the selection of the bicycle as a viable alternative to the car. Bicycle route choice models provide directions to this search, but findings from these models depend on the observation of actual route choices and the generation of plausible alternatives. While the former is a challenge that recent enhancements in technology and software help tackling, the latter is a challenge that recent advances in path generation help confronting, not

[1] A: Bygningstorvet 116B, 2800 Kgs. Lyngby, Denmark T: +45 45256563 F: +45 152 788 546 E: katha@transport.dtu.dk
[2] A: Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland T: +41 446333085 F: +41 446331057 E: nadine.rieser@ivt.baug.ethz.ch
[3] A: Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland T: +41 446333943 F: +41 446331057 E: axhausen@ivt.baug.ethz.ch
[4] A: Bygningstorvet 116B, 2800 Kgs. Lyngby, Denmark T: +45 45251514 F: +31 152 788 546 E: oan@transport.dtu.dk
[5] A: Bygningstorvet 116B, 2800 Kgs. Lyngby, Denmark T: +45 45251514 F: +31 152 788 546 E: cgp@transport.dtu.dk

without the uncertainty related to the dependency of model estimates on choice set composition (see, e.g., Bekhor et al., 2006; Bliemer and Bovy, 2008; Prato and Bekhor, 2007).

The literature in bicycle route choice shows that most studies have used stated preference (SP) data (e.g., Axhausen and Smith, 1986; Bovy and Bradley, 1985; Hopkinson and Wardman, 1996; Hunt and Abraham, 2007; Krizek, 2006; Sener et al., 2009; Stinson and Bhat, 2003; Tilahun et al., 2007), while only a few studies have used revealed preference (RP) data (e.g., Aultman-Hall et al., 1997; Broach et al., 2011; Broach et al., 2012; Hood et al., 2011; Howard and Burns, 2001; Hyodo et al., 2000; Menghini et al., 2010; Shafizadeh and Niemeier, 1997). On the one hand, SP data trade the easiness in individuating alternative routes with the possible bias in predefining the factors being relevant to route choices of cyclists. On the other hand, RP data trade the realism of observing actual cyclists' behaviour with the challenge of generating plausible alternative routes prior to model estimation.

While the challenge of collecting RP data (i.e., actual route choices) has greatly benefitted from enhancements in GPS device technology, GPS data post-processing (e.g., Schüssler and Axhausen, 2008, 2009b; Stopher, 2009; Tsui and Shalaby, 2006), and highly detailed network digitalization, the challenge of generating plausible alternative routes is still testing. Most studies have focused on implementing path generation methods for cars or public transport, which are normally generated on a simplified network, and only few studies have focused on bicycle route choice sets, which require a highly detailed network. Menghini et al. (2010) applied a Breadth First Search on Link Elimination (BFS-LE) method (Rieser-Schüssler et al., 2012) while limiting the cost function to the route length. Broach et al. (2010) compared a modified route labelling method to a K-shortest paths link penalty (Cascetta et al., 1996; de la Barra et al., 1993; Ramming, 2001), a simulated shortest paths (Bekhor et al., 2006; Ramming, 2001), and labelled routes method (Bekhor et al., 2006; Ben-Akiva et al., 1984; Ramming, 2001). The modified route labelling method performed best out of all four methods, however obtained lower coverage in comparison with studies focusing on car route choice. This is likely because the network used in the study was a realistic "all streets" network, while most car route choice studies cited used much coarser auto networks. Hood et al. (2011) implemented a doubly stochastic generation function (DSGF) (Bovy and Fiorenzo-Catalano, 2007; Nielsen, 2000) with a multi-attribute cost function while observing better performance than a BFS-LE method with single-attribute cost function. They obtained a slightly higher coverage than Broach et al. (2010). Notably, the first two studies defined cost functions either containing only one attribute or containing only travel time and distance. Although Hood et al. (2011) included a multi-attribute cost function, they only manage to reproduce one-third of the observed routes. This emphasises the importance of both identifying factors that are important in the choice set generation process for bicycle route choice as well as exploring algorithmic issues in generating a plausible set of path choice alternatives in a highly detailed network.

The current study extends the body of knowledge on choice set generation for bicycle route choice. Firstly, the current study applies to the bicycle context the three most effective path generation methods in the car context: BFS-LE, DSGF, and branch & bound (B&B) (Hoogendoorn-Lanser et al., 2006; Prato and Bekhor, 2006). These methods are chosen in order to investigate their transferability in the context of choice set generation for bicycle routes choice, given their proven ability in reproducing observed car route choices. Secondly, the current study evaluates the efficiency of the three path generation methods in a high-resolution network by using different evaluation methods, such as replicating the observed routes while also generating realistic alternatives that take into account taste heterogeneity across cyclists. Lastly, the current study extends the path generation literature by proposing multi-attribute cost functions that account for scenic routes, dedicated cycle lanes, and road type. These attributes are relevant to the route choice of cyclists and are important when intending to estimate models providing insights into the factors relating to cyclists' choices of routes.

The remainder of this paper is organised as follows. Section 2 describes the data collected for this study. Section 3 describes the path generation methods applied in this study, the bicycle-tailored multi-attribute cost functions, and the methods used to evaluate the efficiency of the methods. Section 4 presents the results from the implementation and the comparison of the path generation methods. Section 5 discusses the results and draws conclusions.

2. Data

The current study uses a dataset consisting of 778 bicycle trips, traced by GPS and carried out by 139 persons living in the Greater Copenhagen Area in Denmark. In addition to collecting GPS tracks, travel diaries were collected from a sample of the participants by means of a web-based survey.

Extensive data processing was required to obtain data that could be used for choice set generation of bicycle routes, as the GPS data collection and the travel diaries focused on all modes of transport in the Greater Copenhagen Area. The post-processing procedure used is described in detail in Schüssler and Axhausen (2008, 2009b). Initially, various criteria were first used to filter the data, e.g. the number of satellites in view, altitude value, the Horizontal Dilution of Precision (HDOP) value, sudden jumps in position, etc., followed by a Gauss kernel smoothing approach to remove random errors. Then, different criteria were applied to identify trips and activities and trips were divided into single-mode stages. Last, modified fuzzy logic rules were applied to identify the transport mode for each stage using the median speed, 95th percentile acceleration, and 95th percentile speed. The original fuzzy logic rules had to be altered in order to better fit the travel behaviour of Danes and extended with more disaggregate components using GIS software (Rasmussen et al., 2013). The travel diaries were used for validation of the post-processing procedure.

