EJTIR

Issue 18(2), 2018
pp. 158-177
ISSN: 1567-7141
http://tlo.tbm.tudelft.nl/ejtir

Modelling bicycle route choice using data from a GPS-assisted household survey

Muhammad Ghanayim[1]
Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Israel.

Shlomo Bekhor[2]
Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Israel.

This paper considers bicycle route choice for commuter trips. Bicycle route preferences are analysed using a dataset from a GPS-assisted household travel survey conducted in the Tel Aviv metropolitan area. Different choice set generation methods were applied to generate alternative routes for each observation, and the matching with the actual route is discussed. Model estimation is performed for different route choice sets to test the sensitivity of the parameter estimates. The results obtained are quite consistent, and indicate an expected tendency to ride in longer routes, but with separated bike lanes. In the absence of such lanes, riders prefer to use local streets and avoid riding on busy arterial streets and highways.

Keywords: *Bicycle route choice, set generation, GPS-based travel survey.*

1. Introduction

1.1 Motivation

There has been an increasing interest in bicycle network planning, in particular as a sustainable transportation mode in urban areas. In several European cities, bicycle is treated as an essential transportation mode. In contrast, transportation planning studies conducted in Israel consider bicycle as part of non-motorized modes, often joined with pedestrian. An effective planning requires a clear understanding of the preferences of its users. Therefore, this paper focuses on bicycle route choice in urban areas, which is an essential planning component.

Most existing studies on bicycle route choice models assumed a two-stage approach, meaning that first a choice set is generated (using one or a combination of methods), followed by model estimation using the generated choice set. The need to generate a representative choice set is challenging, because of the very large number of possible alternative routes in a dense network.

There are several methods to generate a route choice set, most of them developed for motorized travel modes. These methods are generally based on variations of the shortest path search. For example, the labelling approach, which minimizes generalized cost functions according to link attributes (Ben-Akiva et al., 1984); link penalty, which gradually increases the impedance of all links on the shortest path (De la Barra et al., 1993); link elimination, which removes shortest paths from the network in sequence to generate new routes (Azevedo et al., 1993); simulation method, which produces alternative paths by drawing link impedances from probability distributions (Bekhor et al., 2006). These and other approaches are relatively well known in the literature (Prashker and Bekhor, 2004; Prato, 2009).

[1] A: Technion City, Haifa, 32000, Israel T: +972 4 829 2901 F: +972 4 822 5716 E: smfgmfg9@campus.technion.ac.il
[2] A: Technion City, Haifa, 32000, Israel T: +972 4 829 2460 F: +972 4 822 5716 E: sbekhor@technion.ac.il

1.2 Bicycle route choice studies

Broach et al. (2012) applied different methods to generate routes in Portland, such as labelling method, link penalty, and simulated shortest paths. The coverage results reported in the paper indicated that only 25% of the generated routes replicated the actual routes, and 42.3% coverage assuming 80% overlap. They found that cyclists chose a route that is 11% on average longer than the shortest route for commute trips.

Hood et al. (2010) applied the simulation method to generate routes (the doubly stochastic method) in San Francisco. In contrast to other studies, they added to the cost function environmental and socio-economic variables, such as crime record areas, lighting and weather conditions. However, as with Broach et al. (2012), coverage rates were very low (approximately 30%).

Menghini et al. (2009) used the link elimination method to generate routes in Zurich. The cost function used was limited to travel distance. In Halldórsdóttir et al. (2014) study in Copenhagen, cost functions included land use and road type. They added a random component to these values to create a representative route choice set. The coverage results (70%) were superior in comparison to the previous studies.

In the above studies, planning road networks were used as a basis for route choice set generation. These networks did not include small local streets with low density traffic (as is often the case for planning networks). This is because of the need to have a simpler network for planning purposes. Such a planning network is not suitable for bicycle route set generation, precisely because there is an advantage in these roads for cycling (related to the small traffic volume in these streets).

In Los Angeles (2014), a detailed transportation network was used to generate routes. Sensitivity tests indicated that a choice set consisting of five bicycle paths appears to be a practical size of choice set. The Cross-Nested Logit (CNL) choice model was used for model estimation.

Recently, Chen and Shen (2016) examined built environment factors (land use characteristics and roadway design features) that affected cyclists' route choices in Seattle. Data was collected from a smartphone application, and the labelling approach combined with the k-shortest path method was applied to generate the route choice set for the 543 observations. The results indicated that cyclists preferred bicycle routes of shorter distances, lower speed limits, and flatter roads. Furthermore, routes surrounded by mixed land use, waters or parks were preferred.

Choice set composition is known as a difficult issue to resolve in the context of modelling revealed choice behaviour. Biased model parameters and statistical inconsistency of parameter estimates are related to mis-specification of the choice set. There are recent modelling approaches that avoid the bias induced by generated choice sets. Frejinger et al. (2009) developed a correction term for route sampling, using random walk as the generation method. Fosgerau et al. (2013) proposed the recursive logit (RL) model where path choice is modelled as a sequence of link choices using a dynamic discrete choice framework. The RL model was extended to the Nested RL by Mai et al. (2015). Recently, Zimmerman et al. (2017) estimated bicycle route choice models without the need to generate routes a priori. The estimation is based on the recursive logit model developed by Fosgerau et al. (2013).

Accuracy and efficiency of model estimates for MNL and Mixed Logit have been tested empirically by reducing randomly the size of a synthetic dataset (Nerella and Bhat, 2004). Recently, Faghih-Imani and Eluru (2017) examined the impact of sample size on hourly usage and users' destination choice preferences employing data from New York City's CitiBike. Both studies show that parameter estimates significantly change with respect to the sample size.

1.3 Objectives

In most studies described above, the sample used as basis for model estimation was not representative of the entire population. This is because the studies collected data from specific

population segments (e.g. students). The present paper investigates bicycle route choice from a general purpose survey, which includes bicycle riders from different areas and socio-economic characteristics.

This paper investigates bicycle route choice characteristics, using a dataset from a GPS-assisted household survey and mapped to a detailed urban network. The paper estimates bicycle route choice models, taking into account different network characteristics and land use variables. In contrast to previous papers, this is one of the first attempts to use data from a general-purpose household survey for bicycle route choice modelling. Since we are interested in the analysis of the choice set composition, the present paper will first generate routes, and use the generated routes for model estimation.

2. Methodology

This section describes the methodology and the research steps, consisting of: (i) survey and dataset description, (ii) application of the path generation techniques to the urban network, (iii) choice set composition with special emphasis in defining the overlap measure, and (iv) route choice models estimated.

2.1 Survey Data

A household survey was carried out in 14 major cities of the Tel Aviv metropolitan area between December 2013 and June 2014. Only weekdays were surveyed. The survey method was similar to the Jerusalem household travel survey (Oliveira et al., 2011), and it is briefly outlined as follows.

The survey was composed of two main phases, both of them with surveyor in-home visits. The first (recruiting) phase, in which a surveyor visited the household, collected general information and provided household members older than 14 with a GPS data logger. They were instructed to carry the device on their pockets or purses for 24 consecutive hours. The GPS tracker has enough battery for 24 hours.

In the second phase, the interviewer returned to the household to retrieve the GPS readings and complete the questionnaire about the activities recorded by the GPS logger. The GPS data was retrieved with the use of laptop computers, which assisted the surveyor and the household members to identify their trips and activities.

The overall sample included 8,515 persons living in 2,896 households (average household size 2.94). On average, there are 1.04 cars per household. A total of 39,952 trips were recorded, an average of 4.7 trips per person. The trips are distributed by main mode as follows: 49.7% car (either driver or passenger), 35.4% walk, 11.6% public transport, 1.7% motorcycle and only 1.6% bicycle (618 trips). Table 1 presents selected characteristics of bicycle riders in comparison to car users in the sample.

Table 1. Tel Aviv household survey 2014 - Selected individual characteristics

	Bicycle riders	Car users
Average household size	3.6	2.9
Motorization rate	290	354
Average age	33	47
% over age 40	30%	63%
Male proportion	73%	61%
Female over age 40	7%	23%

The table shows significant differences between bicycle and car users. Clearly, the difference in the average age is related to the fact that driver's license is issued only to those over the age 17

years, and also because bicycle riders characterize younger population. Note that the proportion of women over the age of 40 traveling by bicycle is much lower in comparison to car users. These results are similar to those reported by Plaut (2005), which analysed bicycle commuters in the US.

