The Journal of Transport and Land Use http://jtlu.org
Vol. 11 No. 1 [2018] pp. 883–900

JTLU

Modelling route choice of Dutch cyclists using smartphone data

Silvia Bernardi
DICAM – University of Bologna
silvia.bernardi9@unibo.it

Lissy La Paix Puello
University of Twente
l.c.lapaixpuello@utwente.nl

Karst Geurs
University of Twente
k.t.geurs@utwente.nl

Abstract: This paper analyzes the GPS traces recorded by cyclists in the framework of the Mobile Mobility Panel throughout the Netherlands. The objective of this paper is to analyze bicycle route choice via network attributes and trip length over a sequence of trips by approximately 280 bicycle users, who were asked to register their trips by means of a specific smartphone application. Approximately 3,500 bike trips were recorded throughout the Netherlands over a four-week period in 2014. The bike trips have been matched to a specific bicycle network built and updated by a Dutch cyclists' union. Route choice models were estimated, using both the binomial logit model and the mixed multinomial logit model with Path-size logit model formulation. The chosen alternatives were part of the choice set for the mixed multinomial logit model. Also, the shortest route was generated for each origin-destination pair.

The results show that trip lengths and trip distribution over time reveal a population sample much used to cycling, frequently and over long distances. Furthermore, when considering the composition of chosen routes in terms of link type, the usage of cycleway links is frequent. For repeated trips, the shortest route option tends to be chosen more; frequent cyclists, on systematic trips, tend to optimize their trip and prefer the shortest routes. This is even truer for males and for non-leisure trips. The estimated probabilities for both multinomial and binomial models show that the binomial model tends to overestimate the probabilities of choosing the shortest route. This result is stronger in non-leisure trips, where people tend to choose a more personalized route, instead of the shortest. This research contributes to the generation of a more efficient distribution of bicycle trips over the network. Future research can more specifically address the intrapersonal variation in route — destination choice given the availability of longitudinal data.

Keywords: bike, cycling, GPS data, route choice, choice set

Article history:
Received: February 6, 2017
Received in revised form: November 1, 2017
Accepted: May 7, 2018
Available online: October 22, 2018

http://dx.doi.org/10.5198/jtlu.2018.1143
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

1 Introduction and literature framework

In the last decades, most cities, administrations and urban planners have given growing importance to sustainable transport modes, especially non-motorized modes like cycling. In the EU, many investments and funds have been dedicated to the development of more efficient bicycle infrastructures in medium-sized to large cities. As a result, research interest has been directed towards demand estimation and route choice models to the benefit of cycling as a modal option for urban trips. Findings from bicycle route choice models can point out cyclists' preferences, helping cities to better anticipate them and plan new bicycle infrastructures. In recent years, bicycle route choice modelling has seen a surge in availability of data, mostly coming from person-based GPS devices that provide researchers with travel diaries collected for ever bigger samples of users, and ever wider areas of study. A significant improvement in data accuracy, continuity and quality has been observed, due to the spread of smartphones and mobile applications for self-localization and navigation. Furthermore, highly detailed digitised networks have also become easier to obtain, thanks to the expansion of databases compiled by volunteers (volunteered geographic information). Inevitably, this has led to further complexity in how to treat this data and has raised the need for more efficient post-processing procedures and methodologies.

As for bicycle route choice modelling, most of the studies in the literature have been using stated preference (SP) data (Landis, Vattikuti, & Brannick, 1997; Axhausen & Smith, 1986; Hunt & Abraham, 2007; Krizek, 2006; Sener, Eluru, & Bhat, 2009; Stinson & Bhat, 2003; Tilahun, Levinson, & Krizek, 2007) , while only a few studies have used revealed preference (RP) data (Aultman-Hall, Hall, & Baetz, 1997; Howard & Burns, 2001). Stated preference surveys have been largely criticized, as they do not capture the actual choices of users, but only their intentions. A way to collect data that more reliably describe users' choices is through revealed preference data. Traditionally, two were the main limitations of RP data collection processes: the small size of samples, when obtained by surveys at a specific location, and the difficulty of observing the routes. In recent years, these issues have been partially overcome thanks to Global Positioning System. Since the spread of GPS dataset availability, the number of route choice models calibrated for bicycle use has increased (Menghini, Carrasco, Schüssler, & Axhausen, 2010; Hood, Sall, & Charlton, 2011; Zimmerman, Mai, & Frejinger, 2017; Broach, Dill, & Gliebe, 2012).

Particularly, Menghini et al. (2010) studied the route choices of cyclists in Zürich, Switzerland, concluding that the factor that mostly affected cyclists was the route length. Other factors — such as the presence of bicycle paths, the maximum gradient, and traffic lights — showed little effect. Their study confirmed the strong preference for direct routes and showed that faster cyclists prefer marked routes.

Hood et al. (2011) analyzed GPS traces from cyclists in San Francisco, USA, proving a preference for bicycle lanes to other bicycle facility types — e.g. paths and routes — especially for infrequent cyclists. Their study also confirmed the negative preference for length, as well as frequent turns, consistent with the findings by Zimmermann et al. (2017), who found that cyclists have a preference for simple route. Furthermore, slopes proved to discourage cyclists, especially women and commuters. No significant effects were found for traffic volume; neither did traffic speed, number of lanes, crime rates, rain, and nightfall show significant effects. However, the effect of network attributes on choosing specific (longer) routes still needs to be disentangled.

Another interesting application was carried out by Broach et al. (2012), who estimated the route choice of cyclists in the Portland, USA, metropolitan area. Their study confirmed the negative effect on cyclists' route choice for route length and slopes, and also found significant negative effects caused by traffic volume. A preference, instead, was proven for off-street bicycle paths, bicycle boulevards, and bridge facilities. Furthermore, the study found a significant — and interesting — effect for controlled intersections: the effect of traffic lights was found to be negative, meaning that the presence of a controlled intersections generally decreases a route's utility. Cyclists could feel slowered down due to waiting time. Nevertheless, the authors explained that, where conflicting traffic volumes were high, cyclists were discouraged by unsignalized intersections. This means cyclists tend to avoid controlled intersections, but in high traffic volume situations the presence of traffic lights could play an opposite role and attract cyclists, who perceive them as a safety-increasing feature.