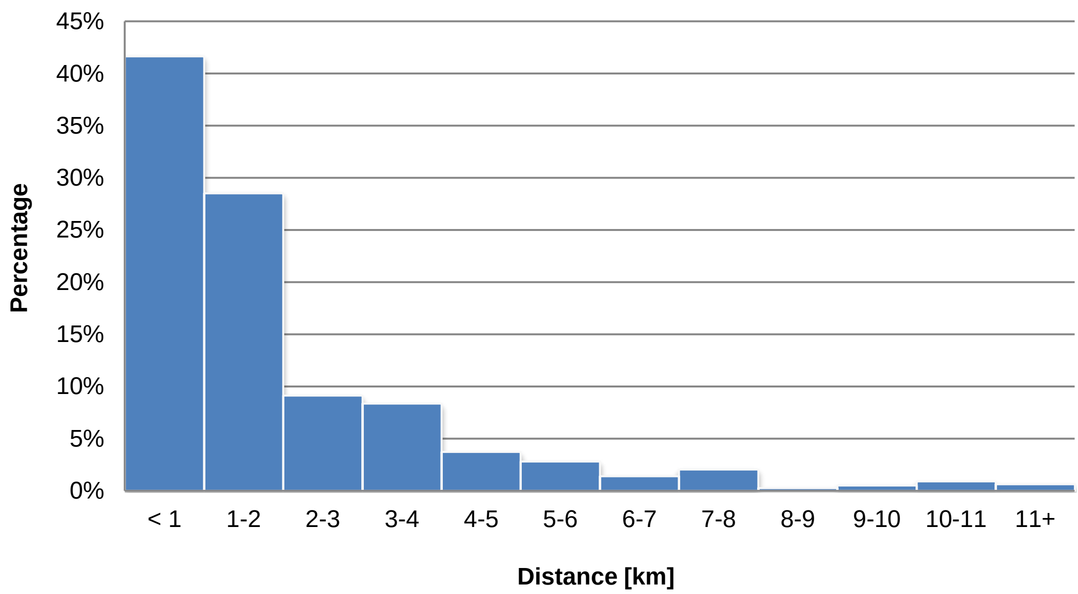

Figure 1. Distribution of the observations over distance travelled.

After the data post-processing, a total of 1,824,034 GPS points and 4,552 stages were identified, and 490,062 points and 1,026 stages were retained for further analysis after the travel mode was identified as bicycle.

After the filtering, the processed bicycle trips were mapped to a high-resolution network using the map-matching algorithm developed by Schüssler and Axhausen (2009a) that extended a previous algorithm developed by Marchal et al. (2005). The network consists of 110,893 nodes and 272,586 directional links for the study area, and the high-resolution is the result of the compilation of various sources in order to consider attributes that were considered relevant to bicycle route choice such as road type, segregated cycle paths, and land use information. The network includes small paths only accessible by bicycles and pedestrians, and does not include motorways and expressways where cycling is illegal in Denmark.

After the map-matching algorithm was run, there were 778 stages remaining for the choice set generation. In some cases, stages were filtered because of missing network links or because the traveller was cycling off-road, which resulted in the lower number of stages. The average distance was approximately 2 km and approximately 90% of the trips were less than 5 km. Figure 1 shows the distribution of the observations over distance travelled.

3. Methods

This section introduces the applied path generation methods, the defined cost functions and the implemented comparison methods. It should be noted that a maximum choice set size of 20 alternative routes was defined prior to choice set generation, and a time abort threshold was predefined in order to move on to the next observation after processing origin-destination pairs for which the choice set generation could not be completed within the time interval.[6]

3.1 Choice set generation methods

Doubly stochastic generation function
The DSGF (Bovy and Fiorenzo-Catalano, 2007; Nielsen, 2000) accounts for variation in travellers' link cost and differences in travellers' attribute preferences by drawing random costs and random parameters from probability distributions. Advantages of this method are the inherent heterogeneity of the generated alternatives and its computational efficiency in large networks, even though the randomisation of link costs and the parameters can be time consuming in a high-resolution network.

In the DSGF method, a shortest path search is carried out iteratively using an implementation of the Dijkstra´s algorithm (Dijkstra, 1959) on a realization of the network. At each iteration, the realization of the network is obtained by randomly drawing the cost of each link from a probability distribution and extracting attribute preferences for each traveller from another probability distribution. After each iteration, only unique routes not generated in previous ones are added to the route choice set as the same route may be found several times during the process, even though the realizations of the network are obtained from different costs and preference parameters. The shortest path search is repeated iteratively until the preselected maximum choice set size is achieved or the predefined time abort threshold is reached.

Although the literature reports the implementation of a multi-attribute cost function for the DSGF method in the bicycle route choice context (Hood et al., 2011), the current study extends this implementation by testing and reporting the results of four different cost functions that consider not only route length and time, but also bicycle-oriented factors such as preferences for road types, dedicated cycle paths and land use. As the correct definition of choice sets is a necessary condition for obtaining meaningful parameter estimates, including these bicycle-oriented factors is essential to the in-depth understanding of cyclists' preferences.

[6] The alternative routes were generated using a tool developed in Java, originally developed for the automatic processing of GPS tracks to reconstruct travel diaries (POSDAP) (see www.sourceforge.net/projects/posdap).

Breadth first search on link elimination

The BFS-LE method (Rieser-Schüssler et al., 2012) combines a breadth first search with topologically equivalent network reduction. The procedure concentrates on generating a route set, which afterwards can be reduced to an individual choice set. Its advantage is a high computational efficiency in a high-resolution network while ensuring a significant level of route variety.

The algorithm searches for the shortest path between origin and destination. Consecutively, the links of the shortest path are removed one by one and the shortest path for the resulting network is determined. Once all links of the original shortest path have been processed, the algorithm proceeds to the next level, where two links at a time are eliminated. The algorithm monitors the generated networks and only keeps unique and connected routes. The algorithm continues until the maximum number of unique routes has been generated, the time abort threshold is met or there are not more feasible routes between origin and destination.

As the literature reports only single-attribute cost functions for the implementation of the BFS-LE method in the bicycle route choice context (Hood et al., 2011; Menghini et al., 2010), the current study proposes to examine the efficiency of the BFS-LE method with the same multi-attribute cost functions implemented for the DSGF method. The same input parameters are applied for the two methods, although obviously for the BFS-LE the error components and the error terms are not extracted from a probability distribution.