The GPS device recorded the position every 3 seconds on average, and the raw data file included over 45 million readings. After performing logical checks and deleting observations with gross errors in GPS, there were 151,392 points related to 545 bicycle trips performed by 221 persons. The logical checks included the time that elapsed between two consecutive data points, and the correspondent speed. The GPS points were map matched to a detailed network of the Tel Aviv metropolitan area, which contains 92,670 nodes and 127,053 links, with a total length of 8384 km. The network data source is from a private company (Mapa), and the matching was performed using the ArcGIS software. Figure 1 displays the matched GPS points to the map.

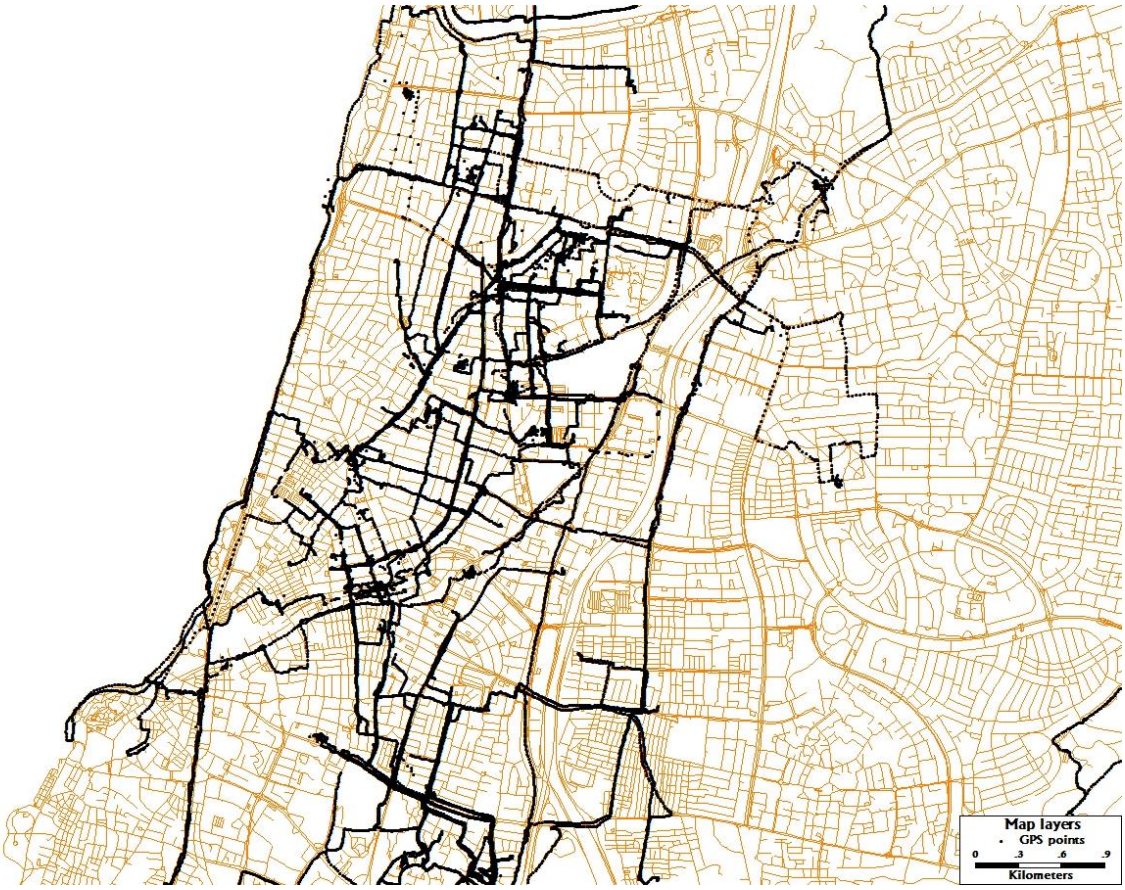

Figure 1. Tel Aviv area – GPS track points of bicycle trips

2.2 Path Set Generation

The methods used to generate routes in this paper are variants of selective generation methods using shortest path algorithm, which may include bias in the estimation results (Frejinger et al., 2009). Special care is placed in generating representative routes to form the choice set for model estimation.

Path generation techniques were applied to the detailed network of the Tel Aviv metropolitan area. 2% of the network roads were omitted, corresponding to freeway links, in which bicycle ride is prohibited. Bikeway facilities (dedicated lanes, bikeways and walking paths that also allow cycling) comprise 3.8% of the Tel Aviv city network. Although this is a relative small proportion, it is significantly higher in comparison to only 0.8% on average for the neighbouring cities.

In this paper, bicycle routes are generated by three main methods: link elimination, link penalty, and simulation method. All these methods essentially change one or more network variables and calculate the correspondent shortest route. The following paragraphs describe the assumptions of each method.

Link elimination method
The link elimination method is applied by successively performing the following procedure: (i) calculation of the shortest path; (ii) deleting a link from the shortest path; (iii) computation of the next shortest path and comparison with the existing routes in the path set. If the generated route is similar to a route in the path set, then an additional link of the shortest path is deleted from the network. In total, up to 10 different routes were generated for each origin-destination pair.

Link penalty method
The link penalty method is applied by successively performing the following procedure: (i) calculation of the shortest path; (ii) increasing travel times on the shortest path links by a factor of 1.10; (iii) computation of the next shortest path and comparison with the existing routes in the path set. In total, up to 5 different routes were generated for each origin-destination pair.

Simulation method
The simulation method is implemented by calculating the shortest path for each draw of link impedances from a log normal distribution (to prevent negative values), with mean equal to the actual link length and variance equal to 50% of the mean. For each generated route, the program compared it to the previously generated routes in the set. In case the generated route was approximately (80%) equal to a route in the choice set, it was not added to the set.

Note that if a strict tolerance is imposed (that is, only removing routes that overlap completely with previous routes), the resulting set will include many similar routes. On the other hand, if only disjointed routes (no links in common) are included, the resulting choice set will be relatively small. The best results were obtained assuming 80% tolerance.

All three methods above were independently run in the network. After removing routes that overlap more than 80% with other routes in the choice set, a total of 20 routes per each origin-destination pair formed the path set for model estimation.

2.3 Choice Set Composition
The objective of the choice set formation is the maximization of the coverage of the collected routes and the consequent composition of choice sets behaviourally consistent with the observed behaviour. Choice sets may correspond to path sets generated by single methods with good performance indexes, or to the combination of path sets produced by different methods with poor individual performances.

The following measures were discussed in Prato and Bekhor (2007) and in more details by Halldórsdóttir et al. (2014) and are presented here for completeness. The coverage measures the percentage of observations for which a path generation technique reproduces the actual behaviour according to a certain overlap threshold, which expresses the degree of similarity between generated and collected routes.

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

where I(•) is the coverage function, equal to one when its argument is true and zero otherwise, O_{nr} is the overlap measure for technique r and observation n, δ is the overlap threshold.

$$O_{nr} = \frac{L_{nr}}{L_n} \tag{2}$$

Where L_{nr} is the overlapping length between generated and observed routes, L_n is the length of the observed path for driver n.

The index of behavioural consistency compares a path generation method to the ideal algorithm that would replicate link-by-link all the routes reported in the survey, with a 100% resulting coverage for a 100% overlap threshold.

$$CI_r = \frac{\sum_{n=1}^{N} O_{nr,max}}{N \cdot O_{max}} \tag{3}$$

Where CI_r is the consistency index of algorithm r, $O_{nr,max}$ is the maximum overlap measure obtained with the paths generated by algorithm r for the observed choice of each driver n, O_{max} is the 100% overlap over all the N observations for the ideal algorithm.

