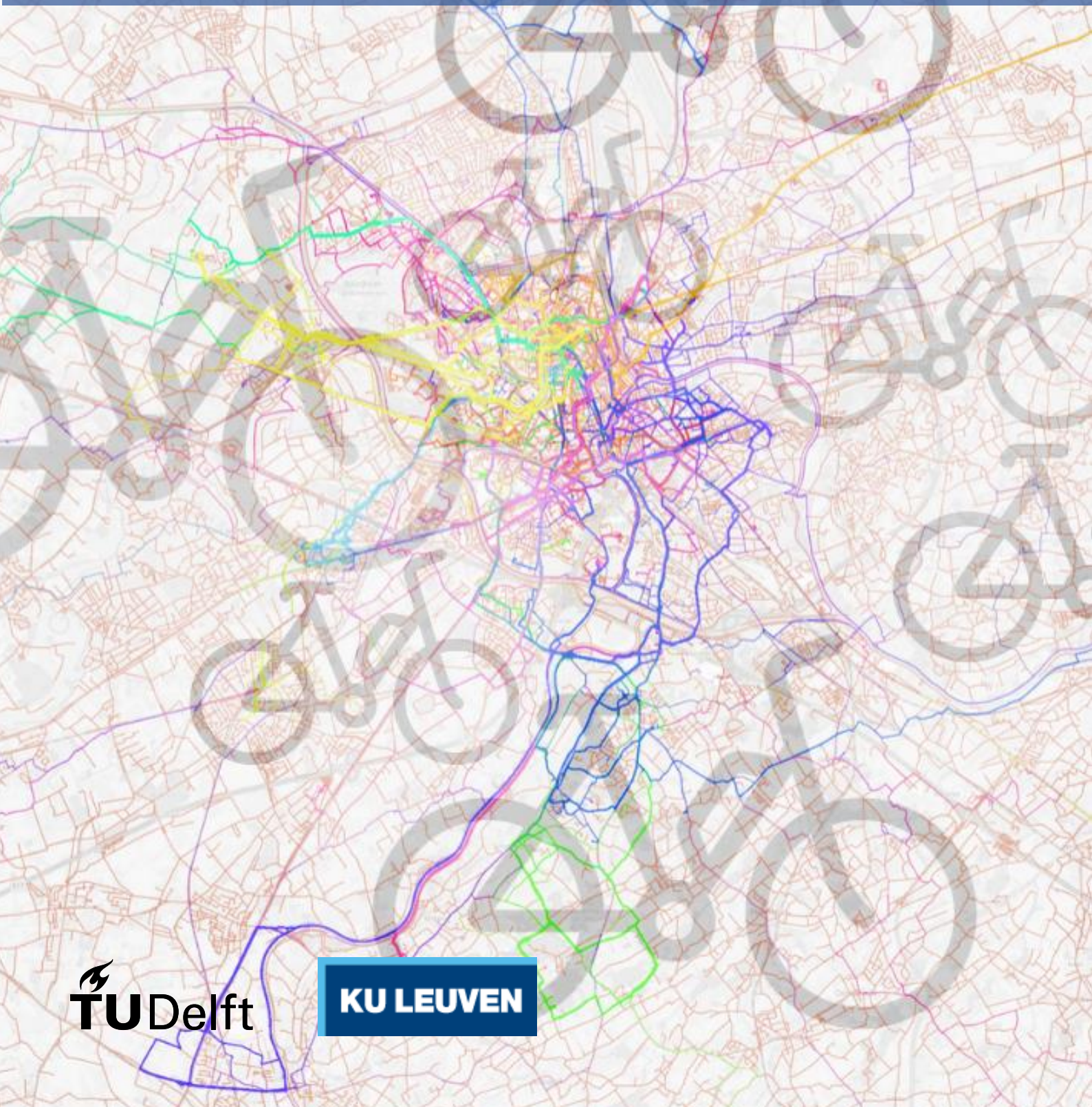
S. Kishoen Misier

Determinants of E-biker's route choice behavior

The case of E-bikers in Ghent.



TU Delft

KU LEUVEN

Determinants of E-biker's route choice behaviour

The case of E-bikers in Ghent.

By

Suraj Kishoen Misier
4625811

in partial fulfilment of the requirements for the degree of

Master of Science

in Transport and Planning

at the Delft University of Technology,

to be defended publicly on Tuesday January 8, 2020 at 02:00 PM.

Supervisor:	Prof.dr.ir. Serge Hoogendoorn	
Thesis committee:	Dr. Dorine Duives,	TU Delft
	Prof. Chris Tampère,	KU Leuven
	Dr. Danique Ton,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

PREFACE

This thesis report is the final step of the master course in Transport and Planning at TU Delft and concludes my study journey in the Netherlands. This project was done in collaboration with KU Leuven. I hope that this report will be of much value for future research, as well as for practice. This final project certainly was of much value for my development in knowledge and skill in research in the field of cycling route choice behavior analysis in particular.

I am grateful for all the people that guided, inspired and helped me along this journey with all its ups and downs. I am grateful for Dr. Danique Ton and Dr. Dorine Duives for the pleasant guidance throughout the whole project, for checking up on me and giving me that boost when my motivation was at a low and inspiring me with their work in the field of active modes. I will certainly miss our regular meetings starting and ending with the funny conversations on various topics. I am grateful for the supervision from a distance by Prof. Chris Tampère for providing me with the SPRINT dataset, in depth guidance through the whole project and very valuable inputs through the many discussions via Skype. A special thanks goes to Dr. Willem Himpe at KU Leuven, who guided me in depth through the data processing procedure, provided me with example codes and for his valuable input through the many discussions via Skype. I am grateful for the guidance of Prof.dr.ir. Serge Hoogendoorn as supervisor and the chair from the university.

I am grateful for the friends I have made at the university along the way who inspired me and made this journey a pleasant one. I am grateful for family and friends outside of the university for providing me with the warmth that I am used to back home in Suriname. My biggest gratitude goes towards my parents, who sacrificed so much over the years just to provide me with this opportunity of having this journey abroad. I hope to make them proud with my ongoing development.

Suraj Kishoen Misier
Paramaribo, January 2020

SUMMARY

One of the trends in the world is that more people and cities see the bicycle as viable alternative to the car, as this mode has many benefits over the car. The increasing popularity of the electric bikes helps this trend, especially in Belgium where almost one of the two bikes sold is an e-bike.

Research in cyclists' route choice behavior should help in modeling travelers' behavior that is needed to make the right decisions for future investments regarding cyclists. Research on cycling route choice behavior is increasingly done in the recent years using revealed preference data in terms of observed GPS tracks. All the route choice behavior studies are about traditional cyclists. To the best of the author's knowledge, there is no published study on route choice behavior of E-bikes yet and thus no statement can be made whether E-bikers consider the same factors as traditional cyclist in their route choice and how the relative importance of the factors differs between the E-bikers and traditional cyclists.

This research aims at analyzing route choice behavior of E-bikers, in terms of which factors play a role and to what extent. The research will be performed based on E-bikers route data in Ghent. Two interesting developments are aimed to include as well, in terms of multi-attribute cost functions as input for route generation and including land use attributes to cyclist's route choice models. The following main question is answered with this research:

"How do E-bikers choose their route?"

A short summary including conclusions and discussions of the different aspects in this research is given in the following sub-sections.

Literature study

A literature study is performed on which *route choice model* and *route generation* method to apply, as well as which *attributes* to consider in the model.

Based on the literature study, the following three main types of *route choice models* exist based on their model structure: Logit-, Generalized Extreme Value- (GEV) and Non-GEV structures. Logit models are models that keep the simple MNL structure and can be extended to cope with correlation of the route alternatives. The GEV model structures deal with correlations in the stochastic part of the utility function, have a tree structure and relate the network topology to the specific coefficients, but they do not consider taste variation or correlation over time of unobserved factors. Non-GEV model structures also allow random taste variation and correlation in unobserved factors over time. These models do not present closed-form expression for the choice probabilities. Their estimation is based on simulation and thus more complex than the other models.
For this research, heterogeneity of preferences and correlation of unobserved factors are neglected. For estimating the general route preferences of E-bikers, the logit structured models should suffice. The simple Logit structured and well-established Path Sized Logit (PSL) model is used in this research. An attempt is made to estimate the relatively new and promising logit structured Recursive Logit (RL) model as well. Several attempts were made to get the RL model to converge consistently for different initial parameter values without success. Because of the complexity and lack of experience with this model, the exact source of the problem could not be identified within this project. However, a qualitative comparison is still made. The PSL model is thus the route choice model chosen for this study.

An overview is made on the different types of *route generation methods*. Four types of route generation methods are distinguished: Deterministic shortest path-based, Stochastic Shortest path-based, Constrained Enumeration methods and Probabilistic methods. Most of the route generation methods are of the Deterministic shortest path-based type. The cyclist's route choice behavior studies apply a wide variety of route generation methods with the majority being the Labeling -, Breadth First Search on Link Elimination (BFS-LE) - and Doubly stochastic generation function (DSGF) methods. The first two methods are of the Deterministic shortest path-based type while the DSGF is of the Stochastic shortest path-based type. Halldórsdóttir et al., 2014 compared, from what they found in literature, the three most promising methods for car route generation in the cycling context. The three methods they compared are: Breadth First Search on Link Elimination (BFS-LE), Doubly stochastic generation function (DSGF) and Branch & Bound (B&B). They applied multi- attribute cost functions by adding different weights to the distance of links with different road types, cycle lanes and land use attributes. They state that multi-attribute cost functions increase the performance of the route generation methods in terms of consistency with observed behavior and heterogeneity of the choice set. The BFS-LE and DSGF method had similar performance in terms of consistency. The DSGF had a slightly better heterogeneity in terms of route choice set than the BFS-LE method, but the BFS-LE was by far the most computationally efficient of the three. The BFS-LE method is not the most applied method in the cyclist's route choice studies, which arguably could be related to the fact that none of the studies considered multi-attribute cost functions except for Prato et al. (2018). Due to the great performance and efficiency of the BFS-LE method when using multi-attribute cost functions, it is the route generation method of choice for the PSL route choice model in this research.

The performed literature study, on studies about cyclist's route choice analysis using GPS data, lead to a list of *attributes* that were found to be of significance on cyclist's route choice behavior. None of the studies included E-bikers route choice behavior, but the same attributes are expected to be of relevance for E-bikers. Based on these studies, a list of attributes with an indication of their relative impacts on the route choice behavior is made. The list of attributes considered for this study is made based on the most included and significant attributes in the related studies and the availability of the data. These are: Distance, Cycleways, Large roads, Small roads, Roundabouts, Traffic lights, Left turns, Right turns, Wrong way, Green area, Water area, Building area, Up-slopes 2-4%, Up-slopes 4-6% and Up-slopes >6%.

Data

GPS tracks of E-bikers from the SPRINT project in Ghent (Belgium) was found suitable to estimate the route choice behavior of E- bikers. Data concerning the cyclist GPS route choice behavior, network and other relevant route choice attributes of the area are gathered and processed. The location data of the E-bikes in the SPRINT research were automatically logged by a GPS tracking device, which was directly connected to the E-bike's battery. The GPS tracks were made by a group of 21 participants that made trips in the Ghent region that were recorded over a year. The network data obtained from OpenStreetMap is simplified and enriched by adding additional attributes from the Grootschalig Referentie Bestand of Basiskaart Vlaanderen' (GRB) in terms of land-use data on network-link level.
The GPS tracks are processed as well. Trips needed to be identified and are filtered such that only trips that are of interest are included. The filtered GPS data is then map-matched to the processed network. This processed data is then used further for route generation and the choice model.