The aim of the study is to explore the link between GPS-observed route choices made by cyclists and a considerable number of attributes of the transport network. In order to achieve this, the added value of the present study is threefold: Firstly, this research contributes to the investigation on cyclists' preferences by analyzing a large sample of GPS-recorded bicycle trips, a database including the repeated choices made by cyclists. Secondly, the transport network model available for the study provided unique details on the operational and qualitative features of the bicycle facilities in the Netherlands (such as quality, beauty of the context, brightness and illumination, etc.), constituting one of the most detailed bicycle network models in existence, built and updated by volunteer cyclists. Finally, the choice set built for the multinomial logit model increases the accuracy of choice sets, by only containing realistic options.

The paper is structured as follows. The next section describes the data available for this study: the data collection and all the relevant features of the dataset, the network model and details about the post-processing operation that transformed the recorded GPS points into actual revealed preferences.

2 Description of the database

2.1 The GPS trips database

The available database consisted of over 50,000 trips (28 million GPS points) recorded all over the Netherlands, for a period of four weeks between April and May 2014. About 600 respondents were recruited from an existing online panel, the Mobile Mobility Panel, providing a true representation of the Dutch population. The respondents were asked to register all their trips by means of a smartphone application called MoveSmarter (for iPhone and Android), specifically elaborated for this data collection. During the monitoring period, respondents also participated in a web-based prompted recall survey to check and revise trip characteristics, delete or add trips if necessary. For each registered trip, additional information such as origin, destination, mode and purpose were also checked by users and registered. This has been useful in "correcting" the automatic trip splitting and mode detection operations following the data collection stage. For more information about the data collection and the automated trip detection see Geurs, Thomas, Bijlsma, & Douhou (2015). Furthermore, for each participant, a rich set of socio-economic information was available: sex, age, family composition, home location, activity type and location, income etc.

For the application described in this paper, we selected only the trips made by bike, so the size of the database used for this study is approximately 4,000 trips, registered by 280 users. This limited number of bike trips registered — not limited per se but compared to the total trips in the original database — is due to data filtering operations, which are necessary to exclude incomplete and inconsistent records, based on the quality and completeness of the single-trip data.

Figure 1 shows the percentages of bike trips made by the participants in the monitoring period, grouped by purpose: the trips made for work and study-related reasons constitute a 12% of the total, while 25% of trips were made for leisure (including in the category physical exercise and enjoying the view, as these were other purposes the application proposed to participants). Shopping trips were the

11% of the total, while another small number of trips was labelled with other available purpose categories that can fall into the class of "errands" (visiting people, taking people somewhere etc.). A significant percentage of trips was categorized as "going home," and it is very likely that part of these include going home from work; unfortunately, it is not possible to determine where participants went home from for all these instances, as participants only for a small number of these trips provided a label for the origin location.

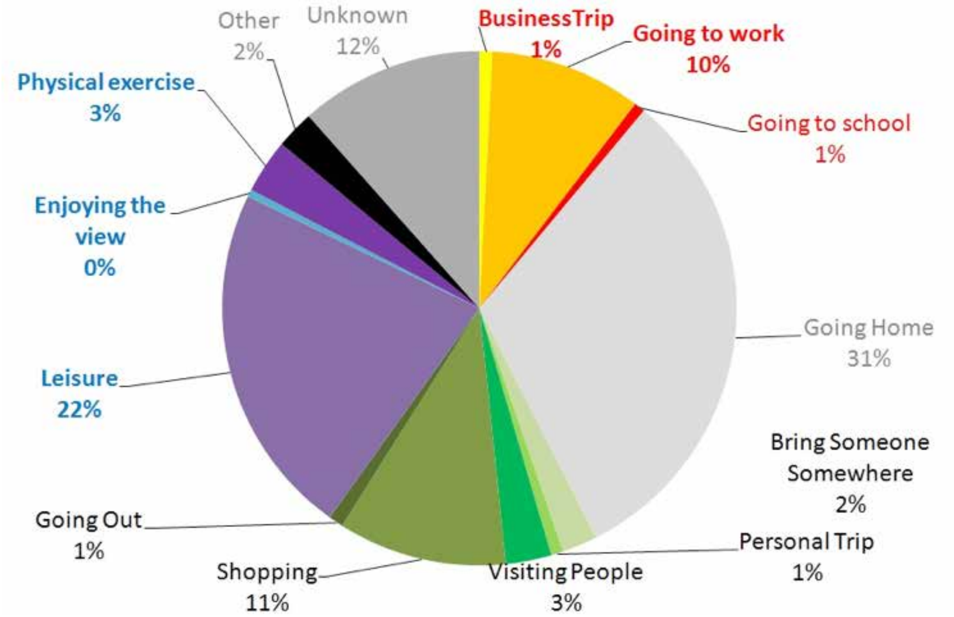

Figure 1: Percentages of bike trips by purpose

2.2 The network database

The network database available for this application is the bicycle network has been provided by the Fietsersbond (2015), a Dutch cyclist union, which realized different types of bicycle-related studies and initiatives within the Netherlands. With the help of volunteers, it annually updates a representation of the bicycle network throughout the whole country. This includes roadways (only those accessible to cyclists, so motorways etc. are not included) and all types of bicycle facilities available. Overall, the Fietsersbond network consists of more than 1.5 million links, representing a total network length of over 180,000 km. A wide range of attributes are associated with each link, including geometrical features (length, width), type of facility, type of pavement, and other features — described using qualitative scales — such as quality, beauty of the context, brightness and illumination, etc.

In addition, land-use data was available from Open Street Map (2015) and was joined to the network links using spatial analysis tools, thus determing which land-use typology every link intersected.

Table 1 shows the most frequent land-use typologies where the network links is included: a vast majority of the network links are crossing residential areas, and there is no strong variety in land-use types, considering the network.