The overlap measure indicated in Equations (2) and (3) accounts for the physical common path length. For bicycle route choice, it is important to account also for other path characteristics, such as the overlap in specific bikeways or other dedicated places. Therefore, a generalization of the overlap measure O_n (the technique r is omitted for simplicity) is proposed in the following equation:

$$O_n = \max_{i=1,I} \sum_{k=1}^{K} \alpha * \left(1 - \left|\frac{M_{nki} - M_{nk}}{\max(M_{nk}, M_{nki})}\right|\right) + (1 - \alpha) * O_{ni} \tag{4}$$

Where M_{nki} represents the measure of characteristic k for alternative route i of observation n. M_{nk} is the measure of characteristic k for the chosen route of observation n. O_{ni} is the overlap measure of generated route i of observation n (similar to equation 2). K is the number of characteristics and I is the number of routes in the choice set. The parameter α ($0 \leq \alpha \leq 1$) indicates a weighting factor between the two overlap measures. In this paper, we assume that α is equal to 0.5.

2.4 Model Specification

Model specifications that account for correlation among alternatives are preferable than the Multinomial Logit (MNL) model to represent route choice behaviour. MNL modifications, such as C-Logit (Cascetta et al., 1996) and Path Size Logit (PSL), include a correction term in the deterministic part of the utility function and maintain a simple logit structure (Ben-Akiva and Bierlaire, 1999; Ramming, 2001). Generalized Extreme Value (GEV) specifications, such as Cross-Nested Logit (CNL) and Generalized Nested Logit (GNL), relate the network topology to model parameters in the stochastic term of the utility function and present a more complex structure (Bekhor and Prashker, 2001). Probit and Logit Kernel (LK) assume that the covariance of path utilities is proportional to overlap lengths (Yai et al., 1997; Bekhor et al., 2002).

In line with previous papers, this paper estimates models using 3 simple model forms: the multinomial logit (MNL), C-logit, and Path Size-logit (PSL). C-Logit and PSL maintain the MNL structure and present very similar functional forms, but each model interprets differently the correction term that measures the degree of similarity of each route with respect to all the other routes in the choice set (Prashker and Bekhor, 2004). In the C-Logit, the commonality factor indicates that the utility of a path must be reduced because of the similarity with other routes, and is always greater or equal to one, as can be seen in the following equation (Cascetta et al., 1996):

$$CF_k = ln\left[1 + \sum_{\substack{l \in C_n \\ k \neq l}} \left(\frac{L_{kl}}{\sqrt{L_k L_l}}\right)\left(\frac{L_k - L_{kl}}{L_l - L_{kl}}\right)\right] \quad (5)$$

Where CF_k is the commonality factor of route k, L_k is the length of route k, L_l is the length of each route l in the choice set C_n; L_{kl} is the common length between routes k and l.

In the PSL, the path size presents the fraction of the path that constitutes a "full" alternative, and is always less or equal to one. This paper calculates the path sizes are calculated according to the following formulation (Bovy et al., 2008):

$$PS_k = \sum_{a \in \Gamma_k} \frac{L_a}{L_k} \frac{1}{\sum_{l \in C_n} \left(\frac{L_k}{L_l}\right)^{\gamma} \delta_{al}} \quad (6)$$

where PS_k is the path size of route k, Γ_k is the set of links belonging to route k, L_a is the length of link a, δ_{al} is the link-path incidence dummy (equal to one if route l uses link a, and zero otherwise), γ is a positive parameter that accounts for different size contributions due to routes with different lengths.

Since the dataset is composed of 545 trips performed by 221 individuals, the choice data is of panel type. Therefore, the mixed logit versions of the MNL, C-Logit and PSL models are used as basis for model estimation. The detailed network representation enables the testing of additional variables to the route length. After performing several tests, the utility function for an individual n choosing an alternative k is specified as follows:

$$u_{nk} = \beta_1 Length_k + \beta_2 Length\_A_k + \beta_3 Length\_C_k + \beta_4 Street_k + \beta_5 Dwell_k + \beta_6 Near\_Sea_k + \beta_7 Near\_Park_k + \xi_n \quad (7)$$

Where:

$Length_k$ - Total length of route k (km)

$Length\_A_k$ - Length of route k in category A streets (bike paths) (km)

$Length\_C_k$ - Length in route k in category C streets (urban arterials and highways) (km)

$Street_k$ – Average street length, defined as the ratio between the route length and the number of intersections (m)

$Dwell_k$ – Number of dwelling units per meter along the streets of route k

$Near\_Sea_k$ – Length of route k passing along or up to 100 m of the sea shore (km)

$Near\_Park_k$ – Length of route k passing along or up to 100 m of a green land use (km)

ξ_n – random variable with zero mean and standard deviation σ_n

Note that most variables interact with the route length, which increases the correlation of the coefficients. The variables included in the utility function exhibit correlation less than 0.6.

3. Results

3.1 Choice Set Generation

As indicated in the methodology section, a total of 20 routes were generated for each observation between origin and destination. The first quality test for the choice set routes is the coverage,

since the actual (observed) route may not be exactly replicated by the 20 routes generated. Figure 2 shows the results of the coverage for different overlap thresholds.

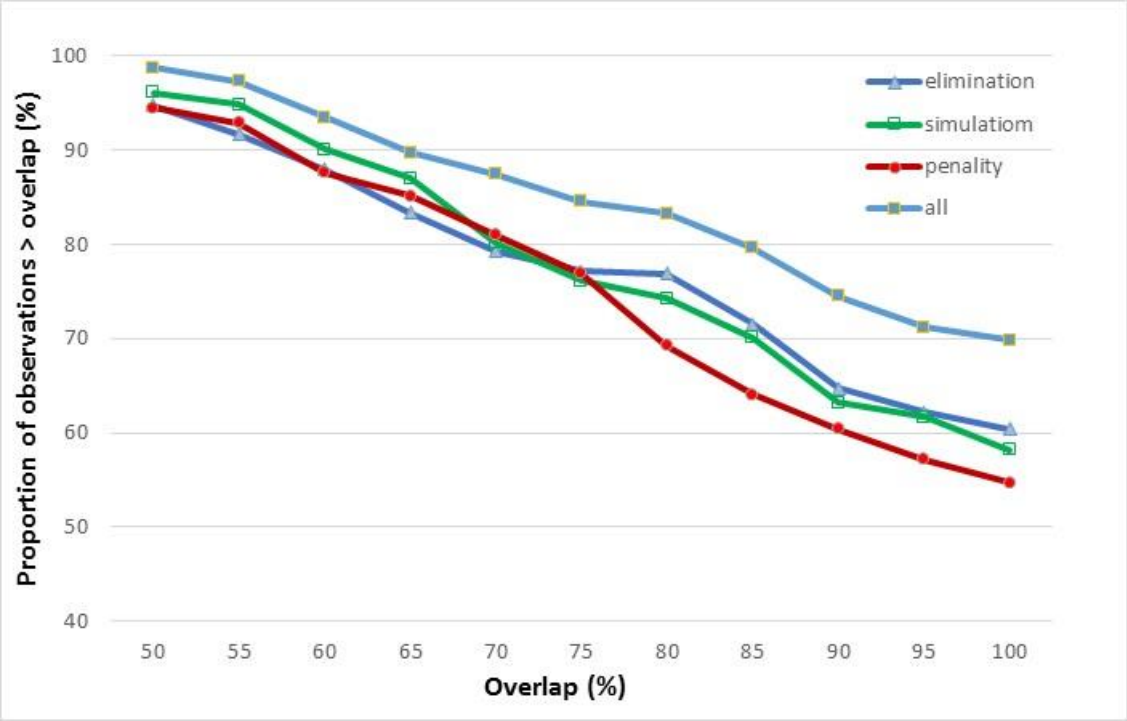

Figure 2. Coverage distribution for different overlap thresholds

The overall coverage for the 3 methods together, for an overlap threshold of 80% of the route length, is 83%. If a strict threshold of 100% overlap is imposed, the coverage drops to 70%. It is noticeable that with only 5 routes generated by the simulation method, the coverage is superior in comparison to both link penalty and link elimination methods. Figure 3 shows an example of the generated routes for a specific origin-destination pair. The dark points correspond to the actual chosen route, along the seashore.