Branch & bound

The B&B method (Hoogendoorn-Lanser et al., 2006; Prato and Bekhor, 2006) constructs a connection tree between the origin and destination of a trip by processing link sequences according to a branching rule accounting for logical constraints devised to increase route likelihood and heterogeneity. The strength of this algorithm is the realism and the heterogeneity of the generated routes, but a disadvantage is the high computation time in a high-resolution network. Rieser-Schüssler et al. (2012) tested the method with GPS observed car trips in a high-resolution network where the average number of links per chosen route was 65.69. This proved to be too high for the B&B algorithm, as within a reasonable computation time it only managed to produce alternatives for origin-destination pairs connected by very short paths. However, empirical results have shown that the algorithm can be applied to networks of different sizes by applying different design parameters (Bekhor and Prato, 2009).

In the current study, different behavioural constraints were tested to exclude links that: (i) take the cyclists farther from the destination and closer to the origin (directional constraint); (ii) cause the travel time to be excessively high when compared to the shortest path (temporal constraint); (iii) cause the cyclist to have a detour larger than an acceptable value (loop constraint); (iv) are shared by other routes that would not be considered as separate alternatives (similarity constraint). Table 1 presents the input parameters. Firstly, basis values were implemented according to the indications by Prato and Bekhor (2006). Then, relaxed values were applied to allow for higher travel times and greater overlap, and restrictive values were used to allow for lower travel times and less overlap. It should be noted that the time abort threshold was applied differently than the BFS-LE and the DSGF, namely the B&B searched through the network tree before checking whether the time restriction has been met. As the search tree was rather large because of the highly detailed network, at times the algorithm took longer to compute and the choice set size exceeds the maximum. Consequently, this method could not be compared to the other two algorithms in relation to computation time or number of unique routes created for each chosen route. However, the structure of the derived choice set could be compared.

Table 1. Implementation of the branch & bound generation technique

	Basis	Restricted	Relaxed
Directional factor	1.10	1.10	1.10
Temporal factor	1.50	1.33	1.67
Loop factor	1.20	1.20	1.20
Similarity factor	0.80	0.75	0.85

3.2 Cost functions

Four cost functions were tested for the BFS-LE and the DSGF methods with various input parameters being considered with the aim to get realistic route alternatives, to obtain the best coverage over the average of the sample, and to capture heterogeneous preferences across cyclists. The parameters accounted for preferences of road types, cycle lanes and land use. Table 2 presents an overview of the calibrated parameters and the variation factors applied in the cost functions. The following sup-sections describe in more detail the tested cost functions and the calibration of the parameters.

Table 2. Overview of the calibrated parameters and variation factors

	Cost function			
	Road type	Cycle lanes	Land use	Road type, cycle lanes, and land use
Parameters				
Large roads	0.167	–	–	0.167
Small roads	0.333	–	–	0.333
Other roads	0.500	–	–	0.500
Segregated cycle lanes	–	0.400	–	0.400
No segregated cycle lanes	–	0.600	–	0.600
Scenic roads	–	–	0.125	0.125
Non-scenic roads	–	–	0.250	0.250
Forest roads	–	–	0.500	0.500
Non-forest roads	–	–	0.125	0.125
Variation factors				
$\xi_{Large\ roads}$	2	–	–	2
$\xi_{Small\ roads}$	10	–	–	10
$\xi_{Other\ roads}$	10	–	–	10
$\xi_{Segregated\ cycle\ lanes}$	–	2	–	2
$\xi_{No\ segregated\ cycle\ lanes}$	–	10	–	10
$\xi_{Scenic\ roads}$	–	–	2	2
$\xi_{Non\text{-}scenic\ roads}$	–	–	3	3
$\xi_{Forest\ roads}$	–	–	7	7
$\xi_{Non\text{-}forest\ roads}$	–	–	2	2
ε_a	2	2	2	2

Road type

Cost functions should consider information regarding different road types for cyclists, and in the case of absence of specific information, regarding different road types in general. For example, three different road types may be considered (e.g., large traffic roads, small traffic roads, other roads). Larger roads are usually properly equipped with segregated cycle paths and cycle lanes (e.g., Hunt and Abraham, 2007; Stinson and Bhat, 2003) in the Greater Copenhagen Area, while smaller roads have usually fewer or non-segregated cycle lanes, thus resulting in a mixed traffic scenario between cyclists and motorists, but in return have lower speed limits and traffic volumes (e.g., Axhausen and Smith, 1986; Antonakos, 1994). Other roads are a mixture of pedestrian paths or shared bicycle and pedestrian paths.

The first cost function represented individuals with different preference for the three road types:

$$C_a = \sum_k \left(\left(\beta_{RoadType_k} + \xi_{RoadType_{ak}} \right) \cdot RoadType_{ak} \cdot Length_a \right) + \varepsilon_a \quad (1)$$

where C_a is the random cost of link a, $Length_a$ is the length of link a, $RoadType_{ak}$ is the road type k that link a belongs to, $\xi_{RoadType_{ak}}$ are error components related to road type k of link a, $\beta_{RoadType_k}$ are coefficients related to road type k, and ε_a is the random error term for link a.

The implementation of the cost function differs between DSGF and BFS-LE. In the DSGF method, the error terms ε_a express the random link costs and the error components $\xi_{RoadType_{ak}}$ express the heterogeneous preferences. Each error term ε_a was computed as the link length multiplied by a standard normal distribution and a variation factor that determined the "width" of the distribution from which the link lengths were drawn. Each error component $\xi_{RoadType_{ak}}$ was calculated as the corresponding parameter β multiplied by a standard normal distribution and a distribution factor that determined the "width" of the distribution from which the utility parameter itself is drawn in the doubly stochastic case. Calibration of the parameters resulted in the large traffic roads having the lowest cost and capturing preferences for high quality bicycle facilities, and the small and the other traffic roads having double and triple cost and capturing preferences for the most direct route and for paths not shared with pedestrians. In the DSGF method, each error term ε_a was computed with a variation factor of 2, and each error component $\xi_{RoadType_{ak}}$ were computed with a variation factor of 2 for large roads and 10 for small and other roads. In the BFS-LE method, variation in in travellers' link cost and differences in travellers' attribute preferences are not considered, and hence each error term ε_a is equal to zero for every link a and each error component $\xi_{RoadType_{ak}}$ is equal to one for every road type k and every link a.

Cycle lanes
Cost functions could consider information on cycle lanes, and in particular on Copenhagen-style lanes that are segregated lanes with raised curbs separating the cycle lane from the motorized traffic on one side and from the pedestrians on the other side. Studies have shown that some cyclists prefer routes that are separated from motorized traffic (e.g., Hunt and Abraham, 2007; Stinson and Bhat, 2003) and are willing to take detours to travel on bicycle paths, while others prefer the most direct route (e.g., Stinson and Bhat, 2005).