Table 1: Land-use typologies crossed by Fietsersbond network links

Land-use typologies	**% of links crossing**
Residential	76.87%
Grass	11.65%
Industrial	7.22%
Farm	1.87%
Commercial	0.41%
Farmland	0.36%
Meadow	0.28%
Retail	0.25%
Orchard	0.22%
Greenhouse	0.20%
Construction	0.15%
Farmyard	0.10%

2.3 Post-processing

GPS data as recorded in smartphone applications requires a considerable amount of post-processing operations in order to become usable for further analysis of revealed preferences. In general, post-processing consists of individual stages accounting for data filtering, trip and activities detection, mode stage determination, mode identification, and map-matching (Schüssler & Axhausen, 2009).

As for data filtering, the trip and the mode detection stages, details on methodologies and results are available in Thomas, Geurs, Koolwaaij, & Bijlsma (2015). As already pointed out in the data collection methodology description, participants also filled out a web-based travel diary. Therefore, the automatically detected trip features (such as origin, destination, time, purpose and mode) could be verified using the information provided by participants themselves. Finally, the GPS points were matched to the network using the map matching algorithm described in Marchal, Hackney, & Axhausen (2005).

2.4 Descriptive analysis of the results

As we mentioned in the previous section, the network provided by the Fietsersbond (Dutch Cyclists' Union) is richer in those attributes describing the perceived quality of bicycle paths. Table 2 illustrates the kilometres cycled by respondents, grouped in terms of link type, link quality, link beauty and traffic nuisance perceived on links.

Table 2: Kilometers cycled by link type, link quality level, link beauty level and link traffic nuisance

Link feature	Km cycled per type of link	%	Km in the network database	%
Link type				
Roadway	6205.4	36.7%	89880.5	48.3%
Path along road	3360.6	20.6%	22827.2	12.3%
Bike lane	1890.0	11.2%	7143.5	3.8%
Exclusive bike path	1982.7	11.7%	14627.6	7.9%
Service road	478.1	2.8%	2963.3	2%
Bicycle boulevard	134.3	0.8%	172.6	0.1%
Pedestrian	119.5	0.7%	1836.9	1%
Unknown	802.7	4.7%	13337.2	7.2%
Blank	1948.7	11.5%	32633.7	17.5%
Link quality				
Good	9593.8	56.7%	87856.4	47.2%
Fair	4019.7	23.8%	41726.8	22.4%
Low	270.4	1.6%	5253.9	2.8%
Unknown	952.1	5.63%	15600.3	8.4%
Blank	2086.7	12.3%	35639.85	19.1%
Link beauty				
Neutral	8317.0	49.1%	77578.1	41.7%
Beautiful	4935.0	29.2%	51179.4	27.5%
Picturesque	330.1	1.95%	2707.9	1.5%
Ugly/boring	303.6	1.79%	3459.1	1.9%
Very ugly	6.9	0.1%	94.1	0.05%
Unknown	944.2	5.6%	15419.1	8.3%
Blank	2086.7	12.3%	35639.9	19.2%
Link traffic nuisance				
Little	6795.0	44.9%	77720.2	41.8%
Reasonable	3049.0	20.4%	23776.9	12.8%
Very little	1831.0	10.8%	30347.4	16.3%
Much	798.3	5.4%	2963.6	1.6%

[1] The Fietsersbond network database is specific for bicycle, thus it provides a description of all types of link where cyclists are allowed to ride. The term "roadway" mean the part of the road where motorized traffic is allowed. A "path along road" refers to a path adjacent to the roadway, but physically separated from it, where motorized traffic is not allowed (but moped are, according to Dutch traffic rules). A "bike lane" is a path obtained from the roadway, dedicated to cyclists, but not physically separated (only by painted marks on the pavement); according to Dutch traffic regulation, in case of high volumes of traffic, motorized vehicles can invade this space, so separation between them and bicycles is compromised. "Exclusive bike paths" are those paths reserved to only cyclists; frequently, these paths are not adjacent to roadways, but they can be found in parks or countryside environment. The "service road" label is used to indicate roads with low traffic volumes, sometimes adjacent to bigger roads, that have the function of access to private houses or activities. With the expression "bicycle boulevard" the author indicates a road that has been optimized for bicycle traffic. This means the road is open to motorized vehicles, but their posted speed is reduced (normally to 30 km/h), and signals indicate that the right of way belongs to cyclists. Finally, the label "pedestrian" is used for those paths reserved to pedestrian, like pavements, where cyclists sometimes ride even without allowance from the Dutch traffic regulation.

Given that also land-use information was available, we were able to explore the connection between perceived quality and beauty of links and the land-use typology those same links were crossing. Table 3 shows both the percentage of "good/fair/bad quality" rated links, and the percentage of "beautiful/neutral/ugly" rated links, considering the main land-use typologies. For some network links, such information was not available ("unknown" record).

Table 3: Percentage of "good/fair/bad quality" rated and unknown links, "beautiful/neutral/ugly/unknown" rated and unknown links, considering the main land-use typologies

Land-use typology	**Quality**				**Beauty**			
	Good quality links	**Fair quality links**	**Bad quality links**	**Unknown quality links**	**Beautiful links**	**Neutral links**	**Ugly links**	**Unknown beauty links**
Residential	48%	27%	1%	24%	14%	60%	2%	24%
Grass	54%	24%	5%	18%	54%	28%	1%	17%
Industrial	51%	17%	1%	31%	4%	46%	19%	31%
Farm	48%	25%	8%	19%	50%	30%	1%	19%
Farmland	28%	31%	30%	11%	67%	20%	1%	11%
Retail	46%	33%	2%	20%	12%	61%	8%	20%
Orchard	48%	28%	12%	12%	58%	29%	1%	12%

3 Travelled distance versus shortest distance

Another evaluation of cyclists' behaviour is provided by the comparison between the chosen paths, and the minimum cost path in terms of length. This gives a measure of how cyclists' trips were "optimized". Shortest paths have been calculated using the network model and ArcGIS, taking into account length, and excluding highways and motorways, as these are unavailable for cycling.