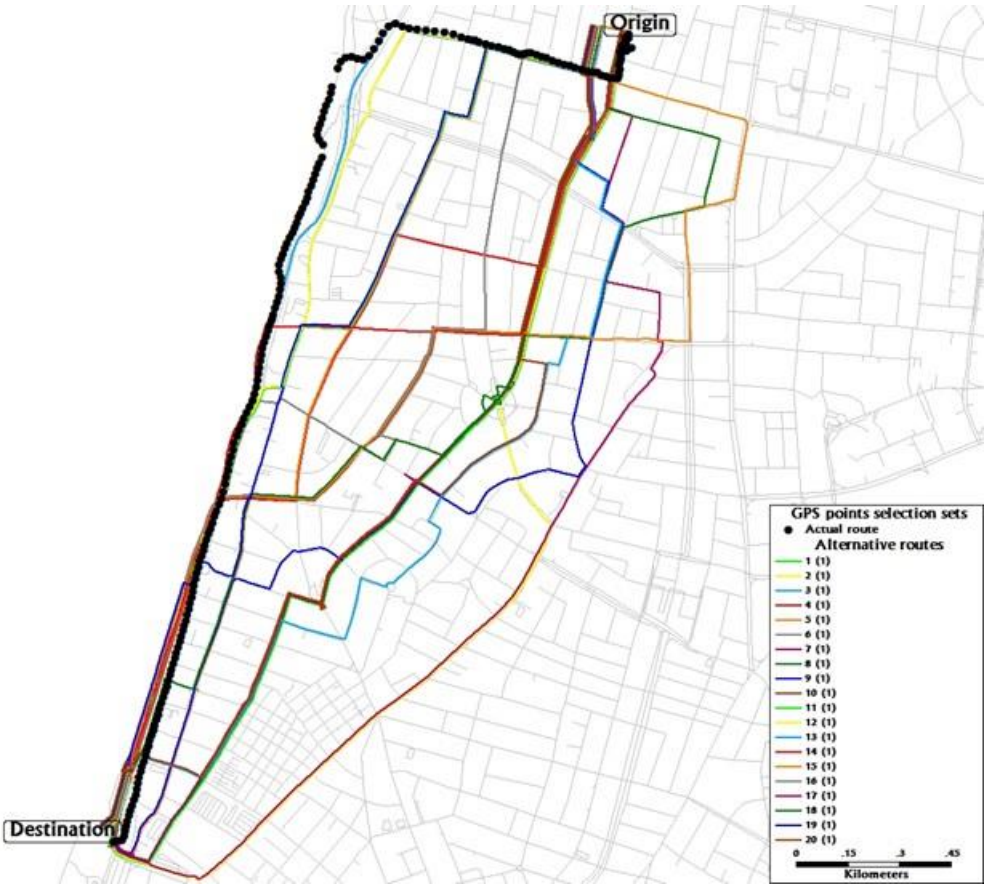

Figure 3. Route choice set example

As discussed in Halldórsdóttir et al. (2014), there is a trade-off between the number of generated routes (in terms of computational costs) and the coverage. Figure 4 shows the overlap progress with the least matched route in the dataset as a function of the number of generated routes. As can be seen, a choice set of 20 routes are sufficient to reach 80% overlap for all actual routes in the dataset.

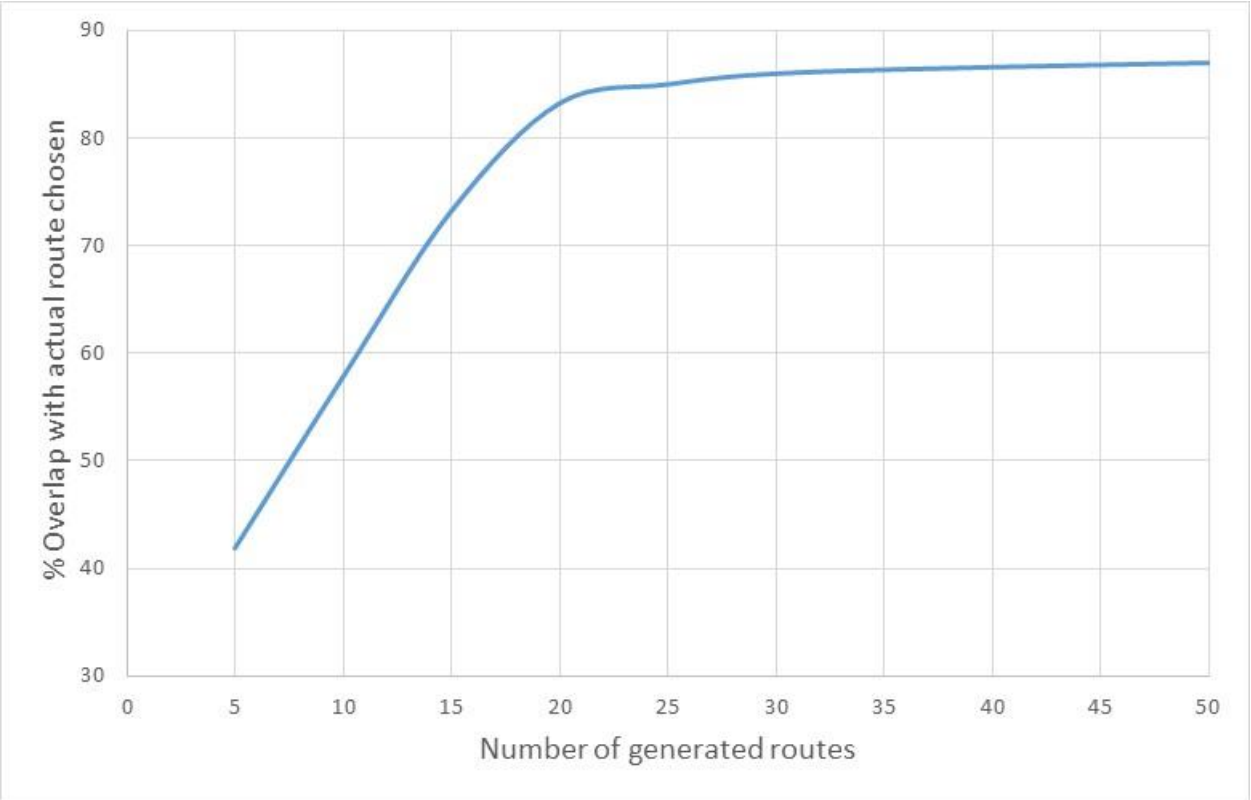

Figure 4. Overlap progress with respect to the number of generated routes

3.2 Choice Set Composition

Ideally, the choice set generation procedure should be able to cover 100% of the actual chosen routes. In practice, even after generating a large number of alternative routes, the coverage will not be perfect. In addition, it is important to analyse the composition of the choice set not only with respect to the overall length, but also with respect to other relevant attributes.

In contrast to previous papers that defined overlap as the route with the maximum common length (equation 3), this paper uses an additional definition (equation 4) that checks the overlap not only with respect to the overall length, but also with respect to other route attributes. After performing several tests in the dataset with a smaller number of observations, the attribute that added most to the overlap explanation is the length of bike paths.

Figure 5 compares the overall coverage between the two methods: the standard method (based on equation 3) and the proposed method (based on equation 4). Similarly, Figure 6 compares the coverage between two methods for the bike path sections of the chosen routes.