Route generation

Different *multi-attribute cost functions* are considered in the route generation using the BFS-LE method. Halldórsdóttir et al. (2014) included attributes related to road types, cycle lanes and land use

attributes. They were able to generate routes that were highly consistent with the observed routes. In line with them, the attributes considered are 'Road category' and 'Land use'. Links are put into mutual exclusive categories for each attribute. Links are put into mutual exclusive categories for each attribute. Calibration of the weights for each category within an attribute is done on a limited set of attributes, a small part of the data that can represent the whole data set and with limited predefined discrete weight values. Different cost functions, with each one attribute as well as a combination of attributes, are defined and their performances compared. None of the multi-attribute cost functions are found to perform substantially better than the conventional distance-only cost function in terms of consistency with observed routes and heterogeneity. A choice for the best function between the tested multi-attribute cost functions and different weight sets is difficult to make, since they are similar in terms of performance. Given this finding that the tested multi-attribute cost functions do not improve the route sets, the distance-only cost function is used to generate the routes for the big dataset. The conventional distance-only cost function does not perform much different but is often used and accepted in previous research. The resulting routes are described in the table below. The differences in mean and standard deviation between the generated routes and the observed routes are given in the last columns as well.

Table 1. Descriptive statistics per considered attribute of the generated routes compared to the observed routes.

		OBSERVATIONS			***GENERATED ROUTES***			***Diff. Mean***	***Diff. St. dev.***
	Unit	***Mean***	***Median***	***St. dev.***	***Mean***	***Median***	***St. dev.***		
Length	*km*	6.97	5.01	6.31	7.55	5.70	6.78	8%	8%
PS	-	0.52	0.54	0.25	0.16	0.12	0.12	-69%	-50%
Wrong way	*km*	0.03	0.00	0.06	0.03	0.00	0.06	-7%	-2%
Cycleway	*km*	0.91	0.31	1.47	1.14	0.43	1.60	25%	9%
Large road	*km*	2.03	1.30	2.22	1.48	1.21	1.34	-27%	-40%
Small road	*km*	2.62	2.08	2.11	2.14	1.64	1.91	-18%	-9%
Other road	*km*	1.40	0.22	2.95	2.79	0.63	4.99	99%	69%
Green	*%. km*	2.92	1.36	3.39	3.67	1.56	4.31	26%	27%
Water	*%. km*	0.59	0.20	0.96	0.89	0.38	1.40	52%	46%
Building density	*%. km*	1.10	0.99	0.72	0.81	0.67	0.53	-26%	-26%
Scenic (G+W based)	*km*	5.52	2.98	5.79	6.49	3.55	6.84	18%	18%
Roundabouts	*#*	0.56	0.00	0.96	0.43	0.00	0.69	-23%	-28%
Traffic Lights	*#*	1.25	1.00	1.65	0.74	0.00	0.97	-41%	-41%
Left turns	*#*	4.77	4.00	3.82	4.60	4.00	3.10	-3%	-19%
Right turns	*#*	4.93	4.00	3.71	4.63	4.00	2.91	-6%	-22%
Intersection cross (no turn)	*#*	46.53	41.00	32.38	27.04	26.00	15.82	-42%	-51%
Up-slope 2-4%	*km*	0.18	0.11	0.20	0.13	0.08	0.15	-26%	-23%
Up-slope 4-6%	*km*	0.02	0.00	0.04	0.02	0.00	0.05	0%	24%
Up-slope >6%	*km*	0.02	0.00	0.04	0.01	0.00	0.04	-15%	20%

The generated routes are similar in terms of length, wrong way distance, and number of turns. However, they differ quite significantly with respect to the other variables. All the differences in these exploratory variables between the routes should give some indication on why not all the observed routes are reproduced.

The generated routes have a much lower PS factor on average compared to the observed routes, indicating that the generated routes are far less unique in their choice sets and differ significantly from the observed routes. The generated routes seem to make use of other roads much more and less of large and small roads. The other roads include unpaved paths through woods as well, which are not used by the E-bikers included in this study. The generated routes make use of links with more greenery, more water and less buildings in their direct environment. There are also less traffic lights and roundabouts included on the generated routes. A big difference in the number of intersection crossings (no turns) is also noticed. Perhaps, the observed cyclists cross many intersections where they have priority over crossing traffic and prefer to cross those intersections instead of making turns on other intersections where they must stop. Information about right of way (or stop signs) was unfortunately not available and hence excluded from this research. In terms of the proportion of up-slopes, the generated routes are somewhat similar to the observed routes.

Choice model

Different permutations of the considered attributes are tested in the Path Size Logit model. These attributes include: the route length, path size factor, one-way restrictions, different road categories (cycleways, large-, small-, other roads), land use (green-, building- and water area nearby the link as well as a scenic attribute combining the previous attributes into dummy values), intersection control (roundabout and traffic lights), turns (left, right, none) and up-slopes (2-4%, 4-6% and >6%).

The final model is given in the table below.

Table 2. Parameter estimates with their respective standard deviations and t-values for the chosen PSL model.

Attribute	***Unit***	***Parameter Value***	***St. dev.***	***t-value***
Length	km	-0.65	0.03	-20.20
Ln (PS)		2.42	0.03	69.50
Cycleway	*	3.92	0.31	12.50
Large road	*	3.24	0.28	11.50
Small road	*	4.53	0.27	16.90
Left turns	per km	-0.23	0.04	-5.32
Right turns	per km	-0.20	0.04	-4.80
Roundabouts	per km	0.56	0.15	3.82
Traffic Lights	per km	0.54	0.10	5.26
Wrong way	*	-5.47	1.42	-3.85
Green area	**	-4.70	0.40	-11.90
Up-slope 2-4%	*	10.40	0.86	12.00
Up-slope 4-6%	*	16.40	2.13	7.71
0 LL			-12772	
Final LL			-5863	
LRS (init. model)			13817	
$\bar{\rho}^2$			0.54	

** proportion of route length*

*** weighted proportion of route length*

The estimated parameters were used to predict the choices of the validation data set. The obtained loglikelihood was -2531 and the hit-ratio was 63%.

The attributes other than the distance that are found to be of importance (in terms of their parameter value) for E-bikers in Ghent ranging from most important to least important are: the proportion of up-slopes within 4-6%, up-slopes 2-4%, one-way restricted roads on the route, roads with greenery in the direct environment, small roads, dedicated cycleways and large roads. The number of roundabouts, traffic lights, left turns and right turns are also of importance.

However, the validity of the resulting model might be questioned, since many of the generated routes are not consistent with the observed routes. The route generation method not being able to reproduce the observed behavior most of the time resulted in these big differences between the generated routes and the observed routes and thus impacting the outcome of the route choice model.

Discussion

The most important discussion points are:

- The validity of the resulting model might be questioned, since it is not able to predict choices much better than a random model given the hit-ratio of 63%. The route generation method not being able to reproduce the observed behavior most of the time resulted in these big differences between the generated routes and the observed routes and thus impacting the outcome of the route choice model. Calibrating the weights for the multi-attribute cost functions for route generation and afterwards estimating the preferences seems like a 'Chicken-Egg' dilemma. Nonetheless, there is some iteration needed between the two, in order to capture the choice behavior in the route generation already. This dilemma is less present in the link-based RL model, since no route generation is needed. This iteration has not been applied in this study but is highly recommended.
- While the BFS-LE method does produce many different routes, it does not always yield realistic routes that people would consider. Generated routes often make a detour around the removed link and follow the original shortest path. This has to do with its principle of randomizing the network by removing random links from the network. An interesting approach would be to generate different sub-route sets using different cost functions and combining them into one route set with more behavioral consistency and more heterogeneity. The different cost functions would represent the different preferences of different people.
- The performance of the generated route sets in this project are lower compared to what Halldórsdóttir et al. (2014) have reached in terms of consistency index and reproduction rate. Most likely this is related to the difference in data. Several reasons might include:
    - This dataset of E-bike trips includes longer trips which are proven to be more difficult to reproduce than shorter ones. It is noteworthy that Halldórsdóttir et al. (2014) showed that reproducing the observed trip becomes more difficult the longer the trip is.
    - It is noteworthy that Halldórsdóttir et al. (2014) don't compare the performance achieved by using the multi-attribute cost functions with that of the basic distance-only function for the same dataset. The dataset might be such that the performance of the later may already be much higher than other studies, because the data is inherently already easy to reproduce by the method.
    - E- bikers are probably less sensitive to distance and therefore might accept detours in exchange for other attributes not captured in cost function. This makes it much harder to reproduce these routes. For short routes this is less of a problem

Recommendations for further research

Some interesting attributes yet to be included in future work include: Light poles, Mini roundabouts, Trip purpose, Bike type (in terms of speed limit of E-bike), Cycle lane, Panel effects and Car Traffic volume. Additional information on traffic priority in terms of stop signs could be also an interesting addition. Crossing an intersection without having priority is hypothesized to result in a lower utility than crossing while giving priority to motorized traffic. Interaction effects are also not investigated but are recommended to investigate.

The RL model requires the data to be correctly processed beforehand. The model is complex and does not provide clear feedback about errors. It is advised to validate the data first with a simple conventional model like the PSL model and to be extra aware of units, since the model cannot cope with link utilities close to zero.

A route set performance indicator that indicates not only the consistency with observed routes in terms of distance of overlapping links, but also in terms of overlapping characteristics like cycleways or land use would be interesting to look at as well. Ghanayim & Bekhor (2018) have proposed such a generalized overlap indicator.

Recommendations for practice

The parameter estimates can be used cautiously as input for the cost functions for E-bikers for route generation in traffic models or navigation software. But it is strongly advised to do so after a feedback loop between the route generation and choice model is applied.

In terms of data collection, it is advised to let participants of route choice data collection study state their trip purpose in future in terms of utilitarian and recreational trips. Or at least let them mention the locations of their home, work and one or two regular activity locations such that the trip purpose can be inferred for the majority of the trips of that person. Also, open source network data of OSM is proven to be very useful for this type of research. Enriching this open source data set with even more attributes is recommended such that future research and projects can easily make use of high-quality data.

In terms of infrastructure investments and policy measures for the city of Ghent, E-bikers in Ghent have a big aversion for cycling in the wrong direction. Adequate route alternatives should be provided for these cyclists when measures concerning one-way restrictions are to be implemented. With respect to intersection control, E-bikers in Ghent have a slight preference for roundabouts over traffic lights. Investments in controlled intersections could be justified using these preferences in terms of perceived distances.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The urge to shift towards more sustainable and healthy cities results in a search for more sustainable forms of transportation systems. One of the trends in the world is that more people and cities see the bicycle as viable alternative to the car, as this mode has many benefits over the car including: cost-, health- and environmental related benefits (Pan-European Programme, 2014). The increasing popularity of the electric bikes helps this trend, especially in Belgium where almost one of the two bikes sold is an e-bike (Oortwijn, J., 2019). In order to stimulate the usage of bicycles, cities must make decisions on where and how to efficiently invest in cycling infrastructure, policy- and financial incentives (van Overdijk et al., 2017; Skov-Petersen et al., 2018; Rupi & Schweizer, 2017). Research in cyclists' behavior should help in modeling travelers' behavior that is needed to make the right decisions for future investments regarding cyclists. Cyclist preferences in planning tools still seem to be lacking (van Overdijk et al., 2017). One of the choices cyclists make is that of the route choice. By observing route choice behavior of cyclists, the infrastructure and environmental factors which they find relevant can be inferred and quantified. This is being done by comparing the characteristics of a chosen route to those of the possible alternative routes that the cyclist could have chosen. These factors can be included via route choice models as input for the transport planning tools. In general, route choice models depend on the chosen routes and the generation method of the route alternatives.