Comparing the chosen routes with the shortest paths, on average per trip, cyclists cycled 1.37 km longer than the minimum cost path, 15% in percentage of trip length. Figure 2 shows the distributions of the difference between chosen routes and shortest path lengths, as a percentage of trip length (or "detour"). Figure 3 shows that, as the trip length increases, the amount of detour increases: the calculated correlation was 0.68, meaning a positive and strong correlation for the two variables, trip length and detour. For longer trips, cyclists tend to detour more from the most direct path.

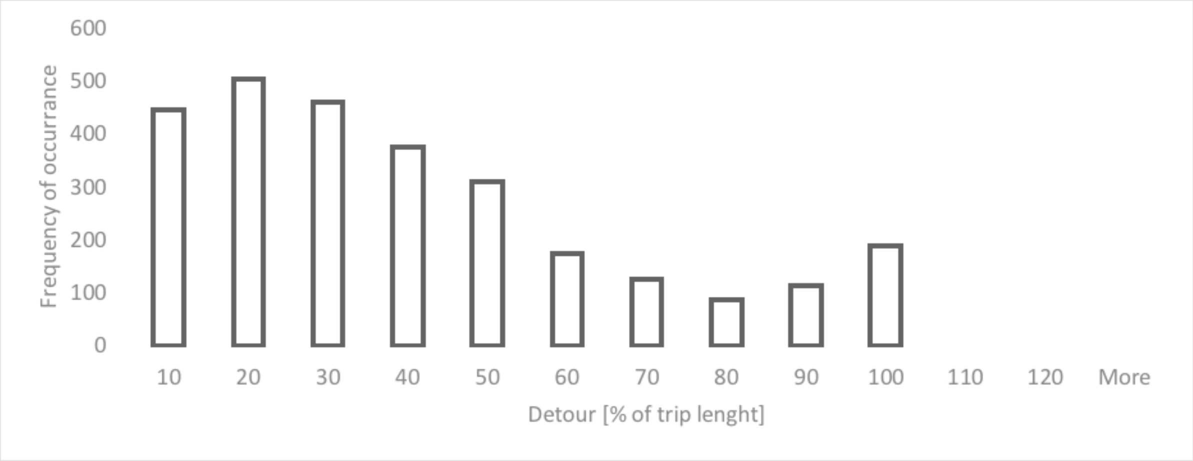

Figure 2: Frequency distribution of the detour, i.e., the difference between chosen routes and calculated shortest routes, expressed in percentage of length

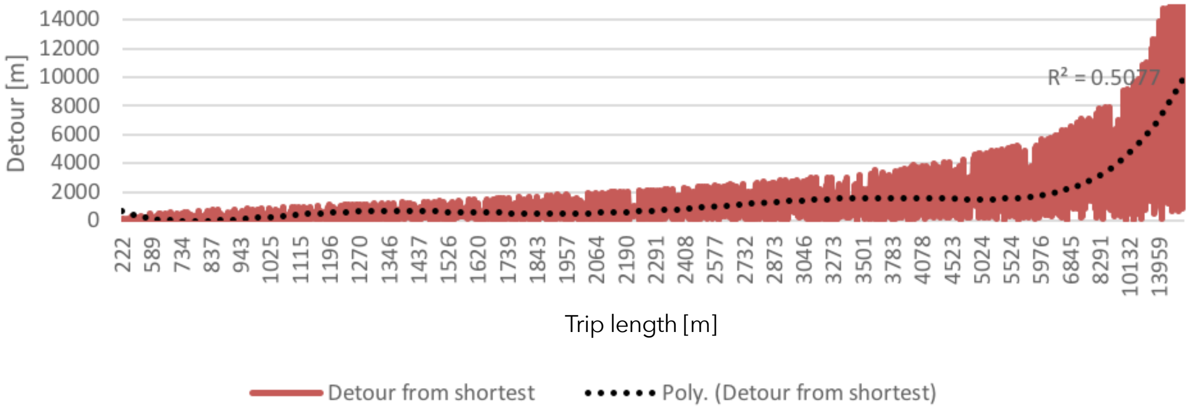

Figure 3: Distribution of the difference between chosen routes and calculated shortest routes, shown against increasing trip length

The "optimization" in terms of length of cyclists' trips has also been categorized by purpose: as expected, trips made for leisure show the highest percentage of detour – calculated as average over all leisure trips (Table 4).

Table 4: Average detour (measured in percentage over trip length) by trip purpose

Trip purpose	**Average detour [% over trip length]**
Leisure	42,03%
Bringing Someone Somewhere	38,87%
Other	37,44%
Shopping	37,14%
Going to work	35,22%
Physical exercise	34,51%
BusinessTrip	33,95%
Going Out	33,65%
Going Home	33,32%
Enjoying the view	32,96%
Visiting People	32,13%
Unknown	31,82%
Personal Trip	29,69%
Going to school	29,30%

The choice of a specific path, though, is not only influenced by its length, or purpose, but also by other factors such as safety of the path, presence of facilities, amenities, quality level etc. For this reason, as will be shown in the following paragraphs, other relevant factors — such as type of road, presence of bicycle facilities, or personal features — have been included in the analysis.