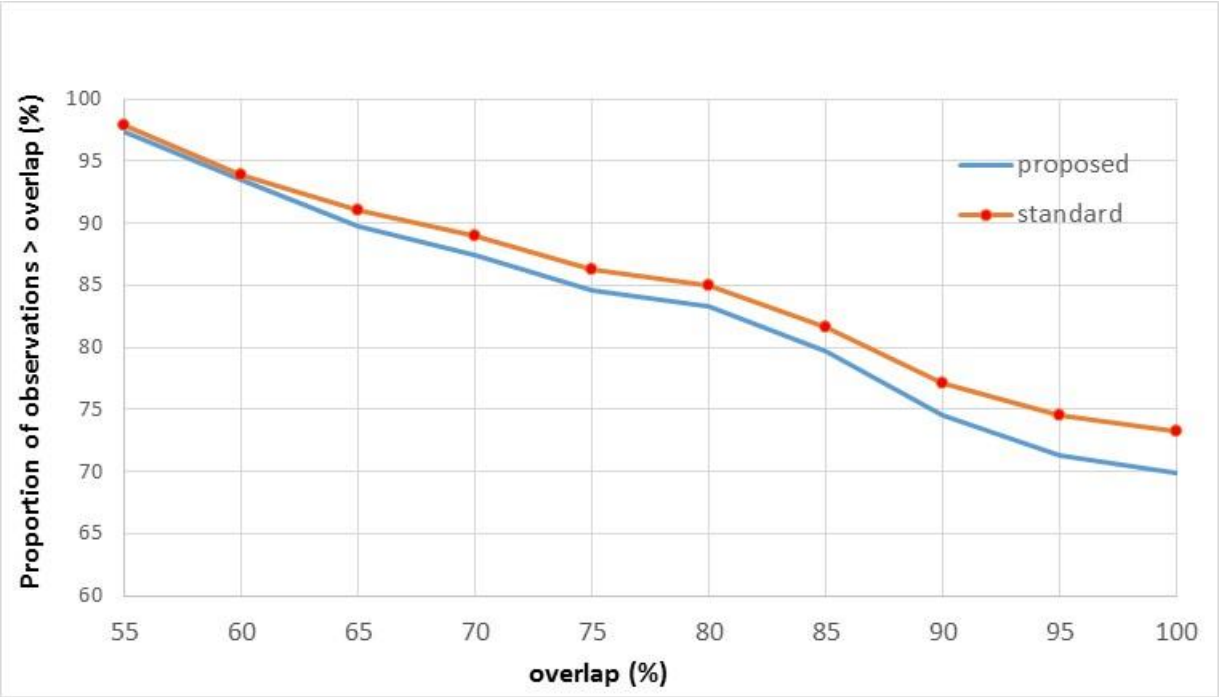

Figure 5. Comparison of the coverage between the two methods

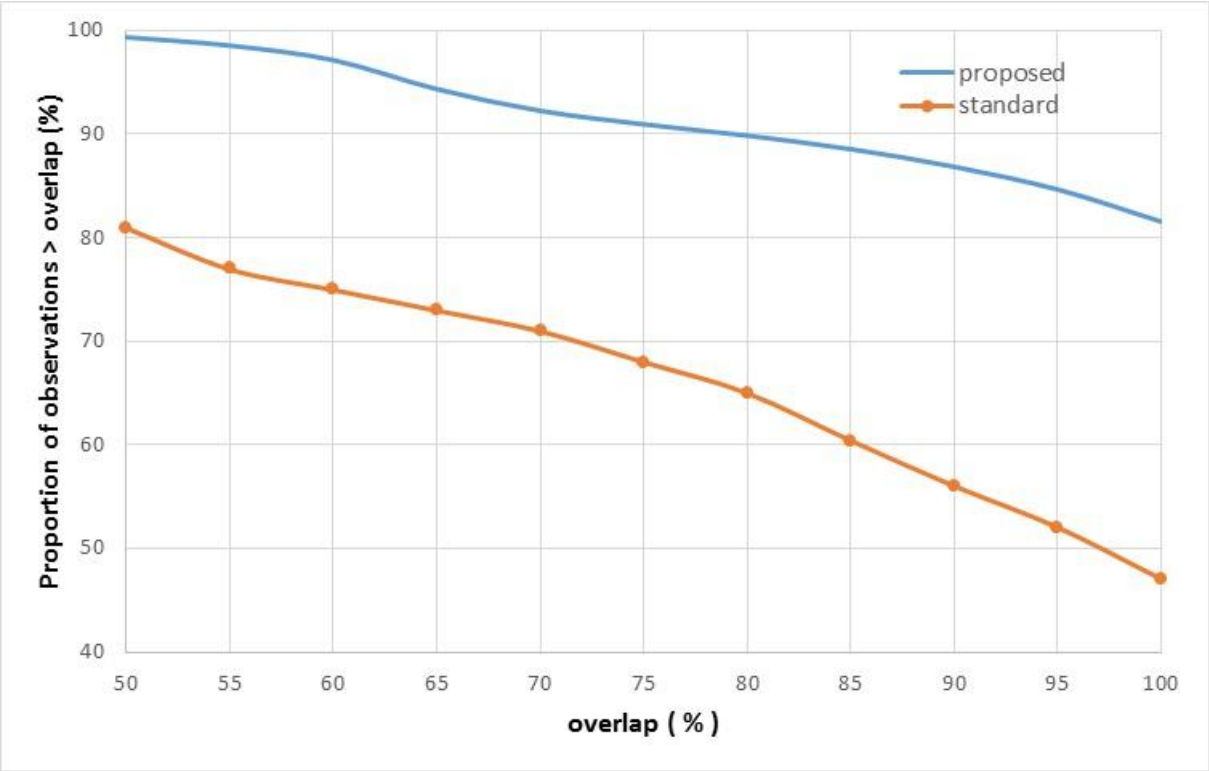

Figure 6. Comparison of the coverage between the two methods – bike path sections

As shown in Figure 5, it is clear that using an additional variable to measure overlap will decrease the overall coverage (in terms of total length). However, the small coverage decrease is largely compensated by the bike path overlapping, as can be seen in Figure 6. Therefore, the

proposed method (equation 4) to measure overlap is more appropriate to evaluate the quality of the choice set (in this case, in terms of the coverage in bike paths).

3.3 Deviation from Shortest Path

According to the literature, commuter cyclists prefer to ride in dedicated lanes, despite of the increase in the route distance. Table 2 shows the ratio of the actual route with the shortest route, and a comparison with results found in the literature.

Table 2. Average ratio of actual route and the shortest route

	Overall sample	Regions with bicycle lanes
Tel Aviv survey 2014	1.13	1.18
Winters et al. (2011) (Vancouver)	1.09	
Broach et al. (2012) (Portland)	1.11	
Mekuria et al. (2012) (San Jose)	1.11	1.16
Hyodo et al. (2000) (Japan)	1.08	

The results presented in Table 2 indicate that the ratio found in the Tel Aviv survey is close to results reported in the literature. Further inspection on the results indicate that this ratio increases to 1.18 in the city of Tel Aviv, which has a higher proportion of bicycle facilities relative to other cities, for which the ratio is equal to 1.07. The above result is consistent with findings by Mekuria et al. (2012).

3.4 Model Estimation Results

The paper reports the model with the explanatory variables that were found significant at the 0.05 level. Variables that were not significant in most runs were not included in the estimation results presented here. Table 3 presents the results for 3 Mixed Logit models (with respectively MNL, PSL and C-Logit as the choice model structure) with 20 alternative routes for each observation. The Mixed Logit model parameters were estimated using Biogeme software with 1,000 Halton draws (Bierlaire, 2003). The explanatory variables are defined in model specification section above.

Table 3. Model estimation results – 20 alternative routes

Variable	Mixed MNL		Mixed PSL		Mixed C-Logit	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Total route length (km)	-6.83	-5.2	-8.04	-14.2	-10.0	-13.8
Route length in category A (km)	1.9	2.1	2.23	7.3	2.78	5.1
Route length in category C (km)	-5.1	-1.7	-4.59	-9.3	-5.9	-6.6
Average street length (m)	0.068	2.1	0.0543	9.0	0.082	2.1
Dwelling units / m	1.21	2.4	2.12	10.7	2.74	2.1
Route length "near sea" (km)	0.56	2.1	1.28	7.5	1.65	4.4
Route length "near park" (km)	2.14	1.3	1.72	5.1	2.24	3.1
PS factor			1.77	9.3		
C-Logit factor					-0.33	-3.7
Sigma (σ)	2.24	4.8	1.14	5.6	1.28	6.34
Hit-ratio	22%		69%		61%	
Initial Log-likelihood	-1420.5					
Final Log-likelihood	-1087.0		-970.5		-1060.0	
Likelihood Ratio	367		900		721	
Rho-bar squared	0.235		0.316		0.253	
Number of observations	545		545		545	
Number of individuals	221		221		221	

As expected, the results presented in Table 3 above show that both Mixed PSL and Mixed C-logit outperform the Mixed MNL model. Note that the Mixed Logit structure accounts only for the panel data and not correlation across alternatives. The PSL and C-Logit models take into account the overlap between routes. The PS correction factor (equation 6) is superior in comparison to the commonality factor (equation 5), and this explains the better performance of the PSL in comparison to the C-Logit.