Research on cycling route choice behavior is increasingly done in the recent years using revealed preference data in terms of observed GPS tracks (Dill, 2008; Menghini, 2009; Hood, 2011; Ehrgott et al., 2012; Broach 2012; Casello, 2014; Halldórsdóttir et al., 2014; Hintaran, 2016; Khatri et al., 2016; Ton et al., 2017; Ghanayim & Bekhor, 2018; Ton et al., 2018; Prato et al., 2018; Oehrlein et al., 2018; Rupi and Schweizer, 2018; Skov-Peterson et al., 2018). Advancements in GPS equipment make it easier to capture revealed preference route data. Alongside this advancement, also route generation methods and route choice models have been developed. Another development is that the list of attributes considered in the route choice models for cyclists are extended.

Cyclists seem to be sensitive towards their environment in their route choice as well. Halldórsdóttir et al. (2014), Ghanayim & Bekhor (2018) and Prato et al. (2018) have shown that land use attributes are found significantly relevant for cyclist's route choice.

One interesting development in route generation methods for cyclists is that of including multi-attribute cost functions for generating route alternatives, such that the route alternatives are more consistent with observed behavior. Halldórsdóttir et al. (2014) made a comparison of what they defined as the three most effective route generation methods for car routes by applying them to cyclist route choice data. They found significantly better route sets using multi-attribute cost functions in their study compared to other cyclist related studies which use distance-only cost functions.

All the route choice behavior studies are about traditional cyclists. To the best of the author's knowledge, there is no published study on route choice behavior of E-bikes yet and thus no statement can be made whether E-bikers consider the same factors as traditional cyclist in their route choice and how the relative importance of the factors differs between the E-bikers and traditional cyclists.

This research aims at analyzing route choice behavior of E-bikers, in terms of which factors play a role and to what extent. The research will be performed based on E-bikers route data in Ghent. This research aims to include the two mentioned interesting developments as well, in terms of multi-attribute cost functions as input for route generation and including land use attributes to cyclist's route choice models.

The following main question and sub questions shall be answered with this research:

“How do E-bikers choose their route?”

- Which route choice model should be used based on literature?
- Which route generation method should be used based on literature?
- Which relevant attributes should be included in the route choice model based on literature?
- To what extent does a multi-attribute cost function in the route generation improve the choice set compared to the distance-only method?
- What attributes do E-bikers find important in their route choice and to what extent?

The report is structured as follows. The results of a literature study presented in chapter 2 aims at answering the questions related to attributes used for cyclist’s route choice behavior, types of route choice models and route generation methods. The research methodology that follows from the literature study is presented in chapter 3, in which the entire process from getting from raw data to the route choice behavior of the E-bikers is discussed. Next, the data is described including how it is collected and processed in chapter 4. This data includes GPS traces from E-bikers in Ghent and network data. The data processing includes network simplification, trip identification & -filtering and map-matching. The most promising route choice model found in the literature was the Recursive Logit (RL) model, however the model could not be successfully estimated. A comparison of the RL model and the Path Size Logit (PSL) model is presented in chapter 5. The results of the route generation method are displayed and discussed in chapter 6. The resulting route set is used to estimate the PSL route choice model. The results of the PSL model are displayed and discussed in chapter 7. Conclusions following the findings of the research are finally discussed in chapter 8. Recommendations for future research and practice are discussed as well.

2 THEORETICAL BACKGROUND

The results of a literature study aimed at answering the following questions are presented in this chapter:

Which route choice model should be used based on literature?

Which route generation method should be used based on literature?

Which relevant attributes should be included in the route choice model based on literature?

A classification of the different route choice models based on their structure is discussed (section 2.1). Route generation methods are crucial for estimating route choice models. An overview of the route generation methods is also given, along with commonly used performance indicators of their resulting routeset (section 2.2). A list of the considered attributes in previous cyclist's route choice behavior studies is made (section 2.3), to help determine which attributes to include for this research.

2.1 ROUTE CHOICE MODELS

Route choice models are developed to simulate route choice behavior of travelers for mainly network load forecasting, further analyzing travel behavior adaptation and for generating route alternatives in in-vehicle navigation systems. Route choice modeling methods are split into two stages. In the first stage, coined route generation, the choice set is formed by generating possible alternative routes. In the second stage, namely the route choice model, the probability of the observed route being chosen within the generated choice set is estimated. The first stage, route generation, will be further elaborated upon in section 2.2.

Before getting into the details of the models, an indication of the required type of input data is discussed (section 2.1.1). An overview of the different types of route choice models based on their structures is given (section 2.1.2). Based on the scope of this research, a choice is made for the logit structured models out of which the Path Size Logit (PSL) and the Recursive Logit (RL) are further discussed (sections 2.1.3 and 2.1.4). Lastly, the performance indicators used for determining the best model are discussed (section 2.1.5)

2.1.1 Route choice data

Route choice data serves as input for the route choice model structures. This choice data can be categorized as either stated preference- (SP) or revealed preference (RP) data depending on whether a hypothetical choice is made or an actual choice is observed. SP route choice data is gathered by letting participants fill in a survey in which they make route choices based on hypothetical situations and alternatives. RP route choice data is obtained by observing where participants ride in real-world situations.

Until recently, many route choice modelling studies have made use of stated preference (SP) data. An overview of 20 of such studies ranging between 1968 to 2012 is given by Casello & Usyukov (2014). Research on cycling route choice behavior based on revealed preference data (based on GPS route tracks) is increasingly done in the recent years (Dill, 2008; Menghini, 2009; Hood, 2011; Ehrgott et al., 2012; Broach 2012; Casello, 2014; Halldórsdóttir et al., 2014; Khatri et al., 2016; Ton et al., 2017; Ghanayim & Bekhor, 2018; Ton et al., 2018; Prato et al., 2018; Oehrlein et al., 2018; Rupi & Schweizer, 2018; Skov- Peterson et al., 2018). Technological advancements in GPS equipment make it easier and cheaper to get revealed preference route data. It is noteworthy that none of the mentioned cycling route choice studies discuss the route choice behavior of E-bikers.

While SP route choice data is easier to handle for analysis and can be used for choices between alternatives that do not exist yet, it may be less realistic when compared to RP route choice data (Halldórsdóttir et al., 2014). There might be a potential hypothetical bias in SP data as participants might not actually make the same choice in reality. They rely on the limited information given in the survey and they do not feel the actual consequences of the hypothetical choice.

Since the purpose of this study lies in analyzing actual route choice behavior and no new alternatives are compared, RP data is used.

2.1.2 Overview of route choice models

A concise overview of the different route choice model structures found in literature is given by Prato (2009). He classifies the route choice models based on their structures into the following three: Logit-, Generalized Extreme Value- (GEV) and Non-GEV structures.

The *Logit structures* maintain the basic Multinomial Logit model (MNL) structure, but are extended with a correction factor in order to cope with the correlations of route alternatives. The MNL model is defined in the box below.

Multinomial logit model for a finite set of alternatives $j \in \{1, \dots, J\}$

A utility u_j is associated with each alternative and is the sum of a deterministic and a random component $v_j + \mu\varepsilon_j$ where ε_j are i.i.d. extreme value type 1 with scale parameter μ. The maximum utility is $u_{\max} = max_j u_j$ and the expected maximum utility is $Eu_{max} = \mu \ln \sum_j e^{\frac{1}{\mu}v_j}$. It is a general fact for additive random utility models (McFadden, 1978; Fosgerau et al., 2013) that choice probabilities can be found as the gradient of Eu_{max} considered as a function of the vector of deterministic utility components v and hence

$$P_j = \frac{e^{\frac{1}{\mu}v_j}}{e^{\frac{1}{\mu}Eu_{\max}}} = \frac{e^{\frac{1}{\mu}v_j}}{\sum_{j'} e^{\frac{1}{\mu}v_{j'}}} \quad (1)$$

The multinomial logit model satisfies the IIA property, since $\frac{P_i}{P_j} = e^{\frac{1}{\mu}(v_i - v_j)}$, which depends only on $v_i - v_j$. This property makes it possible to estimate the model using only a sample of alternatives in order to reduce computational cost (McFadden, 1978).

The C-Logit and Path Size Logit (PSL) models each have a specific correction factor (commonality – and path size factors respectively) added to the simple MNL structure to cater for overlapping route alternatives. The C-Logit model is found to be outperformed by the PSL model for route choice modelling as indicated by Prato (2009) in terms of likelihood. A more recent development is that of the Path Size Correction Logit, which maintains the same simple formulation as the PSL model but with a Path Size Correction (PSC) factor instead of the Path Size (PS) factor. The PSC factor differs slightly in terms of its formulation and is said to be determined in a more systematic way but is found to have similar performance as the PSL model.

A more recent development with respect to the Logit structured models is that of the Recursive Logit introduced by Fosgerau et al. (2013), which does not require separate route generation by allowing the choice to be made on link level instead of path level. All other route choice models do require route choice set generation as discussed in section 2.2.

The *GEV model structures* deal with correlations in the stochastic part of the utility function, have a tree structure and relate the network topology to the specific coefficients. However, they do not consider taste variation or correlation over time of unobserved factors. Two of such models are the Paired Combinatorial Logit and the Cross Nested Logit. Based on the Cross Nested Logit model, the Generalized Nested Logit has been formalized.

Non-GEV model structures also allow random taste variation and correlation in unobserved factors over time. These models do not present closed-form expression for the choice probabilities. Their estimation is based on simulation and thus more complex than the other models. Two of such models are the Multinomial Probit and the Mixed Logit models.

For this research, heterogeneity of preferences and correlation of unobserved factors are neglected. For estimating the general route preferences of E-bikers, the logit structured models should suffice. The simple and well established (applied in 7 out of the 12 cyclist's route choice behavior studies indicated in appendix A) PSL model (see section 2.1.3) is used in this research. An attempt is made to estimate the relatively new and promising Recursive Logit model (section 2.1.4) as well.

2.1.3 Path Size Logit

Ben-Akiva and Bierlaire (1999) presented the Path-Size Logit (PSL) model for an application of discrete choice theory for aggregate alternatives, already used in other transportation contexts such as destination choice. The PSL model expands on the MNL model by including a penalty attribute for overlapping parts of route alternatives. The PSL model expresses the probability of choosing route j within the set of alternative paths C_n in a similar manner to the simple Logit structure:

$$P_{jn} = \frac{e^{\frac{1}{\mu}v_{jn}+\beta_{PS}\ln(PS_{jn})}}{\sum_{j'\in C_n} e^{\frac{1}{\mu}v_{j'n}+\beta_{PS}\ln(PS_{j'n})}} \quad (2)$$

The correction of the path utility is done by adding the $\beta_{PS}\ln(PS_{jn})$ term to the deterministic utility, out of which parameter β_{PS} is to be estimated. Note that in this formulation for a unique path j, the maximum value for $PS_{jn} = 1$ is obtained and thus resulting in $\ln(PS_{jn}) = 0$. For all other values of the path size factor $0 < PS_{jn} < 1$, the natural logarithm would result in negative values. The less unique a path j is, the smaller the value for PS_{jn}, the more negative $\ln(PS_{jn})$ and thus the lower the deterministic part of the utility function $v_{jn} + \beta_{PS}\ln(PS_{jn})$ will be.

The PS attribute is defined for a path j and choice set C_n, where the sum over all links Γ_j of path j is taken. If parts of paths overlap, the PS attribute will reduce the utility for those paths. If link a of path j overlaps with a path i of the choice set C_n, then δ_{ai} will equal one and zero otherwise. The relative importance of link a for the whole path j, is included by dividing the length of the link L_a by the length of the whole path L_j.

$$PS_{jn} = \sum_{a\in\Gamma_j} \frac{L_a}{L_j} \frac{1}{\sum_{i\in C_n} \delta_{ai}} \quad (3)$$

The PSL model is by far the most popular route choice model for cyclist's route choice behavior analysis, being used in 7 out of the 12 studies indicated in appendix A. This has to do with its relative simplicity and efficiency to estimate choice behavior.

2.1.4 Recursive Logit

The formulation of the Recursive Logit model and an overview of the recent studies in which the model is used are given in this section.