4 Choice set generation

The quality of route choice models strongly depends on the choice set considered; especially on its size, its composition, and on how plausible the generated alternatives are. Many studies in recent years have focused on the generation of choice sets (Bekhor, Ben-Akiva, & Ramming, 2006; Prato & Bekhor, 2006; Bliemer & Bovy, 2008; Bovy & Fiorenzo-Catalano, 2007; Broach, Gliebe, & Dill, 2010; Rieser-Schüssler, Balmer, & Axhausen, 2012; Halldórsdóttir, Riesler-Schüssler, Axhausen, Nielsen, & Prato, 2014), where the main issue is to generate alternatives that are plausible and relevant, i.e., that most likely represent the actual routes that users take into consideration. Methods proposed in literature, such as the breadth-first search on link elimination (BSF-LE) (Rieser-Schüssler et al., 2012) or the branch and bound (B&B) (Bovy, 2007), propose quite straightforward algorithms based on the search for shortest paths, but can generate alternatives that do not consider cyclists' preferences for other factors like safety and quality of paths. Most advanced methodologies try to take into account — in addition to the geometry and topology of the network — other attributes that have proven to be relevant for cyclists: doubly stochastic generation function (DSGF) methods, as seen in Nielsen (2000), Bovy and Fiorenzo-Catalano (2007), Halldórsdóttir et al. (2014), allow to consider multi-attribute cost function. For example, Halldórsdóttir et al. (2014) not only included length in the cost function, but also road type, presence of dedicated bicycle paths, and land-use attributes.

Even so, the risk of considering as part of the choice set routes that are not actually considered by cyclists as relevant options, is considerable. When a substantial number of repeated observations are available, over a significant period of time, the alternative routes available to each participant of the sample population can be directly deducted from the observed routes. In other words, if a certain number of repeated trips made by the same user between the same origin and destination pair has been registered, and a variety of different chosen routes can be observed, it is reasonable to generate the choice set of the user, for that specific origin and destination pair, as a selection of those routes. To avoid too high a correlation between the different alternatives in the choice set, observed routes can be grouped following different criteria of similarity, such as length, percentage of dedicated bicycle facilities, number of intersections, quality or land-use attributes etc. The choice sets generated are likely to be smaller than those created by means of a choice set generation algorithm. However, all the alternatives in the set are realistic and have been considered by cyclists as available options.

For the present study, in order to individuate the trips made by the same user between the same OD pair, origins and destination locations were grouped using their 5-digit postcodes. Over the 3502 bicycle trips available, recorded by 262 different participants, 1442 (41%) were repeated more than once. These trips were then grouped by length, in order to define up to 4 groups of registered trips — for each user and each OD pair — which represented the four alternatives available to cyclists. The attributes of each alternative have been obtained in the shape of an average for all the recorded trips belonging to the group. For each user and OD pair, one additional alternative has been introduced, represented by the shortest path between the OD pair. Thus, each choice set has a minimum size of 1 (i.e., the shortest path, calculated for all OD pairs) and a maximum size of 5 alternatives (4 generated from the observed routes plus the shortest path) (Table 5).

Table 5: Alternatives and their choice rates

Alternative	Description	Choice rate
Shortest	Up to 15% longer than the shortest	26.61%
Alt 1 (chosen 1)	16% to 30% longer than the shortest	23.87%
Alt 2 (chosen 2)	31% to 60% longer than the shortest	17.65%
Alt 3 (chosen 3)	61% to 75% longer than the shortest	17.05%
Alt 4 (chosen 4)	More than 76% longer than the shortest	14.82%

5 Model estimation

5.1 Discrete choice modelling framework

When dealing with route choice, the independence of irrelevant alternatives (IIA) property of the Multinomial Logit (MNL) model makes it inappropriate for estimating discrete choices among similar alternatives, such as different paths available to a cyclist between an OD pair in an urban context. In such cases, an overlapping path may not be perceived as a distinct alternative. For bicycle route choice modelling this problem has been overcome by introducing a similarity measure in the utility function, as done by Ben-Akiva and Bierlaire (1999) for the so-called Path Size Logit (PSL) model. This model measures the similarity using a Path-Size term in the deterministic component, which indicates the fraction of the path that constitutes a "full" alternative. For the PSL model, the expression of the probability of choosing route k within the alternative paths reflects the simple Logit structure:

$$P(i/C_n) = \frac{e^{V_{in}+lnPS_{in}}}{\sum_{j\in C_n} e^{V_{jn}+lnPS_{jn}}} \qquad (1)$$

where the Path-Size factor is defined as:

$$PS_{in} = \sum_{a\in\Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j\epsilon C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}} \qquad (2)$$

where Γ_i is the set of all links in path i, l_a is the length of link i, L_i is the total length of path i, C_n is the set of all the alternatives for user n, $L^*_{C_n}$ is the length of the shortest path in C_n, L_j is the total length of path j; δ_{aj} is the link-path incidence variable, and equals 1 if link a is part of path i and 0 otherwise.

Another adaptation of the Logit model that accounts for similarities in the stochastic part of the utility (error correlations) while maintaining a closed-form formula for probabilities is called Mixed Logit (Nielsen, 2000). The defining characteristic of the Mixed Logit model (also called "Logit Kernel") is that the unobserved factors can be decomposed into a part that contains correlation and heteroscedasticity, and another part that is assumed to have an extreme value distribution (Ben-Akiva & Bierlaire, 1999). Mixed Logit is a highly flexible model that can approximate any random utility model (Train, 1986). It resolves the three limitations of standard Logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time.

A mixed logit model can be used to represent error components that create correlations among the utilities for different alternatives (Bierlaire, 2003). Utility is specified as:

$$U_{ni} = \alpha' x_{ni} + \mu' z_{ni} + \varepsilon_{ni} \qquad (3)$$

where x_{ni} and z_{ni} are vectors of observed variables for alternative i, α is a vector of fixed coefficients, μ is a vector with random terms and zero mean, and ε_{ni} is distributed i.i.d. extreme value. $(\mu' z_{ni} + \varepsilon_{ni})$ is the unobserved portion of utility, which can be correlated over alternatives depending on the specification of z_{ni}.

For the estimation of route choice models, we referred to the formulation of the Mixed Logit model and estimated the distribution of the error component. This allowed us to account for the different participants to the travel survey, who repeated a number of choices (panel data).

In addition, for the Multinomial Logit model, where various alternatives are considered, we introduced the Path-Size term as defined by Ben-Akiva and Bierlaire (1999), thus the probability of choosing path *i* is:

$$P(i|C_n) = \frac{e^{\beta(x_{in}+lnPS_{in})}}{\sum_{j\epsilon C_n} e^{\beta(x_{jn}+lnPS_{jn})}} \quad (4)$$

6 Binomial logit model

The first model we estimated was a binomial mixed logit model, in which the two alternatives are represented by the shortest route (Alt0) and the chosen route, in case it differs from the shortest.