In line with previous studies, the estimation results show that on average, bicycle riders prefer to ride on shorter routes (recall that the sample is mostly composed of commuter trips). However, riders are willing to extend their trip in order to ride on bike paths (as indicated from the positive coefficient of the category A variable). In contrast, cyclists avoid riding on busy streets with large car traffic (as indicated by the category C variable).

Two variables that are related to the building environment were found significant: the average street length and the number of dwelling units. The first variable indicates that cyclists prefer riding on longer streets, in the sense that there are fewer junctions. The second variable indicates that cyclists prefer to ride along routes passing through residential areas, which are mostly composed of local streets.

Two additional variables found significant are related to a pleasant environment: "near sea" and "near park" variables. These variables indicate the tendency to ride along the seashore or riding close to parks. Similar to the category A variable (bikeways), these variables indicate the tendency to extend the ride in order to ride in pleasant regions.

Note that a variable indicating slope or gradient was not found significant in the present study, which is surprising. In most regions of the Tel Aviv area there is not much variation in the slope, but the region is not completely flat. We have also tried variables related to bicycle speeds, but the estimation results did not improve, as with the slope.

3.5 Choice Set Sensitivity Tests

As indicated in the previous section, the choice set generation methods applied in this paper created up to 20 different routes for each observation. It is interesting to examine the impact of choice sets (in terms of size and composition) on the calibration results, following Prato and Bekhor (2007).

In order to examine the sensitivity of the parameters with respect to the choice set, we estimated Mixed Logit models according to a variable choice set size, using the PSL model structure for each test. The reference case is the "original" set composed of 20 routes (presented in Table 3). Two alternative choice sets were tested: the first was formed by drawing 10 routes from the 20 set, and the second by including 10 routes that were generated by link penalty and simulation methods only (that is, not including the routes generated by elimination). Table 4 summarizes the results for the two choice sets.

Table 4. Model estimation results for variable choice sets

Variable	Mixed PSL (set 1)			Mixed PSL (set 2)		
	Coeff.	t-stat	deviation	Coeff.	t-stat	deviation
Total route length (km)	-7.94	-12.1	-1%	-8.24	-8.8	+2%
Route length in category A (km)	2.05	4.8	-8%	2.13	6.8	-4%
Route length in category C (km)	-4.42	-5.9	-4%	-4.61	-5.1	-1%
Average street length (m)	0.06	4.7	+10%	0.058	3.9	7%
Dwelling units / m	2.1	3.3	-1%	2.31	8.1	9 %
Route length "near sea" (km)	1.3	6.4	+2%	1.23	7.25	-4%
Route length "near park" (km)	1.61	5.1	-6%	1.66	4.66	-4%
PS factor	1.68	9.1	-5%	1.7	8.79	-1%
Sigma (σ)	1.68	7.61	47%	1.4	6.71	23%
Hit-ratio	72%			76%		
Initial Log-likelihood	-1254.9					
Final Log-likelihood	-914			-903		
Likelihood Ratio	681			704		
Rho-bar squared	0.27			0.28		
Number of observations	545			545		

The results presented in Table 4 show that the coefficients do not significantly differ between the models, which indicate the robustness of the results with respect to different choice sets.

The evaluation of model robustness with respect to the choice set composition was further investigated for additional reductions of the choice set (35%, 50%, 65% and 80% of the original 20 routes). A total of 15 models are estimated and successively applied to the initial data sets to assess the impact over choice probabilities and likelihood values. Different levels of comparison

are available from this numerical experiment: across choice sets, across sample sizes, and across model specifications. Figure 7 and Figure 8 respectively present the computational results for Root Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE) of the coefficient estimates.

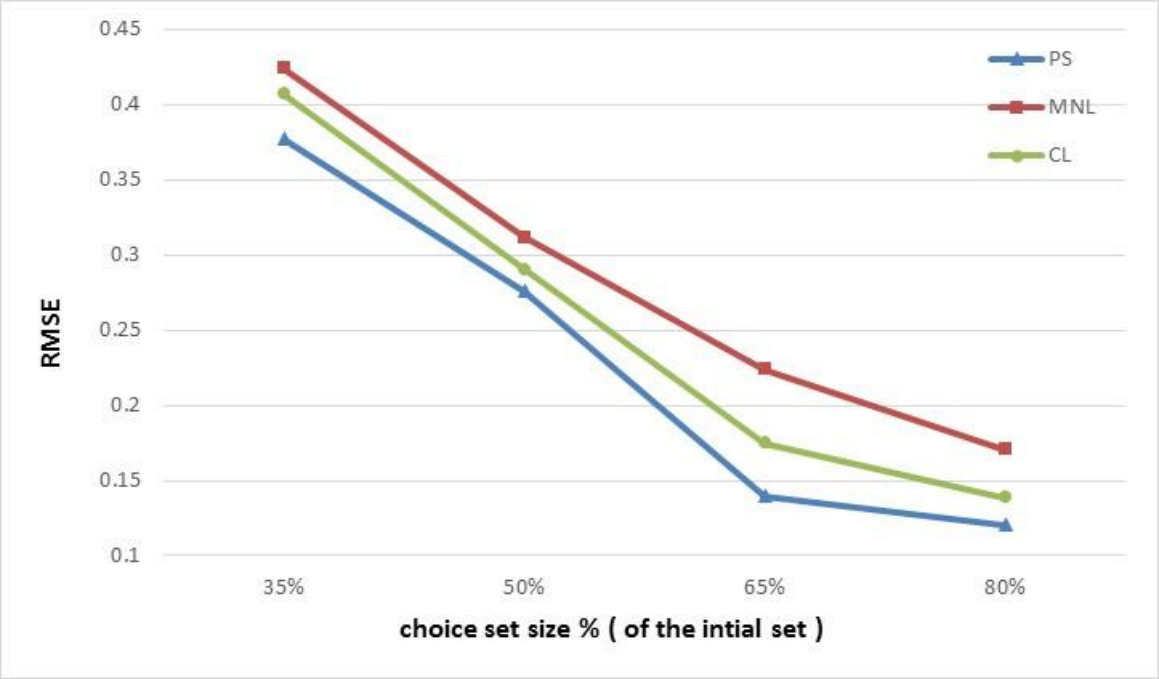

Figure 7. Ability to Replicate Parameter Estimates (RMSE)

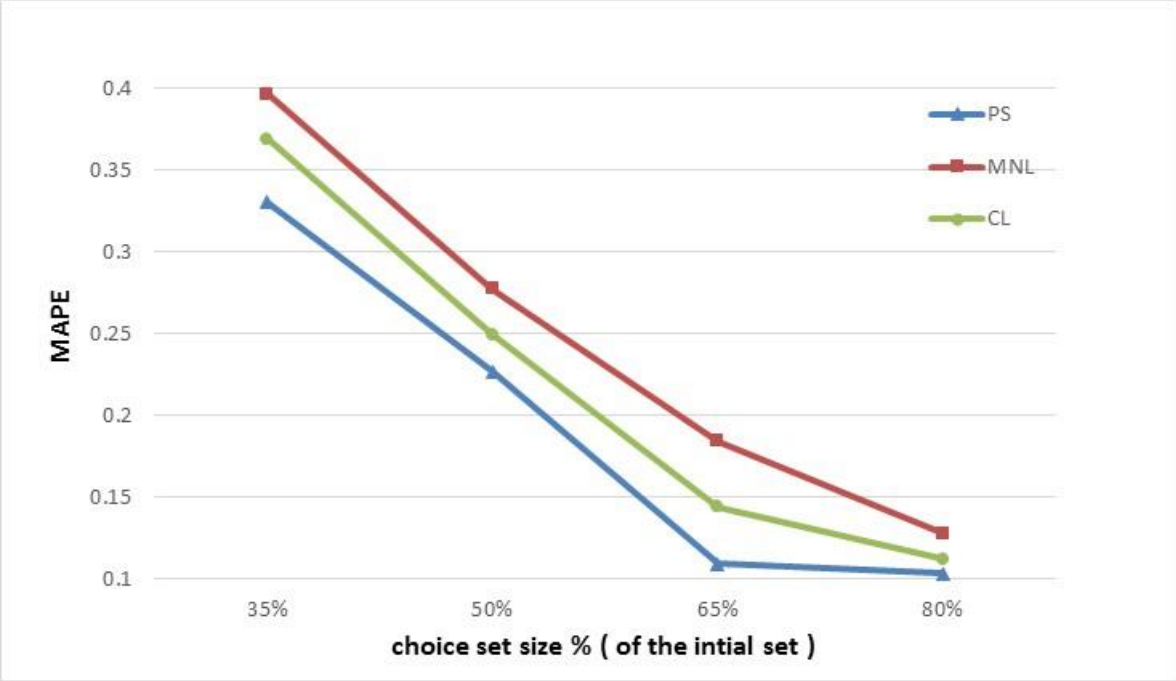

Figure 8. Ability to Replicate Parameter Estimates (MAPE)

The results presented in Figure 7 and Figure 8 indicate that if the choice set is reduced by half (that is, up to 10 routes per observation), the parameter estimates are still quite similar to the original choice set.