Model formulation

The Recursive Logit (RL) model is considered as a dynamic discrete choice model in which the choice of a path is based on a sequence of link choices (Fosgerau et al., 2013). An individual chooses at each node the utility-maximizing link. The utility is considered as the sum of the instantaneous link cost, the maximum expected utility to the destination and i.i.d. extreme value type I error terms. One requirement for attribute values in the RL model is that they should be link additive. The utility of a path should be the sum of the utility of each of its links. The link probabilities are given by a multinomial logit model and the expected downstream utilities are identified from Bellman equations.

Consider a directed connected graph $G = (A, V)$, defined in terms of a set of links (A) and a set of nodes (V). Consider links k, $a \in A$ out of which $a \in A(k)$ and $A(k)$ being the set off outgoing links from the sink node of k (see Figure 1 for the illustration).

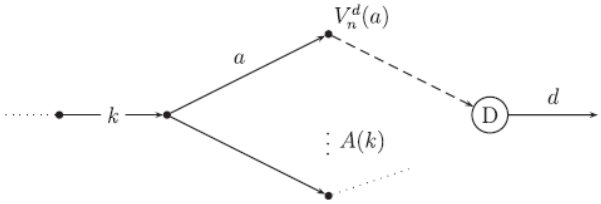


Figure 1. Illustration of notation (Fosgerau t al., 2013)

A path is a sequence of links $\sigma = (k_0, \ldots, k_l)$ with $k_{i+1} \in A(k_i)$ for all $i < l$. Each link pair (k, a) has a deterministic utility component $v_n(a|k)$, based on attributes $x_{n,a|k}$ of the link pair. These may include characteristics of the traveler n. Path attributes are assumed to be link additive. In the terminology of dynamic programming, k is a state and a is a potential action given state k.

The instantaneous random utility for the individual n of a link a conditionally on being in state k can then be defined as:

$$u_n(a|k) = v_n(a|k) + \mu\epsilon_n(a) \tag{4}$$

where $\epsilon_n(a)$ are i.i.d extreme value type 1 error terms with zero mean and μ is a fixed scale parameter.

Consider now an individual traveling from an origin to a destination node. The network is extended with a dummy link to include the destination. More precisely, we define an absorbing state by adding a link d without successors from the destination node (see Figure 1). The deterministic utility is $v(d|k) = 0$ for all k that have the destination as sink node.

The full utility of link a conditionally on being in state k is obtained by adding the maximum expected utility to the destination $V_n^d(a)$ to the instantaneous utility $u_n(a|k)$. The maximum expected utility to the destination is therefore defined by the Bellman equation as follows:

$$V_n^d(k) = E\left[\max_{a \in A(k)} \{v_n(a|k) + V_n^d(a) + \mu\epsilon_n(a)\}\right] \tag{5}$$

Thus, the individual chooses out of all the considered links $A(k)$ the link a which has the highest utility $u_n(a|k) + V_n^d(a)$.

The probability of choosing a link a given state k conditionally on going to destination d is then given by the multinomial logit model:

$$P_n^d(a|k) = \frac{e^{\frac{1}{\mu}v_n(a|k)+V_n^d(a)}}{\sum_{a' \in A(k)} e^{\frac{1}{\mu}v_n(a'|k)+V_n^d(a')}} \quad (6)$$

The value function can then be written as the following logsum:

$$V_n^d(k) = \mu \ln \sum_{a \in A(k)} \delta(a|k) e^{\frac{1}{\mu}v_n(a|k)+V_n^d(a)} \quad (7)$$

To simplify the matrix notation for solving the Bellman equation, indicator $\delta(a|k)$ is introduced. It equals to one if $a \in A(k)$ and zero otherwise.

As mentioned earlier, a path is denoted as $\sigma = \{k_i\}_{i=0}^{l}$, where k_0 is the origin and $k_l = d$. The probability of a path observation is $P(\sigma) = \prod_{i=0}^{l-1} P(k_{i+1}|k_i)$ by the Markov property of the model. Note that the denominator in (4) simplifies to $e^{\frac{1}{\mu}V_n^d(k)}$. The probability of choosing a path σ is thus

$$P(\sigma) = \prod_{i=0}^{l-1} e^{\frac{1}{\mu}v_n(\mathrm{k_{i+1}}|k_i)+V_n^d(\mathrm{k_{i+1}})-V_n^d(\mathrm{k_i})} = \frac{e^{\frac{1}{\mu}\sum_{i=0}^{l-1} v_n(\mathrm{k_{i+1}}|k_i)}}{e^{\frac{1}{\mu}V_n^d(k_0)}} \quad (8)$$

Assume the deterministic utility of the path $v(\sigma) = \sum_{i=0}^{l-1} v_n(\mathrm{k_{i+1}}|k_i)$, than

$$P(\sigma) = \frac{e^{\frac{1}{\mu}v(\sigma)}}{\sum_{\sigma' \in \omega} e^{\frac{1}{\mu}v(\sigma')}} \quad (9)$$

where ω is the set of all possible paths. This set is infinite according to Fosgerau et al. (2013), as long as loops are considered in paths. The RL model is similar in form to the multinomial logit model, but with an infinite number of alternatives. In the form presented above, the RL satisfies the IIA property. This IIA (Independence from Irrelevant Alternatives) property indicates that the ratio of probabilities between any two alternatives only depend on their respective utility values. The alternatives are thus independent from other alternatives. Satisfying this property keeps the model simple.

Two important assumptions made by Fosgerau et al. (2013) for the RL model are that state transitions are deterministic and that the discount factor is equal to one. These assumptions make it possible to solve the value functions as a simple system of linear equations, which can be computed within reasonable time.

$$\boldsymbol{z} = \boldsymbol{M}\boldsymbol{z} + \boldsymbol{b} \Leftrightarrow (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z} = \boldsymbol{b} \quad (10)$$

Where vector $z_k = e^{\frac{1}{\mu}V_n^d(k)}$, matrix $M_{ka} = \delta(a|k)e^{\frac{1}{\mu}v_n(a|k)}$, $b_k = 0$ if $k \neq d$ and $b_d = 1$.

Correlation of utilities

Due to overlapping of paths in the network, the IIA property of the RL model may not be satisfied. Unobserved attributes may be shared among paths with overlapping links and thus leading to inaccurate predictions. Inspired by the Path Size Logit model, Fosgerau et al. (2013) have proposed a so-called Link Size (LS) attribute

that should heuristically cope with the correlation of utilities. In this way the model keeps the logit structure and the IIA property is relaxed. The LS attribute as proposed by Fosgerau et al. (2013) uses the expected link flow as a proxy for the amount of overlap, instead of the number of paths using a given link (as is done in the PSL model). [1]

Recursive Logit applications

Since its formulation by Fosgerau et al. (2013), the RL model has been mainly applied on car data. The dataset on which they tested the model was of a road network of the Swedish city, Borlänge, containing 3,077 nodes and 7,459 links. The GPS data consisted of 1,832 car trips corresponding to simple paths that are longer than four links. The attributes included in the model were: travel time, left turns, crossings, U-turns and the LS attribute.

The same dataset and attributes were used for further development of the RL model in terms of nested structures, correlated structures and even regret minimization models (Mai et al., 2015; Mai et al., 2016; Mai et al., 2017).

Meyer de Freitas (2018) applied the RL model on multimodal route choice behavior within the city of Zurich, including transit, car, bike and walking. Trip data was inferred from Microcensus survey data. Multiple models were estimated consisting of different mode specific attributes. The common attributes were travel time and -costs.

The first published application on cyclist's route choice behavior has been that of Zimmerman et al. (2017). They applied the model on GPS observations within the city of Eugene, Oregon. The data consisted of 648 bike trips after excluding duplicate trips. The network contained 16,352 nodes and 42,384 links. The attributes included are: length, slope, estimated traffic volume, bike facility, crossings and turns. These attributes are all related to the infrastructure of the network and lack attributes related to the cyclist and environment around. In addition to the standard RL model, they also estimated a Nested Recursive Logit (NRL) model without and with LS attribute.

2.1.5 Performance indicators for route choice models

The performance of choice models in general is evaluated based on their fit to the data and their predictive capability. Other generic performance indicators include the computation time for estimation and whether the parameters have a reasonable sign and are within an expected order of magnitude.

Model fit

The *model fit* indicates the likelihood of the observations being chosen with the given parameter estimates. Since the likelihood results in very small numbers, the log-likelihood $LL(\beta)$ of a model is determined as follows:

$$LL(\beta) = \ln(\prod_n \prod_i P_n(i|\beta)^{y_n(i)}) = \sum_n \sum_i y_n(i) \cdot ln(P_n(i|\beta)) \#(1) \quad (11)$$

where $P_n(i|\beta)$ is the probability of alternative i within the choice set being chosen, based on parameter estimates β and $y_n(i)$ equals 1 if alternative i is chosen and 0 otherwise.

McFaddens' adjusted rho-squared measure allows for comparing the final model with the null-model as follows:

[1] The details of this overlap factor are not discussed further in this report, since this research did not even reach that state where the basic RL model could be estimated to satisfaction.

$$\bar{\rho}^2 = 1 - \frac{LL_\beta - K}{LL_0} \quad (12)$$

where LL_β is the log-likelihood of the final estimated model with parameters β, the number of estimated parameters K and LL_0 is the log-likelihood when all the parameters $\beta = 0$. The $\bar{\rho}^2$ can range from 0 to 1 as the model fit increases. This indicator can be used to compare models.

Other measures for comparing choice models in general include Likelihood Ratio Test (LRT) and the Ben-Akiva & Swait test. Both of which test whether the difference in final log-likelihood between two models also holds for the population. The LRT can be applied when comparing nested models. In other words, only if one model can be obtained by constraining parameters from the other model.

The LRT involves calculating a likelihood ratio statistic LRS which is defined as:

$$LRS = -2 \cdot (LL_A - LL_B) \quad (13)$$

Where LL_A and LL_B are the final log-likelihood values for models A and B. This LRS is found to approximately follow a chi-squared distribution. The more attributes a model has, the higher its LL will be compared to its simpler predecessor. To justify whether this improvement is statistically significant with respect to the added attributes, the distribution that the LRS follows, is dependent on the degree of freedom q. In which q is the extra number of additional attributes added to the simple model. The significance level of the improvement is thus dependent on the values for LRS and q within a chi-squared distribution.

Predictive capability

The *predictive capability* indicates how well a model can predict choices on an out of sample choice set. The predictive capability is measured by out of sample testing, which goes as follows. The choice dataset is divided into two sets. One for estimation (between 60-90% of the data) and the other for validation. The model is estimated based on the estimation dataset and used to predict the choices within the validation set. The predicted log-likelihood (log-likelihood of the validation set) or the hit-rate (1 if the observation has the highest probability of being chosen, 0 otherwise) give indications of the predictive capability of the model. Both the predicted log-likelihood and the hit-rate are used in this research.

2.2 ROUTE GENERATION METHODS

In order to estimate path-based route choice models, a set of alternative routes is required as an input. These alternatives are generated using a route generation method. Various route generation methods have been developed over the years. An overview of these methods is given, such that a choice can be made for a method for this research (section 2.2.1). In order to evaluate route generation methods, certain performance indicators related to the generated choice set have been developed. These performance indicators are discussed as well (section 2.2.2).

2.2.1 Overview of route generation methods

Prato (2009) provides an extensive overview of route generation methods used for path-based route choice models as well. He distinguishes four types of route generation methods: Deterministic shortest path-based, Stochastic Shortest path-based, Constrained Enumeration methods and Probabilistic methods.

Most of the presented methods fall under the *Deterministic shortest path-based methods*. These methods generate a route by applying a shortest path algorithm on the network after deterministically modifying variables such as link impedances, route constraints and search criteria. Some well-known methods that fall under this category are: K-shortest path, Labelling approach, Link elimination and Link penalty.