We made the assumption that if the chosen trip length corresponded to the shortest length between the same OD pair, or it was longer by no more than 15%, then it could be said the cyclist chose the shortest path for travelling between that specific OD pair. This assumption included, the percentage choosing the shortest path amounted to 27% of the trips, while longer options were chosen for 73% of instances.

The attributes considered were the percentage of link types in the trip (keeping roadway as a reference, and considering the different types of bicycle links), the number of traffic signals, percentage of good quality links, beautiful links, and little traffic nuisance links in the trip. As personal attributes age and sex were considered, and finally trip purpose. For all the parameters, the shortest path alternative was considered as reference. The parameters were estimated in BIOGEME (Bierlaire, 2003). Not all estimated parameters appeared significant. For the significant parameters, results of the estimation are shown in Table 6.

For what concerns the different link types, only the presence of bike lanes, bicycle boulevards and service roads turned out to be significant. Given the negative signs within their parameters, all three of these link type proved to be avoided by cyclists (in relation to roadway links). In addition, the presence of traffic signals turned out to be positively affecting the choice for longer paths, compared to the shortest alternative.

Regarding the personal characteristics of cyclists, age did not prove to be significant, while gender did. Gender was introduced as a dummy variable, keeping "female" as a reference. The results indicate that male cyclists perceive longer alternatives as less attractive, in comparison to the shortest available option, and when compared to their female counterparts.

Table 6: Results from the binomial mixed logit model estimate

Utility parameter	**Value**	**T-test**
ASC_shortest	1	--(fixed)
ASC_chosen	-0.79	-4.82
β_male (dummy)	-0.68	-3.24
β _signals	0.07	3.98
β _(bike_lane)	-1.13	-2.04
β _(bike_boulevard)	-4.87	-2.40
β _service	-1.88	-2.09
β _(good_quality)	6.01	1.52
β _(low_quality)	7.81	1.9
σ_(panel_shortest)	0.79	3.09

Table 6: Results from the binomial mixed logit model estimate (cont.)

Utility parameter	**Value**	**T-test**
σ_(panel_other)	-0.77	-2.96
Number of observations	2798	
Number of draws	200	
Initial log-likelihood:	-1939.00	
Final log-likelihood:	-929.40	
Rho-square:	0.520	

As for the quality levels of links, considering fair-quality links as the reference case, both the presence of good-quality links and low-quality links resulted significant and positive. A positive sign was to be expected for the proportion of good-quality links in the longer paths while contrarily a negative sign could be expected for the proportion of low-quality links. This can be explained if we refer to Table 2: in the descriptive analysis, we already highlighted that low-quality links represented only the 1.6% of the chosen links, so the choice-sample for low-quality links is too small to result in representative findings. For this context, where the vast majority of links in the network database are of good quality, quality itself is not actually determining cyclists' preferences.

7 Mixed multinomial Path-Size logit model

For the multinomial Path-Size model, the previously described choice sets were used. The Path Size formulation expressed in Eq.1 was considered. The attributes considered were the percentage of link types in the trip, the number of traffic signals, percentage of good quality links, beautiful links, and little traffic nuisance links within the trip. As personal attributes age and sex were considered, and finally trip purpose. Specific parameters for the 5 alternatives available were estimated, again using BIOGEME (Bierlaire, 2003). Not all estimated parameters resulted in significant findings; for the significant parameters, results of the estimation are shown in Table 7. For this model, none of the parameters describing the type of link along the route reached the significance level.

For the shortest alternative, the visual attractiveness (beauty) of the path, calculated as the percentage of beautiful links for the route, resulted significant and negative, considering a fair level of beauty as reference. This could be interpreted as a negative preference for beauty along the route for those cyclists that choose the shortest available path, meaning they choose speed over pleasantness. Another unexpected coefficient that reached the level of significance is the one for the attribute "percentage of links with little traffic nuisance along the route", which resulted negative for alternative 4. It means cyclists using the longest alternatives tend to avoid links with low traffic nuisance, with respect to links with fair levels of traffic nuisance. It could be related to leisure trips, without time constraints, where cyclists would not mind to pass by traffic-congested areas.

Table 7: Results from the mixed multinomial Path-Size logit model estimate

Utility parameter	**Value**	**T-test**
ASC_shortest	1	--(fixed)
ASC _alt1	2.44	11.69
ASC _alt2	2.85	13.38
ASC _alt3	2.61	11.39
ASC _alt4	2.32	7.58
β_(signals_Alt1)	0.34	2.25
β _(signals_Alt3)	0.30	1.78
β _(beautiful_shortest)	-0.86	-1.63
β _(little_nuisance_Alt4)	-0.70	-1.61
β _(Leisure_Alt1)	0.67	2.33
β _(Leisure_Alt3)	1.02	2.66
β _(Leisure_Alt4)	1.67	4.36
β _(PathSize_Alt1)	0.45	2.86
β _(PathSize_Alt2)	0.55	3.59
β _(PathSize_Alt3)	0.44	3.05
β _(PathSize_Alt4)	0.29	2.24
σ_(panel_shortest)	1.01	6.46
Number of observations	2798	
Number of draws	200	
Initial log-likelihood:	-1797.25	
Final log-likelihood:	-883.50	
Rho-square:	0.508	

As seen for the binomial logit model, the presence of traffic signals turned out to be positively affecting the choice for longer paths, with respect to the shortest alternative. Referring to the personal characteristics of cyclists, age did not prove to be significant, while gender did: using gender as a dummy variable, and keeping "female" as a reference, male cyclists result to perceive longer path alternatives as less attractive. Nevertheless, the corresponding parameter failed to reach significance.

The positive sign of the Path-Size factor, as calculated for all of the alternatives (except for the shortest path that is the reference) indicates that if a route is less similar to the alternatives, its chances of being chosen will be high. This positive effect of the Path-Size factor was also reported in other past studies on route choice models (Prato & Bekhor, 2006, 2003).