The second cost function characterized individuals with different perspectives to use segregated bicycle paths:

$$C_a = \sum_k \left(\left(\beta_{BikeLanes_k} + \xi_{BikeLanes_{ak}} \right) \cdot BikeLanes_{ak} \cdot Length_a \right) + \varepsilon_a \quad (2)$$

where C_a is the random cost of link a, $Length_a$ is the length of link a, $BikeLanes_{ak}$ indicates the presence of cycle lane configuration k on link a, $\xi_{BikeLanes_{ak}}$ are error components related to cycle lane configuration k in link a, $\beta_{BikeLanes_k}$ are coefficients related to cycle lane configuration k, and ε_a is the random error term for link a.

Calibration of the parameters resulted in roads with no segregated cycle lanes having almost two times larger cost than roads with segregated bicycle lanes. In the DSGF method, each error term ε_a was computed with a variation factor of 2, and each error component $\xi_{BikeLanes_{ak}}$ was calculated with a variation factor of 2 for roads with segregated cycle paths and 10 for roads without segregated cycle paths. In the BFS-LE, each error term ε_a was equal to zero for every link a and each error component $\xi_{BikeLanes_{ak}}$ was equal to one for every cycle lane configuration k and every link a.

Land use
Cost functions could also consider land use attributes, in particular when referring to scenic areas such as the lakes in the Greater Copenhagen Area, as scenic routes are considered attractive for cyclists (e.g. Antonakos, 1994; Axhausen and Smith, 1986). Major forest areas should also be considered, as the bicycle paths in forests are usually dirt paths and in some cases only suitable for mountain bicycles.

The third cost function assumed that cyclists have different preferences when travelling in different areas:

$$C_a = \sum_k \left(\left(\beta_{LandUse_k} + \xi_{LandUse_{ak}} \right) \cdot LandUse_{ak} \cdot Length_a \right) + \varepsilon_a \tag{3}$$

where C_a is the random cost of link a, $Length_a$ is the length of link a, $LandUse_{ak}$ indicates the land use type k associated to link a, $\xi_{LandUse_{ak}}$ are error components related to land use type k for link a, $\beta_{LandUse_k}$ are coefficients related to land use type k, and ε_a is the random error term for link a.

Calibration of the parameters resulted in non-scenic roads (i.e., not alongside lakes) having double the cost compared to scenic roads, thus capturing the preference for scenic routes. Also, calibration resulted in links in forest areas having four times the cost of roads in non-forest areas, thus capturing the disutility of cycling on gravel or dirt paths. In the DSGF method, each error term ε_a was computed with a variation factor of 2, and each error component $\xi_{LandUse_{ak}}$ was calculated with a variation factor of 2 for roads alongside lakes and not in forests, 3 for roads not alongside lakes and 7 for roads in forests. In the BFS-LE, each error term ε_a was equal to zero for every link a and each error component $\xi_{LandUse_{ak}}$ was equal to one for every land use type k and every link a.

Road type, cycle lanes, and land use
The fourth cost function included all three cost attributes, thus capturing a multi-attribute heterogeneous preference structure across individuals:

$$\begin{aligned} C_a = &\sum_k \left(\left(\beta_{RoadType_k} + \xi_{RoadType_{ak}} \right) \cdot RoadType_{ak} \cdot Length_a \right) \\ &+ \sum_k \left(\left(\beta_{BikeLanes_k} + \xi_{BikeLanes_{ak}} \right) \cdot BikeLanes_{ak} \cdot Length_a \right) \\ &+ \sum_k \left(\left(\beta_{LandUse_k} + \xi_{LandUse_{ak}} \right) \cdot LandUse_{ak} \cdot Length_a \right) + \varepsilon_a \end{aligned} \tag{4}$$

Calibration of the parameters resulted in small traffic roads having double and other roads having triple the cost of large traffic roads, roads with segregated cycle lanes having slightly over two times higher cost than large traffic roads, and roads with no segregated cycle lanes having almost four times larger cost. Moreover, roads alongside lakes resulted having slightly lower cost than large traffic roads, while not having lakes along the route implied almost double cost. Also, roads in forests resulted having a cost three times higher than large traffic roads. In the DSGF method, each error term ε_a was computed with a variation factor of 2, and each error component $\xi_{RoadType_{ak}}$, $\xi_{BikeLanes_{ak}}$ and $\xi_{LandUse_{ak}}$ was calculated with the same variation factors applied to the cases with a single attribute in the cost function. In the BFS-LE, each error term ε_a is equal to zero for every link a and each error component $\xi_{RoadType_{ak}}$, $\xi_{BikeLanes_{ak}}$ and $\xi_{LandUse_{ak}}$ was equal to one.

3.3 Evaluation methods

The effectiveness of the three choice set generation methods was evaluated by comparing the generated choice sets to the observed routes. If the choice sets contained the actual chosen route among paths produced with the generation methods, they were considered consistent with the observed behaviour. The consistency of a route choice set generation method was evaluated with respect to the observed behaviour by considering the length of the links shared between the generated route and the observed route for each choice set:

$$O_{nr} = \frac{L_{nr}}{L_n} \tag{5}$$

where O_{nr} is the overlap measure, L_{nr} is the overlapping length between the path generated by choice set generation method r and the observed path for cyclist n, and L_n is the length of the observed path for cyclist n. The coverage is defined as the percentage of observations for which an algorithm generates a route that satisfies a particular threshold for the overlap measure:

$$\max_r \sum_{n=1}^{N} I\left(O_{nr} \geq \delta\right) \tag{6}$$

where $I(\cdot)$ is the coverage function, where when its argument is true it is equal to 1 and when false it equals to 0, and δ is a threshold for the overlap measure.

In order to investigate the heterogeneity of the choice set composition, the path size was calculated for each route in each choice set (Ben-Akiva and Bierlaire, 1999):

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

where PS_{in} is the path size factor, Γ_i is the set of all links of path i, l_a is the length of link a, L_i is the length of path i, C_n is the generated choice set for cyclist n, and δ_{aj} equals 1 if link a is on path i and 0 otherwise. The path size ranges between 0 and 1, indicating the portion of the route that constitutes a completely independent alternative. Thus, a unique route will have a path size equal to one, while two duplicate routes will each have a path size factor of ½, three routes that completely overlap will each has a size of ⅓, and so on. The path size distribution over the choice sets C_n indicates whether the route choice sets contains heterogeneous routes by representing their average degree of independence.