Stochastic Shortest path-based methods also make us of shortest path algorithms on the network, but find optimal paths based on random draws for link impedances and individual preferences from probability distributions. These methods include the Simulation approach and the Doubly stochastic generation function (DSGF). The difference between the two lies in the number of stochastic assumptions. The Simulation approach only assumes a random variation in the traveler's perception of the attributes of a path. The DSGF method also assumes variation in the traveler's tastes.

Constrained enumeration methods assume that travelers not only aim at minimizing cost when choosing a route, but also consider other behavioral constraints as well. Some examples of these constraints include: avoiding routes with large detours, avoiding links that substantially increase the distance or travel time to the destination, not considering similar routes as different routes. Drawbacks of these exhaustive methods include them being computational expensive and the difficulty to determine the thresholds for the different behavioral constraints due to the lack of a theoretical basis.

Probabilistic methods generate routes based on the probability of a route within a network. Frejinger et al. (2009) attached a probability on each link within a network based on the shortest path between each origin-destination pair. Links on the shortest path get a probability of one, while other links get a probability between zero and one depending on their distance or cost to the shortest path. The route probability is the product of its link probabilities. These methods were not mature enough for route choice modeling purposes at the time that Prato (2009) published his review. Undeniably, this probability method by Frejinger et al. (2009) forms the basis for the Recursive Logit model, which is further discussed in section 2.1.4.

The 12 cyclist's route choice behavior studies apply a wide variety of route generation methods with the majority being the Labeling -, BFS-LE - and DSGF methods (see appendix A). Halldórsdóttir et al., 2014 compared, from what they found in literature, the three most promising methods for car route generation in the cycling context. The three methods they compared are: Breadth First Search on Link Elimination (BFS-LE), Doubly stochastic generation function (DSGF) and Branch & Bound (B&B). They applied multi- attribute cost functions by adding different weights to the distance of links with different road types, cycle lanes and land use attributes. They state that multi-attribute cost functions increase the performance of the route generation methods in terms of consistency with observed behavior and heterogeneity of the choice set (further explained in section 2.2.2). However, they do not compare the performance achieved by using the multi-attribute cost functions with that of the basic distance-only function for the same dataset.

The B&B method was outperformed by far by the other methods in terms of consistency, heterogeneity and computational time. The BFS-LE and DSGF method had similar performance in terms of consistency (approx. 65%). The DSGF had a slightly better heterogeneity in terms of route choice set than the BFS-LE method, but the BFS-LE was by far the most computationally efficient of the three (5 min vs 24+ hours).

The BFS-LE method is not the most applied method in the cyclist's route choice studies, which arguably could be related to the fact that none of the studies considered multi-attribute cost functions except for Prato et al. (2018). Due to the great performance and efficiency of the BFS-LE method when using multi-attribute cost functions, it is the route generation method of choice for the PSL route choice model in this research.

2.2.2 Performance indicators for route generation methods

Halldórsdóttir et al. (2014) provide a variety of measures to evaluate the generated route choice set based on its consistency with observed routes and its heterogeneity in composition. The computation time for the different route generation methods is also used in the evaluation of the methods.

The *consistency with observed routes* indicates whether the observed routes are also generated by the method and included in the generated choice set. To determine this consistency Halldórsdóttir et al. (2014) defined an overlap measure O_n in terms of shared lengths between a generated route and observed route as follows:

$$O_n = \frac{L_{overlap_n}}{L_{obs_n}} \tag{14}$$

where $L_{overlap_n}$ is the overlapping length of the route generated with the observed route for an observation n, and L_{obs_n} is the length of the observed route for observation n. Ideally the amount of overlap would be 100%, but the generated routes differ slightly from the observed ones. Therefore, a threshold value δ was included to decide whether an observed route is reproduced by the route generation method or not as follows:

$$RR = \frac{\sum_{n=1}^{N} I\left(O_{n,max} \geq \delta\right)}{N} \tag{15}$$

where RR is the reproduction rate, $O_{n,max}$ is the overlap of a generated route that overlaps the most with observation n, N being the total amount of observations and $I(\cdot)$ is the coverage function, which equals to either 1 or 0 when its argument is respectively satisfied or not.

Prato & Bekhor (2007) extended on the previous measures for the consistency with the observed routes by adding another measure, called the consistency index CI. They compare the amount of overlap $O_{n,\max}$ for the most overlapping generated routes with each observation n, with the ideal overlap $O_{\max}$ of 100% for each observation n as follows:

$$CI = \frac{\sum_{n=1}^{N} O_{n,max}}{N \cdot O_{max}} \tag{16}$$

The *heterogeneity of the choice set composition* indicates how much the generated routes vary from each other. Halldórsdóttir et al. (2014) used the distribution of the path size factor PS_{jn} (as described in equation (3)) over the generated route set as measure for the heterogeneity. The path size ranges between 0 and 1, which indicates how much a route is independent of other routes. For example; if there are two routes that completely overlap, the path size factor will be equal to $\frac{1}{2}$. For three completely overlapping routes that will be $\frac{1}{3}$. The more unique a route is, the higher the path size factor PS. The average path size factor over all choice sets should give an indication of how heterogenous the generated routes are.

2.3 BIKE ROUTE ATTRIBUTES AND THEIR IMPACTS

An overview of studies on cyclist's route choice behavior using GPS data is given in appendix A. The overview includes for each study a description of the dataset, the included route generation method, choice model and the list of attributes with an indication of their relative impacts on the choice behavior. The overview is made based on estimates made on utilitarian trips only. Notice that cyclist's route choice analysis using GPS data has been picking up in the last 10 years. None of the studies includes E-bikers route choice behavior. Most of the studies have found some interaction effects to be of significance, but these are not included in this overview for simplicity. Attributes that the researchers of the respective studies have tested and found insignificant are also indicated in the overview. These insignificant attributes are either mentioned explicitly in their discussion or are assumed if they are left out of the results after mentioning them in the data description. This overview of the relevant attributes is used to determine which attributes to include for this study. Although E-bikers are

expected to have slightly different preferences than conventional cyclists, the same attributes are expected to be of significance in their route choice decisions.

For convenience, the attributes are grouped into the following clusters: infrastructure related -, socio-demographic -, environmental - and land use attributes.

All studies included attributes related to length and bike facility (bike path or -lanes). The rest of the *infrastructure related* attributes (from most common to least common) included junction control (traffic lights and roundabouts), elevation (in terms of up slopes), car traffic volume, turns (left/ right), one-way restrictions, car speed limits and intersections. Some less common infrastructure related attributes include bridges, number of lanes, surface quality (paved/ not) and traffic safety (number of crash incidents).
As expected, all studies indicate that cyclists want to minimize distance during their utilitarian trips and maximize bike facility usage as much as possible. The studies that included the attribute, all indicate a great dislike for positive gradients. Traffic lights seem to be preferred by cyclists. Prato et al. (2018) show that there is a dislike towards traffic lights and that roundabouts are preferred. High car traffic volume is often found to be disliked as expected and seems to be of more significance when combined with turns attributes. Car speed limits seem to have no impact on route choice. There seems to be a relatively small dislike for intersections. One exception is Ghanayim & Bekhor (2018), where the average street length was included as an attribute which is related to the number of intersections. They found a positive impact, which probably has to do with trips being made within the city center where lots of short streets are inevitable. One-way direction streets are disliked, but are often found to be insignificant. In these cities, where cycling has a large mode share, cyclists are allowed to go in one-way directed streets with little disturbance of incoming traffic. Bridges seem to be disliked in general, but bridges with bike facilities seem to be very attractive. These bridges are often inevitable without large detours and therefore often included in the chosen routes.

Cyclist specific (*socio-demographic*) attributes are included in some studies, but are mostly used only to describe the data set. This has to do with either the data not being specified on the individual level (Ton et al., 2017 & 2018) or the sub samples being of too small size to estimate significant preferences (Broach et al., 2012). Hood et al., (2011), Prato et al. (2018) and Ghanayim & Bekhor (2018) did not mention the socio-demographics after model estimation, but are expected to have run into similar problems. Many studies indicate that including these socio demographics would give more valuable insights in the route choice behavior of different groups of cyclists.

Environmental attributes like weather, temperature, wind and daylight have been included in some studies (Broach et al., 2012; Ton et al., 2017; Prato et al., 2018), but are found to be barely of impact on route choice. They are suspected to have more impact on mode choice instead, since these factors are arguably the same for all the route alternatives within a choice set.

Land use attributes, in terms of scenery and building density, have not been included since recently. Ghanayim & Bekhor (2018) included dwelling units per meter and whether a route is near a sea or park. They found all three attributes to have a positive impact on route choice. Prato et al. (2018) included a wide range of land use aspects including, residential (incl. density), industrial, sports and scenic areas. They found that people experience distance less when riding through these areas, especially in combination with high temperatures. Then they have a positive impact. Skov-Petersen et al. (2018) included greenery in the environment and

shopping streets as attributes, out of which they found that only 'shopping streets' was found significant and had a strong negative impact. Participants avoided these busy shopping streets.

A selection has been made from the list of attributes mentioned above. From the list of *infrastructure related* attributes, all attributes are considered in this research except for cycle lanes, traffic volume, car speed limits, bridges, number of lanes, surface quality (paved/ not) and traffic safety (number of crash incidents). Data for these attributes was not available. *Socio-demographic* attributes are not considered since the data consists of a small set of 21 people (out of which socio-demographics of 18 are available) and are therefore not expected to be of significance. *Environmental attributes* like weather, temperature, wind and daylight were not considered, as they are expected to be of little to no significance based on the findings of previous studies and are expected to have more impact on the mode choice than the route choice. *Land use attributes* are included.

The list of the considered infrastructure and land use attributes with their definitions are further discussed in section 3.4.1. The relevance of each attribute is given as well in terms of the number of times the attribute is found significant and the number of studies it is included in.

3 METHODOLOGY

An overview of the methodology applied in this study is given in this chapter (see schematic overview in Figure 2). The methodology explains the whole process on how the get from raw data to the factors and their impact on route choice behavior.