Following Nerella and Bhat (2004) and Faghih-Imani and Eluru (2017), we present the results of the standard deviation of the coefficients for the smaller samples. Figure 9 and Figure 10 respectively present the computational results for Root Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE) of the standard deviation of the coefficient estimates.

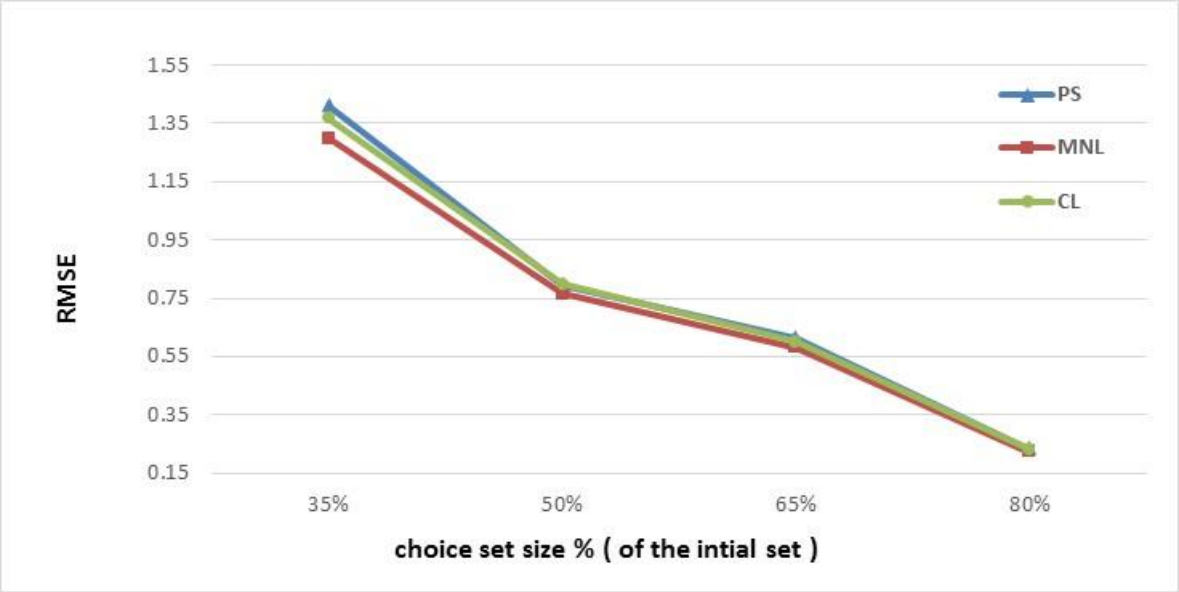

Figure 9. Ability to Replicate the Standard Deviation of the Parameter Estimates (RMSE)

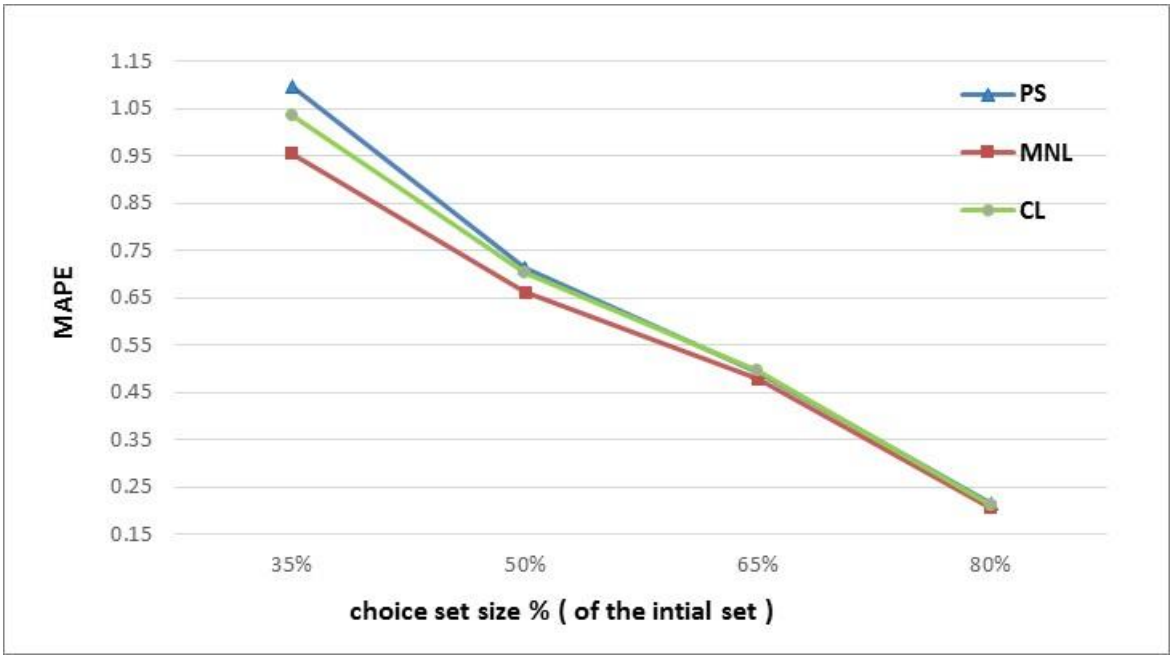

Figure 10. Ability to Replicate the Standard Deviation of the Parameter Estimates (MAPE)

As can be inferred from Figure 9 and Figure 10, the choice set size has a bigger impact in the standard deviation of the coefficient estimates, in comparison to the average coefficients. For example, if the choice set is composed of 10 routes, the standard deviation of the coefficients, the average percentage error is 0.65, more than double in the case of the average coefficient estimates.

3.6 Equal Utility Routes

The equivalent value of each parameter with respect to the total route length (calculated similarly to value of time) provides interesting insights. Table 5 presents the results for selected models.

Table 5. Equivalent values of selected variables with respect to route length

Variable	Mixed C-Logit	Mixed PSL	Mixed PSL (set 1)	Mixed PSL (set 2)
Route length in category A (km)	-0.278	-0.277	-0.26	-0.258
Route length in category C (km)	0.59	0.571	0.56	0.6
Average street length (m)	-0.008	-0.007	-0.00756	-0.00704
Dwelling units / m	-0.274	-0.264	-0.265	-0.28
Route length "near sea" (km)	-0.165	-0.159	-0.163	-0.15
Route length "near park" (km)	-0.224	-0.214	-0.2	-0.21

The model estimation results indicate that the utility of a bike route is equivalent to a route 38% longer than the shortest route. Similarly, the cyclist is willing to ride a route 57% longer than the shortest path to avoid riding routes with high level of stress.

Figure 11 below exemplifies the willingness to ride on longer routes. It shows 2 equal utility routes: the brown route (shortest path) with 1650 m, and the green route (with 60% length in a bikeway and the seashore) with 2235 m. In this specific example, the chosen route was the one close to the seashore.