Among all possible trip purposes tested, only "leisure" resulted significantly influencing cyclists' route choices, with a significant and positive parameter: this means that for leisure trips cyclists choose paths longer than the shortest available option, which represents a consistent and expected result.
Furthermore, the error component expressing the correlation between choices made by the same user (σ_panel) only resulted significant for the shortest-path alternative, and positive: this means that users that repeated the same choice multiple times (same trip, i.e., same origin-destination pair) showed a tendency to choose the shortest path available. Such a result is somewhat expected and intuitive, as frequent cyclists are more likely to optimize their trips, especially when these trips are repeated, as with commuting trips, for instance.

Comparing the two models, binomial and multinomial, the goodness of fit in the binomial model is slightly superior (rho squared equal to 0.52) compared to the multinomial model (rho squared 0.506). However, the information obtained from the multinomial model is more valuable in terms of definition of alternatives.

Figure 4 shows the estimated probabilities for both binomial and multinomial models. The same models were estimated after excluding leisure trips. The probabilities for those (binomial and multinomial) models are also included in Figure 4. As can be seen, the share for the shortest route in the binomial model is equal to 25%, whereas the average probability of choosing the shortest route in the multinomial model is only 5%.

Finally, the estimated probabilities for non-leisure trips (including work and education) indicate that for these trip purposes, the probabilities of choosing the shortest route increases when the binomial model is used. However, the probability of choosing the shortest route remains almost equal when multiple alternatives are considered. Also, the share for alternative 'chosen 1' increases, indicating a most preferred and personalized route, which is not the shortest, for non-leisure trips. This highlights the importance of considering multiple alternatives in route choice modes, instead of the simple incarnation of binary choices.

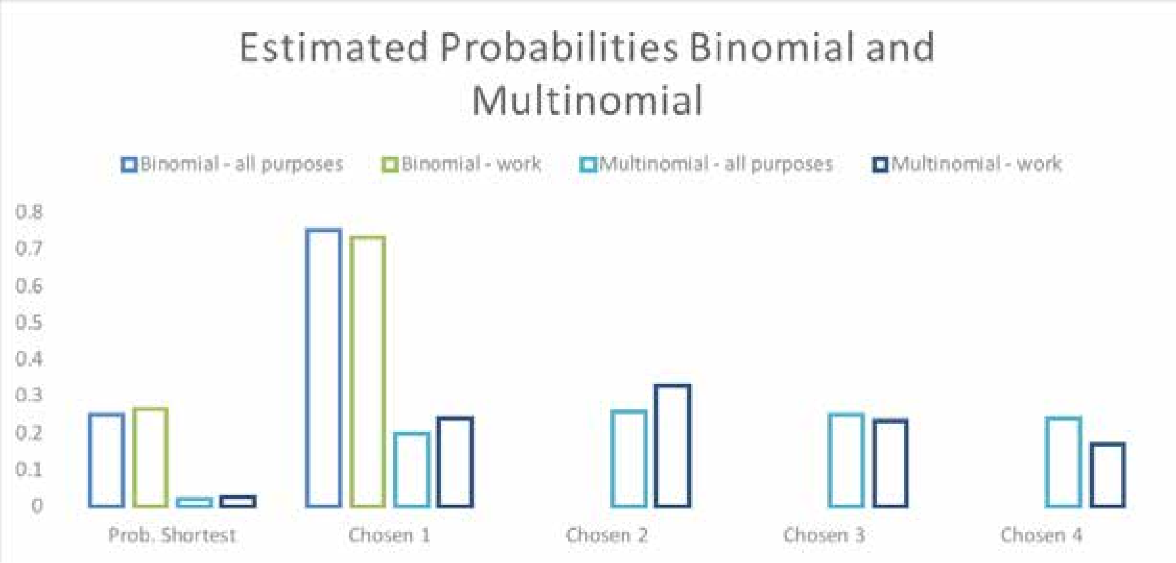

Figure 4: Estimated probabilities for binomial and multinomial models

8 Discussion and conclusions

This paper analyses the GPS traces recorded in the framework of the Mobile Mobility Panel, throughout the whole of the Netherlands. Approximately 280 bicycle users were asked to register all their trips by means of a smartphone application called MoveSmarter that was specifically elaborated for this data collection effort. Furthermore, the recorded trips have been corrected by the users themselves, who checked and corrected their travel information daily on a web-based prompted recall survey. Finally, the bike trips had already been matched to the Open Street Map network, but a second network database was available as well, in the shape of the Fietsersbond (Dutch Cyclists' Union) network database — a specific bicycle network built and updated by this union – which achieves a surprising level of detail, both in terms of link number and in terms of information attached to its links.

From a first descriptive analysis of the Dutch GPS traces, one can clearly spot some main pattern: trip lengths and trip distribution over time reveal a population sample that is much used to cycling, frequently and over long distances. Furthermore, when considering the composition of chosen routes in terms of link type, the usage of cycleway links is much frequent. This can be explained considering that the percentage of dedicated facilities in the Dutch network is much higher than within other European contexts.

The level of detail of the datasets involved allowed us to estimate bicycle route choice models: a binomial logit model and a multinomial logit model. Path-size logit model formulation proved to be the best model formulation for bicycle route choice, as it allows to consider the overlapping of alternatives that characterizes bike trips. Furthermore, we referred to the mixed logit model formulation to introduce an error component that was capable of capturing the influence of trip repetitions. Indeed, many GPS traces from the database consisted of the same trip (origin-destination pair) repeated by the same user throughout the monitoring period. The estimation results have shown that for repeated trips, the shortest route option tends to be chosen more. This finding suggests that frequent cyclists, on systematic trips, tend to optimize their trip and to prefer shortest routes.