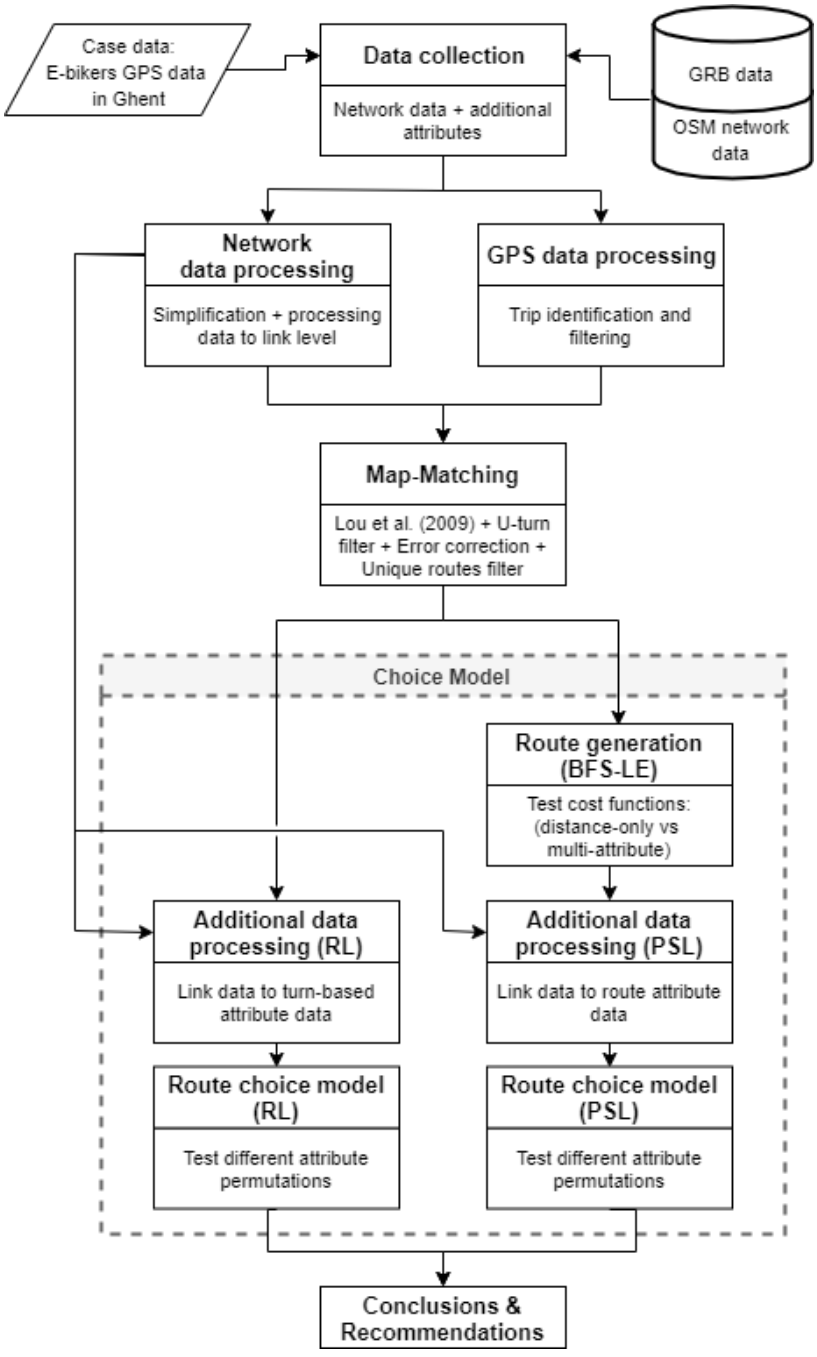


Figure 2. Overview of the methodology.

First the data is collected in terms of GPS observations (pre-collected dataset) and network data. This data is then processed and map-matched. A general description of what processing is needed in order to get the data ready for further analysis is discussed (section 3.1). The map-matched data is used to estimate the choice

behavior with the Recursive Logit (RL) model after some additional data processing (section 3.2). Several attempts were made to get the RL model to converge, without success. The decision was eventually made to not go further with the RL model and move on to the PSL model.

The map-matched observations where used for route generation. The chosen route generation method, Breadth First Search on Link Elimination (BFS-LE), is elaborated on (section 3.3). The test involving the different multi-attribute cost functions is introduced as well. The Path Size Logit (PSL) model with the considered attributes are discussed, along with a description of the process of how the best model formulation (attribute combination) is found (section 3.4).

3.1 DATA PROCESSING

For the purpose of analyzing route choice behavior of E-bikers in the way introduced in section 1 and further discussed in the following sections, GPS data of people using E-bikes for utilitarian purpose is needed as well as the network data for the region where the GPS tracks where recorded. Preferably the data should contain at least 650 utilitarian trips as is the minimum amount used in similar route choice behavior studies on cyclists (see Appendix A).

GPS tracks of E-bikers from the SPRINT project in Ghent (Belgium) was found suitable to estimate the route choice behavior of E-bikers. Data concerning the cyclist GPS route choice behavior, network and the relevant route choice attributes of the area are gathered and processed. A short overview of the processing of the GPS data is given in this section. However, the analysis of the data and the details of the processing procedure are further discussed in section 4.4.

Trips are identified within the GPS data and are filtered. Trips that are recreational, made by car or very short (< 300 m) are removed. Since analyzing panel effects is out of the scope of this research, non-unique trips for the same person are filtered out as well to reduce the impact of the panel data.

Data of the *road network* that is relevant for cyclists is required as well. Highly detailed open source network data is gathered from OpenStreetMaps, because of it being easily available and its high level of detail for the region.

This network data for the more generic attributes (link length, grade, bike facilities, one-way restrictions, etc.) are gathered and processed. The network is simplified and enriched by adding additional attributes in terms of land-use data on network-link level. The inclusion of an additional attribute is based on previous research (attributes that are found relevant and relatively important) and the availability of data. The analysis of the data and the details of the processing procedure are further discussed in section 4.3.

The filtered GPS data is then map-matched to the processed network.

3.2 CHOICE MODEL: RECURSIVE LOGIT

The next step included setting up the Recursive Logit (RL) model. Some additional data processing/ formatting is needed to feed the RL model. Using a small subset of the network and trips, attributes are included to the model. The RL method is link based, thus a small increase in the network size leads to an exponential increase in the number of alternatives. Since the RL model is link-based, attributes must be made link additive for the model to work. This can be done by implementing attributes as interaction terms with length for instance, as has been done by Zimmermann et. al. (2017). The additional attributes are to be included in the model one by one, such that a better model is found.

Optimization and extending the model requires its performance to be measured in a similar way that the best PSL model formulation is obtained (see section 3.4). The model fit in terms of the Log Likelihood and $\bar{\rho}^2$ are

the performance indicators used for determining the best model formulation. The list of performance indicators is indicated in section 2.1.5.

Disclaimer!
Several attempts were made to get the RL model to converge consistently for different initial parameter values without success. Because of the complexity and lack of experience with this model, the exact source of the problem could not be identified within this project. The decision was eventually made to not go further with this model and move on to the PSL model. The results of the attempts are however discussed in section 5.

3.3 ROUTE GENERATION FOR THE PSL MODEL

The next step in the pipeline is the route generation method, which serves as input for the route choice model. As discussed in section 2.1.3, the most promising route generation method for cyclist data seems to be the Breadth First Search on Link Elimination (BFS-LE) method due to its superior performance and low computational effort needed (Halldórsdóttir et al., 2014). This method is thus chosen for route generation in this project.

Inspired by Halldórsdóttir et al. (2014), an attempt will be made to include a multi-attribute cost function to improve the performance of the BFS-LE route generation method. The cost functions (section 3.3.1) and the BFS-LE route generation method (section 3.3.2) are elaborated upon in the following sub-sections.

3.3.1 Cost functions

The idea behind the cost function is that the distance can be perceived differently depending on the attributes of the links. Based on that reasoning, weights are added to the attributes. These weights need to be calibrated. Calibration is done on a limited set of attributes, a small part of the data that can represent the whole data set and with limited predefined discrete weight values. These limitations are set such that the route generation method does not exceed its purpose by estimating the route choice behavior by itself, but does provide a representative choice set for the choice model within reasonable time.

Different cost functions, with each one attribute, are defined and their performances compared. Halldórsdóttir et al. (2014) included attributes related to road types, cycle lanes and land use attributes. They were able to generate routes that were highly consistent with the observed routes. In line with them, the attributes considered are 'Road category' and 'Land use'. Links are put into mutual exclusive categories for each attribute.

The categories for 'Road category' include: cycleway, large road (primary, secondary, tertiary), small road (residential and unclassified) and other (pedestrian/path/tracks, service and unknown). The road categories are further explained in Table 4. The categories for 'Land use' are defined as scenic or non-scenic depending on the amount of the percentage of green- or water area is within the environment of the link. The exact threshold percentage value is determined by visual validation for different threshold values. The processing of Land use data is further explained in section 4.3.6.

The cost function considered for the 'Road category' attribute is:

$$C_a = [\beta_{Cycleway} \cdot Cycleway_a \cdot Length_a] + [\beta_{Large} \cdot Large_a \cdot Length_a] + [\beta_{Small} \cdot Small_a \cdot Length_a] + [Other_a \cdot Length_a] \quad (17)$$

In which the cost C_a for a link a depends on whether it is considered one of the road categories indicated by one of the binary values $Cycleway_a$, $Large_a$, $Small_a$ or $Other_a$ being equal to 1, such that $Cycleway_a + Large_a + Small_a + Other_a = 1$. The relative importance of the length of the first three categories compared to that of the other category are indicated by $\beta_{Cycleway}$, β_{Large} and β_{Small} respectively.

The cost function considered for the 'Land use' attribute is:

$$C_a = [\beta_{Scenic} \cdot Scenic_a \cdot Length_a] + [Non\_scenic_a \cdot Length_a] \quad (18)$$

In which the cost C_a for a link a depends on whether it is considered scenic or not indicated by the binary value $Scenic_a$ being 1 or 0 respectively. The value for Non_scenic_a would be such that $Scenic_a + Non\_scenic_a = 1$. For a scenic link a the weight β_{Scenic} is multiplied with length of the link $Length_a$, indicating the relative importance of the length of a scenic link over that of a non-scenic link. If the link a is not considered to be scenic than the actual length of the link $Length_a$ is considered as cost C_a.

The weights β_k for each type of link k are varied in a discrete manner such that $0.5 \leq \beta_k \leq 2$. The performance of the different cost functions is logged in terms of the route choice set performance indicators, indicated in section 2.2.2. The route choice sets are generated using the Breadth First Search on Link Elimination (BFS-LE) algorithm.

A third cost function is also included which is the sum of the other two cost functions with the most promising weights β_k found for each attribute. This combination of cost functions is proven to increase the performance over the previous separate cost functions by Halldórsdóttir et al. (2014). The best weights for each attribute are decided by the highest Consistency Index CI and heterogeneity in terms of PS found for the generated route set. The performance of this cost function is determined as well and compared to the other cost functions. The best cost function based on this calibration data set is then applied in the BFS-LE algorithm using the whole dataset.

3.3.2 Breadth First Search on Link Elimination

The BFS-LE method (Rieser-Schüssler et al., 2012) involves, as the name suggests, a breadth first search with topologically equivalent network reduction. The method seems to be highly computational efficient while including enough variety in routes (Halldórsdóttir et al., 2014).

The method uses a shortest path algorithm (e.g. Dijkstra, A-Star, etc.) to determine the least cost path between origin and destination. Links of the least cost path are removed one by one from the network, while at each link elimination a new least cost path for the resulting network is determined. When all links of the original least cost path have been removed, the algorithm goes one level deeper by again eliminating links from these reduced networks and determining new least cost paths. The routes generated from these sub-networks at the first depth level are first filtered to keep only the unique routes. All these sub-networks form nodes of a tree (see Figure 3). The tree expands at each depth level, if new least cost paths are found. When a sub-network becomes disconnected and no path can be generated between the origin and destination, this node cannot be expanded in the next depth level and thus is referred to as a leaf node.

The decision to stop the algorithm is set to when the predefined number of unique routes has been generated, the time abort threshold is met or no more feasible routes between origin and destination are present. The predefined number of unique routes to be generated is the maximum number of alternatives that are to be considered in the choice set. This choice set size is set to 21 including the observation (if not generated), which is in line with previous studies (Halldórsdóttir et al., 2014 and Ton et al., 2018).