The behavioural consistency of the route choice set generation methods with respect to an ideal algorithm was measured with a consistency index (see Bekhor and Prato, 2006). Ideally, a choice set generation method would reproduce the observed behaviour perfectly by replicating link by link all the routes collected and would result in 100% coverage for a 100% overlap threshold. However, the actual choice set generation methods only partially reproduce the observed behaviour and generate different numbers of routes. The index measures the behavioural consistency of the methods by accounting for the total overlap over all the observations:

$$CI_r = \frac{\sum_{n=1}^{N} O_{nr,\max}}{N \cdot O_{\max}} \tag{8}$$

where CI_r is the consistency index of choice set generation method r, $O_{nr,max}$ is the maximum overlap measure obtained with the paths generated by method r for the observed choice of each cyclist n, and O_{max} is the overlap 100% overlap over all the N observations for the ideal method.

4. Results

Table 3 presents the coverage results for the three choice set generation methods according to different overlap thresholds varying from complete replication to the reproduction of 70% of the collected routes. The first four rows show the results for the BFS-LE method, followed by the results for the DSGF method, and the results from the B&B method.

Table 3. Coverage measures of the applied algorithms

	Overlap threshold			
Algorithm	100%	90%	80%	70%
BFS-LE method				
Road type	62.2	67.9	72.8	78.8
Segregated bicycle path	66.1	72.0	78.0	82.6
Land use	62.0	67.0	74.6	81.9
All three cost attributes	67.9	74.8	80.1	84.8
Doubly stochastic generation function				
Road type	62.2	69.0	75.3	82.4
Segregated bicycle path	58.6	64.0	70.7	78.3
Land use	58.9	64.5	70.8	76.2
All three cost attributes	63.5	71.1	75.2	79.2
Branch and bound algorithm				
Basis	38.0	40.1	45.2	50.4
Relaxed	38.3	40.2	43.6	49.7
Restricted	40.4	42.7	46.5	51.3

The BFS-LE method duplicated 62% and up to almost 68% of the chosen routes, whereas at least 79% and up to almost 85% were reproduced with a 70% overlap threshold. The DSGF method replicated approximately 59% and up to almost 64% of the chosen routes, whereas more than 76% and up to more than 82% were reproduced with a 70% overlap threshold. The cost function with all three cost attributes had the highest coverage percentage at 100% overlap threshold for both methods and this finding showed the correctness of the hypothesis of selecting attributes other than distance and time. All three tests with the B&B method had very low coverage, and Table 4 suggests the reason.

Table 4 shows the consistency of the applied methods and their computational costs. The B&B method did not manage to generate any alternatives for a large percentage of the observations within the time abortion threshold, which resulted in a very low coverage as seen in Table 3. The majority of these observations were longer than 4 km. In few cases the DSGF method also did not generate any alternative routes. This is not necessarily considered a problem as these are relatively short trips and thus not applicable in route choice modelling.

Both the BFS-LE method and the DSGF method performed quite well in relation to the consistency index, where the BFS-LE with the cost function with all the attributes performed the best, followed by the DSGF with road type as an attribute. Not surprisingly, the B&B did not perform well.

The BFS-LE had a very low computational time, while the DSGF had a lot higher run time. Since the DSGF had larger computational costs, the method did not produce the maximum choice set size in some of the cases because of the time abort threshold. Consequently, this resulted in a lower number of unique routes. Since the B&B method did not follow the same restrictions as the other two methods in terms of time abort threshold and maximum choice set size, the method produced very high number of unique routes for some of the observations, and also resulted in high computation time. Consequently, the method was not comparable to the other two and no further results are presented relatively to the B&B.

Table 4. Computational costs and techniques consistency

	Consistency index	Number of unique routes	Number of observations with no alternative	Computational time*
BFS-LE method				
Road type	86.3	15,560	0	0h 5m
Segregated bicycle path	88.2	15,560	0	0h 4m
Land use	86.9	15,560	0	0h 4m
All three cost attributes	89.5	15,560	0	0h 4m
Doubly stochastic generation function				
Road type	88.5	13,603	0	24h 55m
Segregated bicycle path	85.8	13,570	2	23h 30m
Land use	84.8	12,333	7	27h 51m
All three cost attributes	87.3	11,613	7	38h 58m
Branch and bound algorithm				
Basis	54.8	49,911	345	33h 26m
Relaxed	54.5	47,411	347	33h 08m
Restricted	55.6	53,184	343	33h 28m

*Computations performed on an Intel(R) Xeon(R) COU E5-1650 0 @ 3.20GHz with 32 GB RAM running Windows 7 Professional.

To visualize the consistency of the BFS-LE and DSGF methods with respect to the observed behaviour, the distribution of coverage over the cumulative percentage of observations is presented in Figure 2. It can be seen that the methods follow a very similar pattern and do not exhibit a significant difference.

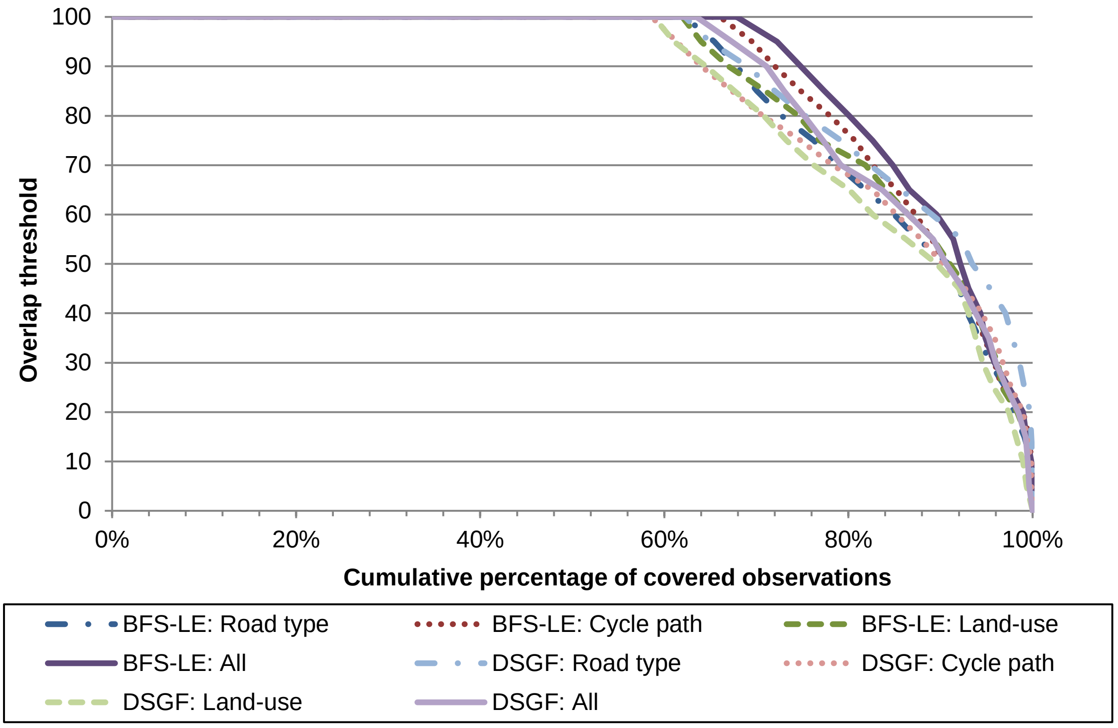

Figure 2. Distribution of coverage over 778 observations.