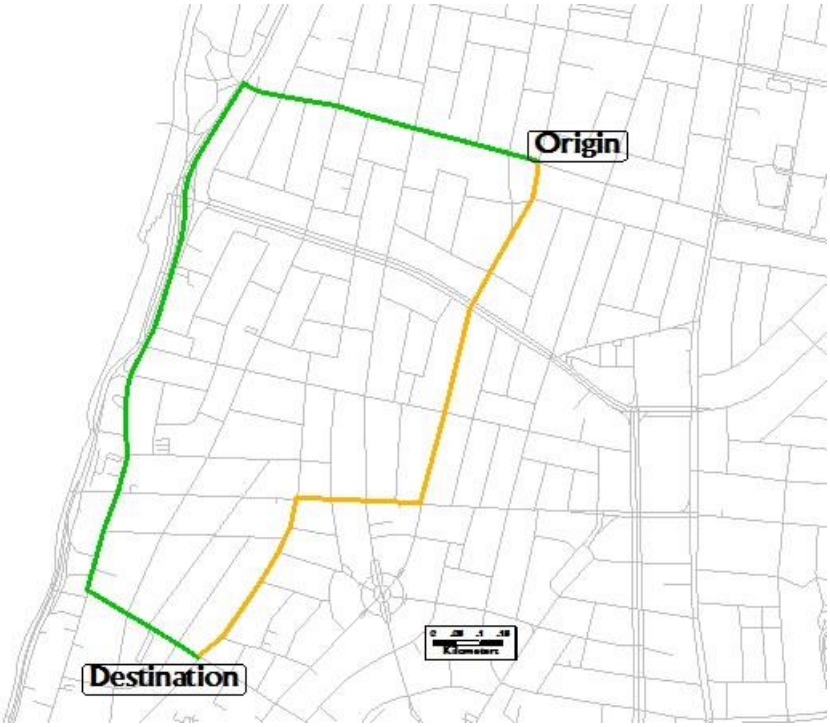

Figure 11. Two equal utility routes

4. Summary and Outlook

The choice set created with up to 20 routes per observation achieved high coverage, compared with reported results from the literature. The high resolution network provides rich information about street characteristics (including pedestrian and bike paths), which enables the formation of alternative routes that cover not only in terms of total length, but also in terms of additional variables, such as the length of bikeways.

Note that for a realistic network, the number of possible routes between each origin-destination pair may be very large. The results presented in this paper, estimated using data from a household survey and high resolution network, indicate that a relatively small number of generated routes can provide a robust choice set for model estimation.

As indicated in the introduction section, the two-stage approach was used in this paper, similar to most studies. The methods used to generate routes are variants of selective generation methods using shortest path algorithm, which may include bias in the estimation results (Frejinger et al., 2009). To overcome in part this problem, we have performed several sensitivity tests, in order to check the robustness of the parameter estimates. Further research will compare the estimation results presented in this paper with the recursive logit method proposed by Fosgerau et al. (2013), and recently applied in Zimmerman et al. (2017).

This paper purposely selected bicycle route observations from a general purpose survey. Survey results indicate a small share of bicycle trips, which are in part related to the lack of dedicated bicycle infrastructure. The model estimation results support this hypothesis, as exemplified in the equal utility route example.

It is possible to extend the model by including multi-modal trips (e.g. bike to train), but in this case there is a need to collect detailed data about the transit network and level of service variables. This is left for further research.

References

Azevedo, J.A., Santos Costa, M.E.O., Silvestre Madera, J.J.E.R., and Vieira Martins, E.Q. (1993). An Algorithm for the Ranking of Shortest Paths. *European Journal of Operational Research*, 69, 97-106.

Bekhor, S., and Prashker, J.N. (2001). Stochastic User Equilibrium Formulation for the Generalized Nested Logit Model. *Transportation Research Record: Journal of the Transportation Research Board*, 1752, 84-90.

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(1), 235-247.

Bekhor, S., Ben-Akiva, M. E., and Ramming, M. S. (2002). Adaptation of Logit Kernel to Route Choice Situation. *Transportation Research Record: Journal of the Transportation Research Board*, 1805, 78-85.

Ben-Akiva, M., and Bierlaire, M. (1999). Discrete Choice Methods and their Applications to Short Term Travel Decisions. In *Handbook of Transportation Science*, Kluwer, Dordrecht, The Netherlands, pp. 5-12.

Ben-Akiva, M.E., Bergman, M.J., Daly, A.J., and Ramaswamy, R. (1984). Modelling Inter-Urban Route Choice Behaviour. In *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*, VNU Science Press, Utrecht, The Netherlands, pp. 299–330.

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

Bovy, P., Bekhor, S., and Prato, C. (2008). The factor of revisited path size: Alternative derivation. *Transportation Research Record: Journal of the Transportation Research Board*, 2076, 132-140.

Broach, J., Dill, J., and Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46(10), 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 *Proceedings of the Thirteenth International Symposium on Transportation and Traffic Theory*, Pergamon, Lyon, France, pp. 697-711.

Chen, P., and Shen Q. (2016). A GPS-Based Analysis of Built Environment Influences on Bicyclist Route Preferences. In *95th Annual meeting of the Transportation Research Board*, paper No. 16-1948.

De La Barra, T., Perez, B. and Anez J. (1993). Multidimensional Path Search and Assignment. Presented at the 21st PTRC Summer Annual Meeting, Manchester, England.

Faghih-Imani, A., and Eluru, N. (2017). Examining the impact of sample size in the analysis of bicycle-sharing systems. *Transportmetrica A: transport science*, 13(2), 139-161.

Fosgerau, M., Frejinger, E., and Karlstrom, A. (2013). A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, 56, 70-80.

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

Halldórsdóttir, K., Rieser-Schussler, N., Axhausen, K. W., Nielsen, O. A., and Prato, C.G. (2014). Efficiency of choice set generation methods for bicycle routes. *European journal of transport and infrastructure research*, 14(4), 332-348.

Hood, J., Sall, E., and Charlton, B. (2010). A GPS-based bicycle route choice model for San Francisco, California. *Transportation letters*, 3(1), 63-75.

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

Los Angeles County Metropolitan Transportation Authority (2014). Bicycle Travel Demand Model: Phase II – Model Description. Staff working paper, 2014.

Mai, T., Fosgerau, M., and Frejinger, E. (2015). A nested recursive logit model for route choice analysis. *Transportation Research Part B: Methodological*, 75, 100-112.

Mekuria, M. C., Furth, P. G., and Nixon, H. (2012). Low-stress bicycling and network connectivity. Mineta Transportation Institute, San Jose State University, Report 11-19.

Menghini, G., Carrasco, N., Schussler, N., and K. Axhausen (2009). Route Choice of Cyclists in Zurich: GPS-based discrete choice models. Institute for Transport Planning and Systems, 2009.

Nerella, S., and Bhat, C. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models. *Transportation Research Record: Journal of the Transportation Research Board*, 1894, 11-19.

Oliveira, M., Vovsha, P., Wolf, J., Birotker, Y., Givon, D., and Paasche J. (2011). GPS-assisted prompted recall household travel survey to support development of advanced travel model in Jerusalem, Israel. *Transportation Research Record: Journal of the Transportation Research Board*, 2246, 16-23.

Plaut, P. O. (2005). Non-motorized commuting in the US. *Transportation Research Part D: Transport and Environment*, 10(5), 347-356.

Prashker, J. N., and Bekhor, S. (2004). Route choice models used in the stochastic user equilibrium problem: a review. *Transport Reviews*, 24(4), 437-463.

Prashker, J.N., and Bekhor, S. (1998). Investigation of Stochastic Network Loading Procedures. *Transportation Research Record: Journal of the Transportation Research Board*, 1645, 94-102.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of choice modelling*, 2(1), 65-100.

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

Ramming, S. (2001). Network Knowledge and Route Choice. Unpublished Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Winters, M., Davidson, G., Kao, D., and Teschke, K. (2011). Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation*, 38(1), 153-168.

Yai, T., Iwakura, S. and Morichi S. (1997). Multinomial Probit with Structured Covariance for Route Choice Behavior. *Transportation Research Part B: Methodological*, 31, 195-207.

Zimmermann, M., Mai, T., and Frejinger, E. (2017). Bike route choice modeling using GPS data without choice sets of paths. *Transportation research part C: emerging technologies*, 75, 183-196.