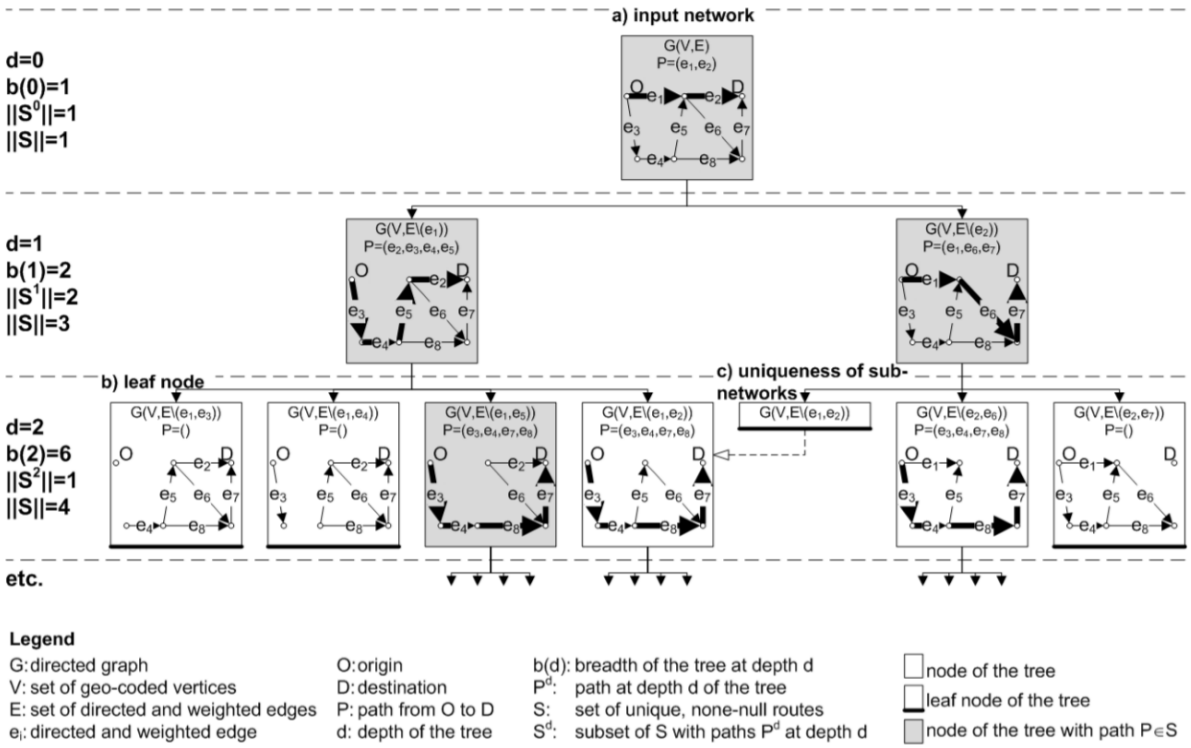


Figure 3. Simple illustration of the process of the BFS-LE algorithm in the form of a tree for two depth levels. Source: Rieser-Schüssler et al. (2012)

The A-star algorithm, that is used to determine the shortest path, is a generalization of the Dijkstra algorithm, such that an extra term is added to the path cost indicating the direction of the destination. This extra term significantly speeds up the path finding. An A-star algorithm with the extra term based on the Euclidian distance to the destination is used.

The algorithm is adapted in order to further speed it up even more. Generating the least cost paths for each adjusted network at each depth d, takes lots of computational effort. This effort is even unnecessary, since only a random set of unique routes will be chosen out of all the least cost paths generated. This waste of effort increases at each depth level, since the amount of tree nodes increases (see Figure 3). The algorithm is adapted such that the randomization of routes takes place before calculating the least cost paths instead of afterwards, and thus reducing computational effort. The order of the tree nodes at depth level d to consider for calculating the least cost path is randomized. Least cost paths are added to the route set if they are unique. This continues for depth level d until all the tree nodes have been considered or enough unique routes have been generated. Only if there are not enough routes generated, the algorithm continuous to the next depth level $d + 1$. The principle of having random diverse routes effectively still holds.

3.4 CHOICE MODEL: PATH SIZE LOGIT

After the routes are generated, the choice behavior can be estimated. As mentioned earlier, the most used route choice model used for cyclists is the Path Size Logit (PSL) model (see Appendix A). The model is proven to work well for this type of data and thus is the chosen path-based route choice model for this project. The basic model formulation is given in section 2.1.3 equation (3).

A list of attributes is made to consider for the model (section 3.4.1), then the procedure for testing different model formulations based on the attributes is discussed (section 3.4.2) and finally the method of measuring the performance of the different model formulations is discussed (section 3.4.3).

3.4.1 Attributes

A selection has been made from the list of attributes considered in literature (see section 2.3). From the list of *infrastructure related* attributes, all attributes are considered in this research except for cycle lanes, traffic volume, car speed limits, bridges, number of lanes, surface quality (paved/ not) and traffic safety (number of crash incidents). Data for these attributes was not available. *Socio-demographic* attributes are not considered since the data consists of a small set of 21 people and are therefore not expected to be of significance. *Environmental attributes* like weather, temperature, wind and daylight were not considered, as they are expected to be of little to no significance based on the findings of previous studies. *Land use attributes* are included.

The list of the considered infrastructure and land use attributes with their definitions is given in Table 3. The relevance of each attribute, based on previous studies mentioned in Appendix A, is given as well in terms of the number of times the attribute is found significant and the number of studies it is included in. These attributes are the most included attributes in all the related studies and are also chosen based on availability of the data.

Table 3. Attribute descriptions.

Attribute	***Model Definition***	***Unit***	***Relevance***
Distance	Total distance (km)	km	12/12
Cycleways	Proportion of route distance on cycleways		12/12
Large roads	Proportion of route distance on large roads (primary-, secondary- and tertiary streets)		12/12
Small roads	Proportion of route distance on small roads (residential and unclassified streets)		-
Up-slope (2-4%)	Proportion of route distance with link up-slopes between 2-4% (i.e. 2-4m/100m elevation increase on links)		7/7
Up-slope (4-6%)	Proportion of route distance with link up-slopes between 4-6% (i.e. 4-6m/100m elevation increase on links)		7/7
Up-slope (>6%)	Proportion of route distance with link up-slopes over 6% (i.e. >6m/100m elevation increase on links)		7/7
Left turns	Number of left turns per km on route ($45° < \delta\alpha < 180°$)	#/ km	6/6
Right turns	Number of right turns per km on route ($-180° < \delta\alpha < -45°$)	#/ km	6/6
Intersection crossings (no turn)	Number of intersection crossings per km on route ($-45° \leq \delta\alpha \leq 45°$)	#/ km	3/5
Roundabouts	Number of roundabouts per km on route	#/ km	1/1
Traffic Lights	Number of intersections controlled by traffic lights per km on route	#/ km	6/7
Wrong way	Proportion of route distance in opposite direction of one-way streets		3/5
Green area	Weighted proportion of buffer area around streets covered with green		2/3

Water area	Weighted proportion of buffer area around streets covered with water	0/0*
Building area	Weighted proportion of buffer area around streets covered with buildings	3/3

α= the angle between two connected links

** not directly included in previous studies, but as combined scenic attribute with greenery*

The road categories are the same used for the cost functions for route generation. These are cycleway, large road (primary, secondary, tertiary), small road (residential and unclassified) and other (pedestrian/path/tracks, service and unknown). The road categories are further explained in Table 4.

The up-slopes are determined by dividing the elevation difference (in meters) between start- and end nodes of each link by the length of the link (in meters). The classification of the up-slopes into 2-4%, 4-6% and >6% is in line with previous studies (Broach et al. (2012), Zimmermann et al. (2016) and Prato et al. (2018)).

Information on turns on the route is calculated based on the angles of the start- and end nodes of the links. The difference $\delta\alpha$ between each of the angles $\alpha_l \; and \; \alpha_k$ of two connected links k and l are determined, while also considering entering a link in the opposite direction. Depending on this angle $\delta\alpha$, the turn is classified as left, right or no turn (see Figure 4).

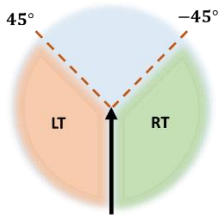


Figure 4. Turn angle criteria illustration.

Gathering and processing of the land use attributes in terms of green-, water- and building areas is discussed in section 4.3.6. The percentage of area of each of the three land use values within a buffer area of a link is determined. This percentage is then multiplied by the link length and summed for the whole route for each land use attribute. This sum product attribute value is normalized by dividing it with the total length of the route.

3.4.2 Model formulation

The PSL model is estimated with PandasBiogeme (Bierlaire, 2018). The utility function of the PSL model is expanded step by step. First, the most relevant attributes are added separately to the length and PS factor to get a 'feeling' of the impact of an attribute on the model fit. Then, different permutations of the attributes are tested, while keeping track of the model fit and the significance of the estimated parameters.

3.4.3 Model performance

The performance of the models in terms of the model fit and predictive capabilities are measured as described in section 2.1.5. The choice dataset is randomly split into an estimation set (70% of all choicesets) and a validation set (the remaining 30% of the choice sets). The model fit in terms of the loglikelihood and $\bar{\rho}^2$, is registered for each tested model using the estimation set. The performance in terms of the model's predictive capability is only done for the final model by making use of the validation set.

4 DATA COLLECTION AND ANALYSIS

This chapter should provide insight of the data used in this study and how it is processed such that it can be used for the route choice analysis.

GPS tracks of E-bikers from the SPRINT project in Ghent (Belgium) was found suitable to estimate the route choice behavior of E- bikers. A short introduction on the SPRINT project is given (section 4.1). Data concerning the cyclist GPS route choice behavior, network and other relevant route choice attributes of the area are gathered (see section 4.2) and processed. The network is simplified and enriched by adding additional attributes in terms of land-use data on network-link level (section 4.3). The inclusion of an additional attribute is based on previous research (attributes that are found relevant and relatively important) and the availability of data. The GPS tracks are processed as well. Trips needed to be identified and are filtered such that only trips that are of interest are included (section 4.4). The filtered GPS data is then map-matched to the processed network (section 4.5) and some descriptive statistics are given of the observations, including each considered attribute and participant (section 4.6).

4.1 INTRODUCTION SPRINT PROJECT

The local authorities in Belgium, like in many other countries, are stimulating cycling more and more. Research to develop effective cycling policy is wanted by these authorities. This involves research on travel behavior of cyclists or of potential cyclists, and factors influencing cyclists' decisions on destinations, whether to bike or not, and on route choices. The SPRINT project of the Policy Support Centre on Traffic Safety, KU Leuven, VITO and the University of Hasselt is one of such research projects which aimed to understand the role of the electric bike. The key goal of the SPRINT project was to understand whether the e-bike may represent a valid alternative for commuting (or functional) trips in the Flanders region (Astegiano et al., 2017). The GPS tracks from E-bike users in the Flanders region was recorded from March 2014 to Aug 2015. Section 4.2 provides more information on the data collection as part of the SPRINT project.

4.2 DATA COLLECTION

The dataset consists of survey results and GPS tracks both capturing the route choice behavior of a group of E- bikers who volunteered to install a GPS system on their bike. The participants of the SPRINT project were initially recruited in Ghent and later also from Leuven, resulting in a total of 61 participants. Out of these 61 participants, 42 explicitly gave their permission to us their data for this follow-up research. They each had access to a personal webpage where they could see their GPS-routes, access related surveys and other cycling and research related tools (such as maps of public re-charge spots, newsletters and forums). No tools for route planning were provided.

There are many more trips logged in the Ghent region since the data was gathered over a longer period. For simplicity, this research focuses only on the Ghent region and only makes use of the data of the 21 participants that made trips within and around Ghent. Data related to the rest of the participants of the Leuven region are excluded.

The data collection of the online-surveys (section 4.2.1), GPS tracks (section 4.2.2) and network data (section 4.2.3) are all addressed in the following subsections.

4.2.1 Online surveys

A survey was performed before the GPS data was gathered. The survey collected socio-demographic information, as well as pre- and post E-bike acquisition information about mode choice for different trip

purposes and the frequency of the mode choice. This survey was not obligatory and resulted in 18 out of the 21 participants that made trips in the Ghent region filling in the survey.

An overview of the social demographics of the 18 participants is given in Figure 5. This overview should give an impression of the composition of the group of participants. The sample is homogenously distributed w.r.t. gender; the age is spread between 25 and 60, with the biggest group between 53 and 60 years; there is a big variety in terms of income; almost all participants have an office job (white-collar employees); most of the participants have a driving license and own their E-bike for longer than a year.