Figure 3 shows the average of the maximum coverage over the choice sets as a function of the length of the chosen route. The figure shows that there is a general trend, namely that the results are good (and obvious) for shorter routes, but there is a large difference for longer routes. The DSGF performed better up to 10 km, where the average coverage started slightly decreasing.

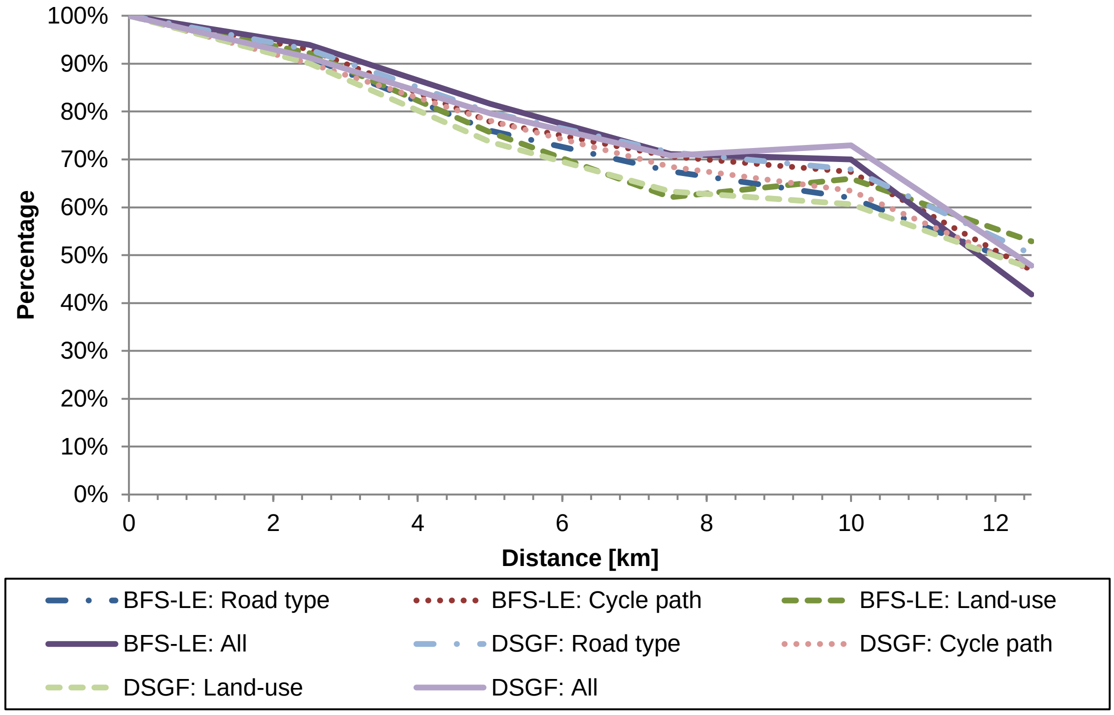

Figure 3. Average maximum coverage over distance.

Figure 4 shows the percentage of completely replicated chosen routes as a function of distance. It can be observed a general trend, namely the number of completely reproduced chosen routes decreases with increasing distance. As intuitively logical, longer routes appear problematic, as routes longer than 8 km were not completely reproduced by any of the choice set generation methods with the exception of one observation getting reproduced by the BFS-LE with the cost function accounting for preferences for specific road types.

Figure 5 shows the distribution of the average path size of all routes in the choice sets generated by the different methods. The path size distribution indicates considerably more diversity between routes generated with the DSGF method than the ones generated with the BFS-LE method.

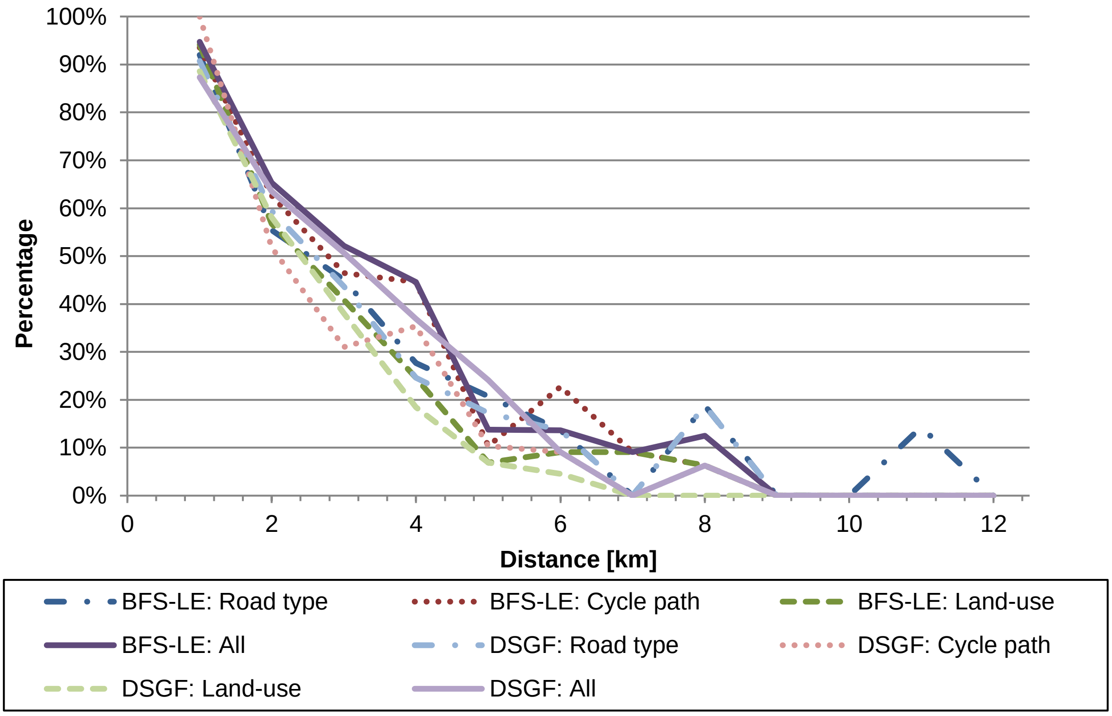

Figure 4. Percentage of chosen route completely replicated as a function of distance.