Unexpectedly, the available information regarding quality, beauty and traffic nuisance for the network links did not result in significant parameters for modelling. Nevertheless, from the descriptive analysis of the GPS trips, we saw that most trips are made on links that present a good-to-very-good level of quality and beauty, and a low-to-very-low level of traffic nuisance. This is probably explained keeping into consideration that the general quality of bicycle links in the Dutch context, especially in urban environments, is rather high. The RP data usage for modelling discrete choice relies on variation in the observed environment so that the statistical model can discern how the various parameters influence the choice. If little variation is found in the data, then the model will fail to quantify the effect of a specific parameter. Consequently, RP data-based models, being based upon observable data in a certain geographical context, could not be suitable for analyzing the effect of the same factors in a different context. For example, for urban areas where the quality — or beauty, or traffic nuisance — of bicycle paths can greatly vary from one alternative to another, such parameters could result to be more significant for cyclists' route choice.

Another important aspect of this research is the choice set generation methodology. Given the high number of trips in the database, and their repetitions, we tried to identify the alternatives that compose the choice set based on the observed routes. Choice sets obtained in this way are certainly smaller than those obtained by topological or probabilistic algorithms we described in the literature review section, but they do represent a more realistic set. In order to build the alternatives, these trips were then grouped by length, and 4 groups of registered trips — for each user and each OD pair — were created, which represented the alternatives available to cyclists. The attributes of each alternative have been obtained as an average of all the recorded trips belonging to the group. For each user and OD pair, one additional alternative has been introduced, represented by the shortest path between the OD pair. Thus, each choice set has a minimum size of 1 (i.e., the shortest path, calculated for all OD pairs) and a maximum size of 5 alternatives (4 generated from the observed routes plus the shortest path).

It should be highlighted that, in order to take advantage of large recorded choice sets like those provided by smartphone-based GPS datasets, the number of repeated trips made by the same user between the same origin and destination pair should be consistent, and a variety of different chosen routes should be observed. Given such premises, the growing availability of smartphone-based GPS data can help overcome the main limitations of RP data collection processes related to small sample sizes and the difficulty of observing different chosen routes.

For future development, it would be interesting to consider other criteria, other than the sole length of trips, for grouping the observed routes and creating alternative sets; for example, groups could be made based on the proportion of good-quality links along the path, or based on the number of traffic signals. A cluster analysis could be performed on multiple parameters in order to define the alternatives based on actual differences between observed routes. Alternatively, the choice set could be enriched by adding to the shortest route, the safest route, or the one entirely made by separated cycleway links, or the most scenic, etc., as calculated by any GIS routing tool.

References

Aultman-Hall, L., Hall, F., & Baetz, B. (1997). Analysis of bicycle commuter routes using geographic information systems: Implications for bicycle planning. *Transportation Research Record: Journal of the Transportation Research Board, 1578* (1), 102–110.

Axhausen, K. W., & Smith, R. L. (1986). Bicyclist link evaluation: A stated preference approach. *Transportation Research Record: Journal of the Transportation Research Board, 1085,* 7–25.

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

Ben-Akiva, M. E., & Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions. *Handbook of Transportation Science,* 5–34.

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

Bliemer, M. C. J., & Bovy, P. H. L. (2008). Impact of route choice set on route choice probabilities. *Transportation Research Record: Journal of the Transportation Research Board, 2076,* 10–19.

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

Broach, J., Gliebe, J. G., & Dill, J. L. (2010). Calibrated labeling method for generating bicyclist route choice sets incorporating unbiased attribute variation. *Transportation Research Record: Journal of the Transportation Research Board, 2197*, 89–97.

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

Fietsersbond. (2015). Fietsersbond routeplanner. Retrieved from http://www.fietsersbond.nl/fietsrouteplanner/

Geurs, K. T., Thomas, T., Bijlsma, M., & Douhou, S. (2015). Automatic trip and mode detection with MoveSmarter: First results from the Dutch Mobile Mobility Panel. *Transportation Research Procedia, 11,* 247–262.

Halldórsdóttir, K., Riesler- Schüssler, 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.

Hood, J., Sall, E., & 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.

Howard, C., & Burns, E. K. (2001). Cycling to work in Phoenix: Route choice, travel behavior, and commuter characteristics. *Transportation Research Record: Journal of the Transportation Research Board, 1773,* 39–46.

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

Krizek, K. J. (2006). Two approaches to valuing some of bicycle facilities' presumed benefits. *Journal of the American Planning Association, 72*(3), 309–320.

Landis, B. W., Vattikuti, V. R, & Brannick, M. T. (1997). Real-time human perceptions: Toward a bicycle level of service. *Transportation Research Record: Journal of the Transportation Research Board, 1578,* 119–126.

Marchal, F, Hackney, J. K., & Axhausen, K. W. (2005). Efficient map matching of large Global Positioning System data sets: Test on speed-monitoring experiment in Zurich. *Transportation Research Record: Journal of the Transportation Research Board, 1935,* 93–100.

Menghini, G., Carrasco, N., Schüssler, N., & 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.

Open Street Map (2015). Retrieved from https://www.openstreetmap.org/

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

Prato, C. G., & 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.

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

Schüssler, N., & Axhausen, K. W. (2009). Processing GPS raw data without additional information. *Transportation Research Record: Journal of the Transportation Research Board, 2105,* 28–36.

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

Stinson, M. A., & Bhat, C. R. (2003). Commuter bicyclist route choice: Analysis using a stated preference survey. *Transportation Research Record: Journal of the Transportation Research Board, 1828,* 107–115.

Thomas, T., Geurs, K., Koolwaaij, J. & Bijlsma, M. (2015). Automatic trip and mode detection with Move Smarter: Results from the Dutch Mobile Mobility Panel. *Transportation Research Procedia, 11,* 247–262.

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

Train, K. E. (1986) *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand.* Cambridge, MA: MIT Press.

Train, K. E. (2003). *Discrete choice methods with simulation.* Cambridge, UK: Cambridge University Press.

Zimmermann, M., Mai, T., & Frejinger, E. Bike route choice modeling using GPS data without choice sets of paths. *Transportation Research Part C, 75,* 83–196.