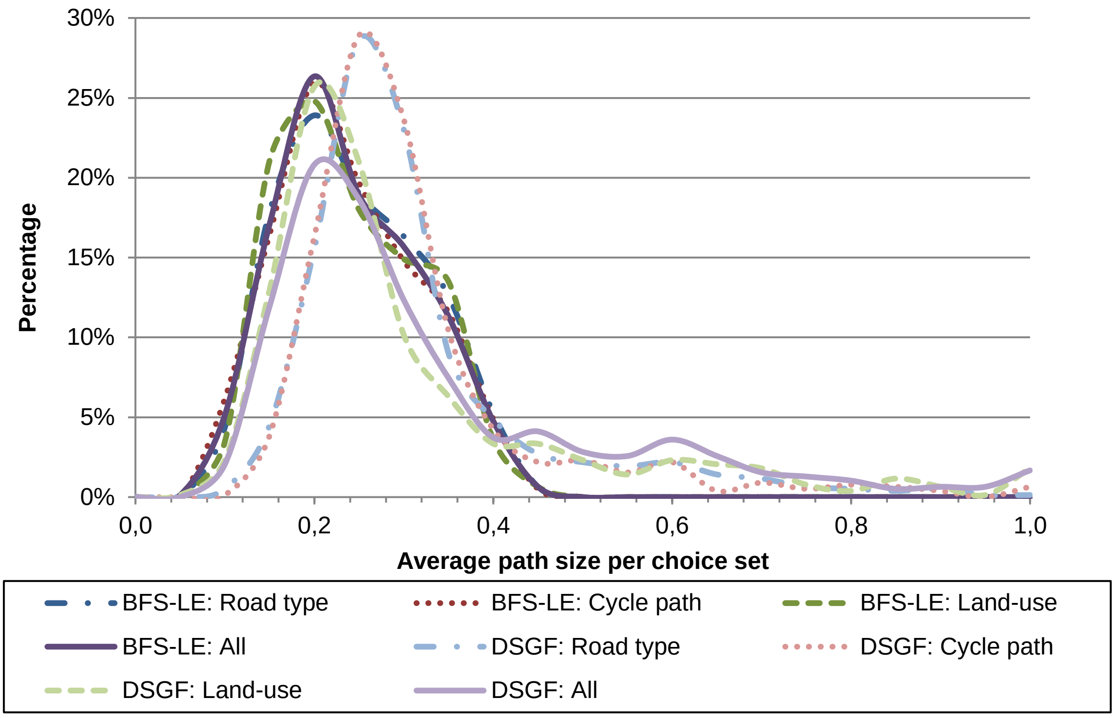

Figure 5. Path size distribution for the resulting route sets.

5. Discussion and conclusions

The current study focused on the efficiency of choice set generation of bicycle routes in a high-resolution network. Bicycle routes were collected with GPS devices and three choice set generation methods were implemented while designing cost functions that included not only distance and time related terms, but also other factors that cyclists would consider relevant, such as scenic routes, dedicated cycle lanes, and road types.

The efficiency of the methods was measured with respected to the observed behaviour of cyclists and the relative ability of generating relevant and heterogeneous alternative routes.

The efficiency of choice set generation methods for bicycle routes revealed similar performances in terms of coverage for the BFS-LE and the DSGF methods, with the BFS-LE method replicating almost 68% of the chosen routes and the DSGF method replicating almost 64%. In perspective, it should be noted that Broach et al. (2010) reproduced only less than one fourth of observed routes, and Hood et al. (2011) only replicated about one-third of observed routes, while using networks with significantly lower number of links (i.e., 3 and 8 times less detailed networks). Clearly, extending the cost function to factors that are relevant to cyclists and that do not pertain only distance and time, but also their preference for scenic routes and safe paths, provided a significant increase in performances when compared to previous findings. Among the cost functions, the fourth function with all three additional cost attributes had the highest coverage percentage at 100% overlap threshold for both methods, which indicated the heterogeneous and complex preference structure for cyclists when considering routes. The B&B method had lower coverage compared to the BFS-LE and the DSGF, as it reproduced approximately 40% of the observed routes and was more in line with previous findings. The problem with this method is the high computational time that did not allow reaching the destination within the time abort threshold for a large percentage of the observations.

When looking at the average maximum coverage over distance, a general trend emerged from the results. Specifically, shorter routes illustrated expected results in having very good coverage for all methods, while longer routes exhibited larger differences across algorithms, with the DSGF method performing better up to 10 km routes and the average coverage decreasing. Hood et al. (2011) found similar results of better coverage with the DSGF for longer observations. Moreover, the number of completely reproduced chosen routes decreased with increasing distance, and routes longer than 8 km were not reproduced.

When looking at the distribution of the average path size of the routes in the choice sets generated with the different methods, the DSGF method produced more heterogeneous alternative routes.

When looking at computational costs, the BFS-LE clearly outperformed the DSGF and the B&B. Hood et al. (2011) found instead that the DSGF (using a uniform probability distribution to draw the random error term) had lower computation time than the BFS-LE in a far less detailed network. A possible reason for this difference is that using a uniform probability distribution might cause the randomly drawn travel cost to deviate substantially from the real travel cost and hence might increase substantially the likelihood of generating unrealistic routes. In the current study, error terms and error components were randomly drawn from a standard normal distribution function while discarding the instances where a negative number was drawn that would explain the longer computation time. A more obvious reason for this difference is that algorithms have been programmed with different resources for diverse stopping criteria, observed routes and considered networks. Two limitations of this study are indeed that a fairer comparison would entail the same programs being developed for the same dataset, and that the coding of the algorithms does not necessarily characterize the definite computation cost, but works only as a proxy. However, these are still relevant results and the better ability in reproducing routes recorded in this study clearly shows the importance of the random draws being from a normal distribution and even more of the cost function being multi-attribute and not only distance and time related.

Results suggest possible directions for further investigation, especially for longer routes. Possible improvements would be to include turn restrictions in the choice set generation, on the line of the turn restrictions constraint included in the original formulation of the B&B method (Prato and Bekhor, 2006). The cost function could also be extended with other attributes considered important for cyclists. Also, the availability of intersection data could contribute to improving further the cost function. Drawing from non-negative distributions (e.g., lognormal, gamma) could help reducing the computation time relating to discarding negative numbers. Last, comparison of model estimation performance and prediction accuracy could be carried out with datasets from the different choice set generation methods.

Acknowledgements

The authors acknowledge the support of Lara Montini at ETH-IVT in Zurich, Michael Balmer at Senozon, and Søren Hasling Pedersen and Jesper Bláfoss Ingvardson at DTU Transport. The authors gratefully acknowledge the financial support of the Danish Road Directorate to the project "The effect of cycling policies". The authors are thankful to the anonymous reviewers who provided insightful comments to the initial version of this paper.

References

Antonakos, C.L. (1994). Environmental and travel preferences of cyclists. *Transportation Research Record*, 1438, 25-33.

Aultman-Hall, L., Hall, F.L. and Baetz, B.B. (1997). Analysis of bicycle commuter routes using Geographic Information Systems: implications for bicycle planning. *Transportation Research Record*, 1578, 102-110.

Axhausen, K.W. and Smith, R.L. (1986). Bicyclist link evaluation: a stated preference approach. *Transportation Research Record*, 1085, 7-25.

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

Bekhor, S. and Prato, C.G. (2009). Methodological transferability in route choice modelling. *Transportation Research Part B*, 43, 422-437.

Ben-Akiva, M.E., Bergman, M.J., Daly, A.J. and Ramaswamy, R. (1984). Modelling inter urban route choice behavior." *Proceeding of the Ninth International Symposium on Transportation and Traffic Theory*, VNU Science Press, Delft, Netherlands, pp. 299-330.

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

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

Bovy, P.H.L. and Bradley, M.A. (1985). Route choice analyzed with stated preference approaches. *Transportation Research Record*, 1037, 11-20.

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

Broach, J., Gliebe, J.G. and Dill, J.L. (2010). Calibrated labeling method for generating bicyclist route choice sets incorporating unbiased attribute variation. *Transportation Research Record*, 2197, 89-97.

Broach, J., Gliebe, J.G. and Dill, J.L. (2011). Bicycle route choice model developed using revealed preference GPS data. Proceedings of the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Broach, J., Gliebe, J.G. and Dill, J.L. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A*, 46, 1730-1740.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In Lesort, J.B. (Ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, pp. 697–711.

de la Barra, T., Perez, B. and Anez, J. (1993). Multidimensional path search and assignment. *Proceedings of the 21st PTRC Summer Meeting*, Manchester, UK, pp. 307-319.

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

Hood, J., Sall, E. and Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters: The International Journal of Transport Research*, 3, 63-75.

Hoogendoorn-Lanser, S., van Nes, R. and Bovy, P.H.L. (2006). A rule-based approach to route choice set generation. Presented at the *11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Hopkinson, P. and Wardman, M. (1996). Evaluating the demand for new cycle facilities. *Transport Policy*, 3, 241-249.

Howard, C. and Burns, E.K. (2001). Cycling to work in Phoenix: route choice, travel behavior, and commuter characteristics. *Transportation Research Record*, 1773, 39-46.

Hunt, J.D. and Abraham, J.E. (2007). Influences on bicycle use. *Transportation*, 34, 453-470.

Hyodo, T., Suzuki, N. and Takahashi, K. (2000). Modeling of Bicycle route and destination choice behavior for bicycle road network plan. *Transportation Research Record*, 1705, 70-76.

Krizek, K.J. (2006). Two approaches to valuing some of bicycle facilities' presumed benefits. *Journal of the American Planning Association*, 72, 309-320.

Marchal, F., Hackney, J. and Axhausen, K.W. (2005). Efficient map matching of large Global Positioning System data sets: tests on speed-monitoring experiment in Zürich. *Transportation Research Record*, 1935, 93-100.

Menghini, G., Carrasco, N., Schüssler, N. and Axhausen, K.W. (2010). Route choice of cyclists in Zurich. *Transportation Research Part A*, 44, 754-765.

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

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

Prato, C.G. and Bekhor, S. (2007). Modeling route choice behavior: how relevant is the composition of choice set? *Transportation Research Record*, 2003, 64-73.

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

Rasmussen T.K., Ingvardson, J.B., Halldórsdóttir, K. and Nielsen, O.A. (2013). Enhancing the use of Global Positioning Systems data from wearable GPS devices as travel survey method: a case study in the Copenhagen Region. Presented at the *2nd Symposium of the European Association for Research in Transportation*, Stockholm, Sweden.

Rieser-Schüssler, N., Balmer, M. and Axhausen, K.W. (2012). Route choice sets for very high-resolution data. *Transportmetrica*, 9, 825-845.

Schüssler, N. and Axhausen, K.W. (2008). Identifying trips and activities and their characteristics from GPS raw data without further information. Presented at the *8th International Conference on Survey Methods in Transport*, Annecy, France.

Schüssler, N. and Axhausen, K.W. (2009a). Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT). Working Paper, 568, IVT, ETH Zurich, Zurich.

Schüssler, N. and Axhausen, K.W. (2009b). Processing GPS raw data without additional information. *Transportation Research Record*, 2105, pp. 28-36.

Sener, I.N., Eluru, N. and Bhat, C.R. (2009). An analysis of bicycle route choice preferences in Texas, US. *Transportation*, 36, 511–539.

Shafizadeh, K. and Niemeier, D. (1997). Bicycle Journey-to-work: travel behavior characteristics and spatial attributes. *Transportation Research Record*, 1578, 84-90.

Stinson, M.A. and Bhat, C.R. (2003). Commuter bicyclist route choice: analysis using a stated preference survey. *Transportation Research Record*, 1828, 107-115.

Stinson, M.A. and Bhat, C.R. (2005). A comparison of the route preferences of experienced and inexperienced bicycle commuters. Proceedings of the *84th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Stopher, P.R. (2009). Collecting and processing data from mobile technologies. In Bonnel, P., Zmud, J., Lee-Gosselin, M.E.H. and Madre, J.L. (Eds.), *Transport Survey Methods: Keeping Up with a Changing World*, Emerald, Bingley, pp. 361–391.

Tilahun, N.Y., Levinson, D.M. and Krizek, K.J. (2007). Trails, lanes, or traffic: valuing bicycle facilities with an adaptive stated preference survey. *Transportation Research Part A*, 41, 287-301.

Tsui, S.Y.A. and Shalaby, A.S. (2006). Enhanced system for link and mode identification for personal travel surveys based on Global Positioning Systems. *Transportation Research Record*, 1972, 38-45.