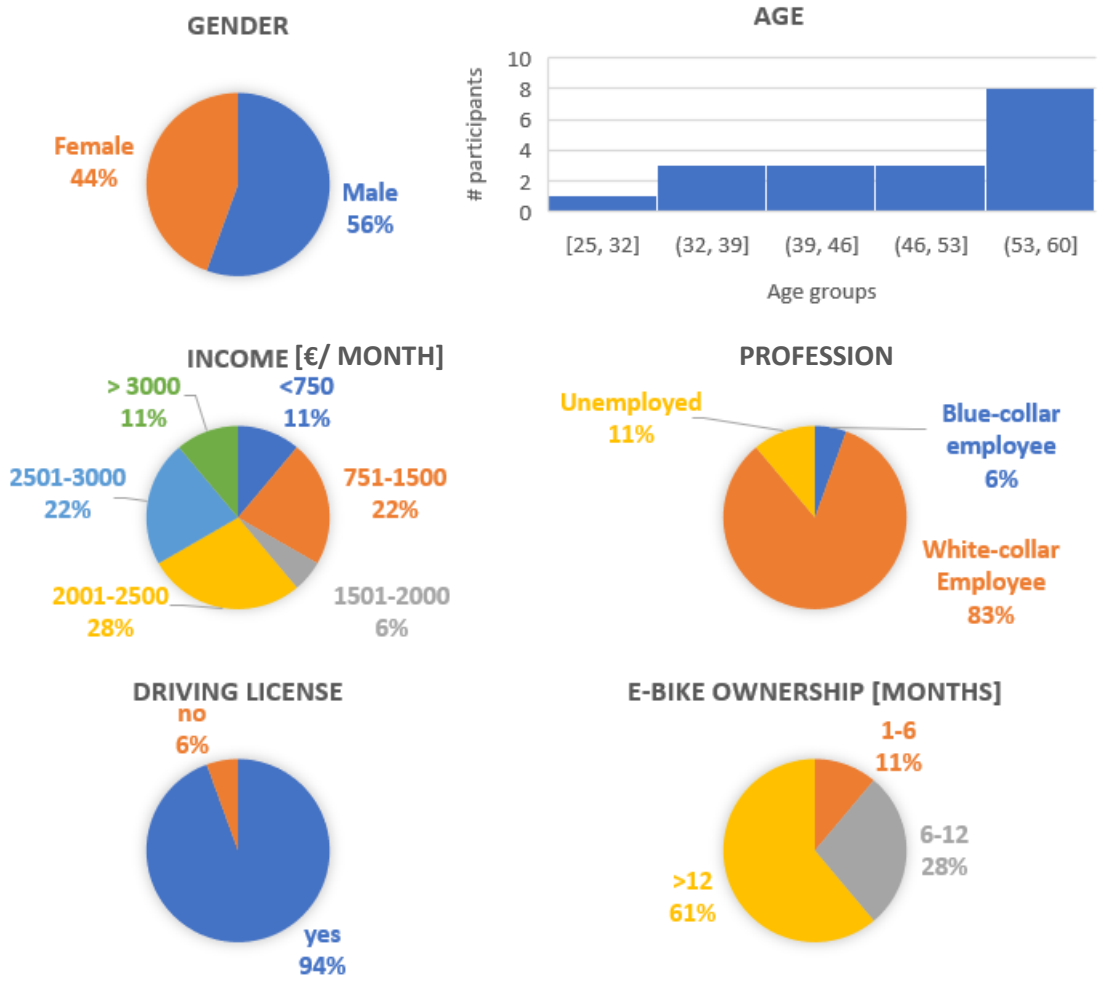


Figure 5. Social Demographic Information

4.2.2 GPS route data

The location data of the E-bikes in the SPRINT research were automatically logged by a GPS tracking device (Erco & gener: GenLoc41e), which was directly connected to the E-bike's battery (see Figure 6). Whenever the E-bike was turned on, the GPS tracker would turn on automatically as well and start logging the location every 5 seconds until the E-bike as turned off. Participants did not have to adapt their normal routine of riding their E-bike in order to start or stop the GPS logging. Each participant was tracked for 30 weeks on average and no travel diary or any other user's annotations were required.

The original dataset contains GPS tracks consisting of 3,694,746 location points mainly in Flanders, made by the 42 people. The data has been collected March 2014 to Aug 2015. Each GPS tracker had a specific device

ID and logged for each observation the time, speed, GPS accuracy, number of satellites and heading (in degrees) along with the exact location (latitude, longitude, altitude).

Figure 6. Tracking device set-up(a) GPS logger GenLoc41e. (b) battery connections. (c) installed logger. Source: Astegiano et al. (2017).

4.2.3 Network data

Data of the road network of the considered region that is relevant for cyclists is acquired. Highly detailed open source network data is obtained from OpenStreetMaps (OSM) using the OSMnx tool (Boeing, 2017). This data is used because of its open availability and high level of detail for the considered region. This OSM network data includes link-specific information such as link length, road category, bike-way, one-way restriction, car speed limit, number of lanes and whether a link is part of a roundabout, tunnel or a bridge. It also includes node-specific information such as traffic signals, pedestrian crossings and bus stops.

The OSM data contains highly detailed network data (see Figure 7). Every link is straight and thus multiple straight links are included to form a curved road. This leads to many unnecessary nodes on a single road that do not represent an intersection. The OSMnx tool includes an option to simplify the network topology by removing nodes where streets curve (node degree $d = 2$) and thus keeping only nodes that represent actual intersections (node degree $d > 2$). This option is used when downloading the OSM network data. The resulting network of the considered region around Ghent contains 61,543 links and 45,280 nodes. The exact measurements of the region are chosen such that most trips are included (further discussed in section 4.4.9).

Figure 7. Overview of considered region and impression of the detail of the OSM network data.

The road categories are described briefly, based on their description on the OSM wiki page[2] and visual checks in Google Maps, in Table 4. An additional 'unknown' road category is added to those links without any of the mentioned categories.

Table 4. Descriptions of road categories.

Category	***Short description***
Primary	Mostly separated directions 1-2 lanes, 70 (50) km/h, cycle lanes/ path, asphalt.
Secondary	Often cycling lane included, 50 (70) km/h, asphalt.
Tertiary	No bike facility, 30-50 km/h, asphalt.
Residential	No bike facility, 30 km/h, bricks/ asphalt.
Pedestrian/Path/ Tracks	Mostly narrow (unpaved) pedestrian tracks trough parks and rural areas.
Service	No bike facility. Alley, bus route, parking's, private access roads, etc.
*Trunk **	The most important roads in a country's system that aren't motorways. Cycling prohibited, 90 km/h, separated directions. Including ramps.
Unclassified	Lowest category road. Mostly rural narrow roads.

** to be removed from cycling network*

Elevation data has been gathered for each node in the network by using the Google Maps API within the OSMnx tool. The grade of a link is determined based on the difference in altitude of the nodes. A map of the elevation data is given in Appendix C and D. Additional information of the environment of the links is gathered from the 'Grootschalig Referentie Bestand of Basiskaart Vlaanderen' (GRB)[3] in terms of land use and buildings (see Figure 8).

[2] Extensive descriptions of all the categories used in OpenStreetMap can be found on this wiki page.
[3] The GRB data is openly available on the following website: https://www.geopunt.be/kaart

Figure 8. Impression of the land use raster data obtained from the GRB with different types of vegetation (each different shade of green representing), buildings (red), water (blue), roofless built areas (brown) and streets (grey).

4.3 Data Processing: Network Data

The gathered network data is processed such that it is useful for the analysis in this research. The level of detail of the gathered OSM network data is still found to be too high. There are many complex intersections consisting of many short links, many parallel links and looping links. These details are found to be problematic during the route generation process (for both the PSL model as well as within the RL model). Therefore, the network data is simplified even more. The network simplification process involves simplifying road categories, roundabouts, detailed intersections, parallel links and looping links (sections 4.3.1 - 4.3.5). The details of the actual implementation of the network simplification in terms of the used MATLAB code are given in Appendix F. A schematic impression of the process is given in Figure 9.

The environmental data is also processed, such that link-level information is obtained (section 4.3.6).

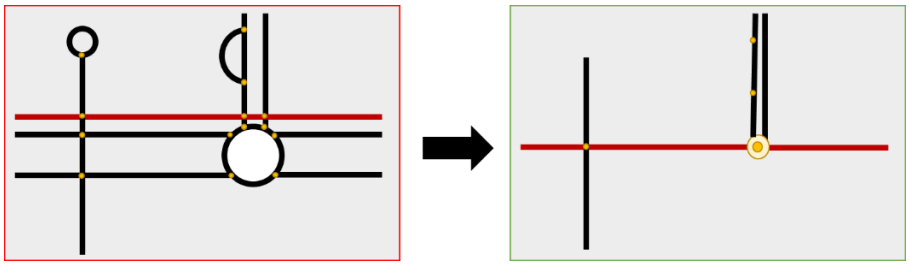


Figure 9. Schematic impression of network before (l) and after (r) simplification process with roads (black links), roundabout, nodes (yellow dots) and cycleway (red links).

4.3.1 Simplify road categorization

Due to the simplification of the network nodes done by the OSMnx tool (Boeing, 2017), links without any intersection between them are combined. The categorical information of these links is combined as well. If these links had different road categories, both categories where combined. These combinations are simplified to the categories given in Table 4. Trunk related links are removed, since it is illegal to cycle on these roads in Belgium. Bridleways, bus stops, steps and planned links are removed as well. An extra ‘unknown’ category is added as well for links without category.

The other attributes such as length and one-way restrictions were correctly combined and did not need further processing.

4.3.2 Simplify roundabouts

Roundabouts consist of multiple links in a circle shape with an intersection category of 'Roundabout'. These links that are clustered, removed and reduced to a single central node with the intersection information of a roundabout. If one of the nodes in the roundabout cluster contained a traffic signal, this information is transferred to the remaining central node.

4.3.3 Simplify intersections

The raw OSM data is detailed such that bikeways and separate lanes are included as separate links. Nodes are included wherever these separate links intersect and interaction between them is allowed for. This level of detail results in complex intersections with multiple separated lanes and bikeways to have lots of short links. These big intersections become less realistic as they appear to consist of many small intersections. These short links could also potentially cause problems for the stochastic shortest path finding algorithms considered in this project, as small loops might be found to have a positive utility. For these reasons, these separate clusters of short links are removed, and all their nodes are combined to one central node. The intersections are identified by a cluster of short links shorter than 20m with at least one link shorter than 10m. The info about the type of junction in terms of roundabouts and traffic signals is preserved when combining nodes.

4.3.4 Simplify parallel links

Roads with separate bikeways are simplified to only one bikeway link (or two if one-way bikeways on both sides or the road). Roads with separated driving directions with no intersections in-between either one of the links are also simplified to one bidirectional link. Parallel lanes which are separated from the main road and rejoined shortly after, e.g. kiss and ride lanes and bus lanes, are removed as well.

These parallel links are only easy to identify after the intersections are simplified to a single node. Links with the same start- and end nodes are identified as parallel links to be simplified. This simplification of parallel links is done multiple times until all parallel links are removed.

4.3.5 Remove looping links

Links with the same start node as end node are removed from the network, as they are of no benefit for the project.

4.3.6 Land use data on link level

The process of gathering land use data on link level is done within the GIS software, QGIS[4]. An overview of the process is given in Figure 10.

The land use raster data is vectorized and simplified. The different categories of vegetation including different grass types, trees and agricultural fields are combined to one category 'Green'. The land use data within a buffer around each link is gathered (see Figure 11), after which the percentage of area of Green, Water and Buildings within the buffer area of each link is calculated. This percentage is then stored on link level.

A rather arbitrary value for the buffer distance of 30m is chosen, such that buildings and water along the roads are captured along the link (axis of the road). In rural areas, the buildings are positioned further away from the streets as can be seen in Figure 11. The full land use map of the area around Ghent considered, is included in Appendix B.

[4] The QGIS process file and data is available upon request via email: surajkm@live.com

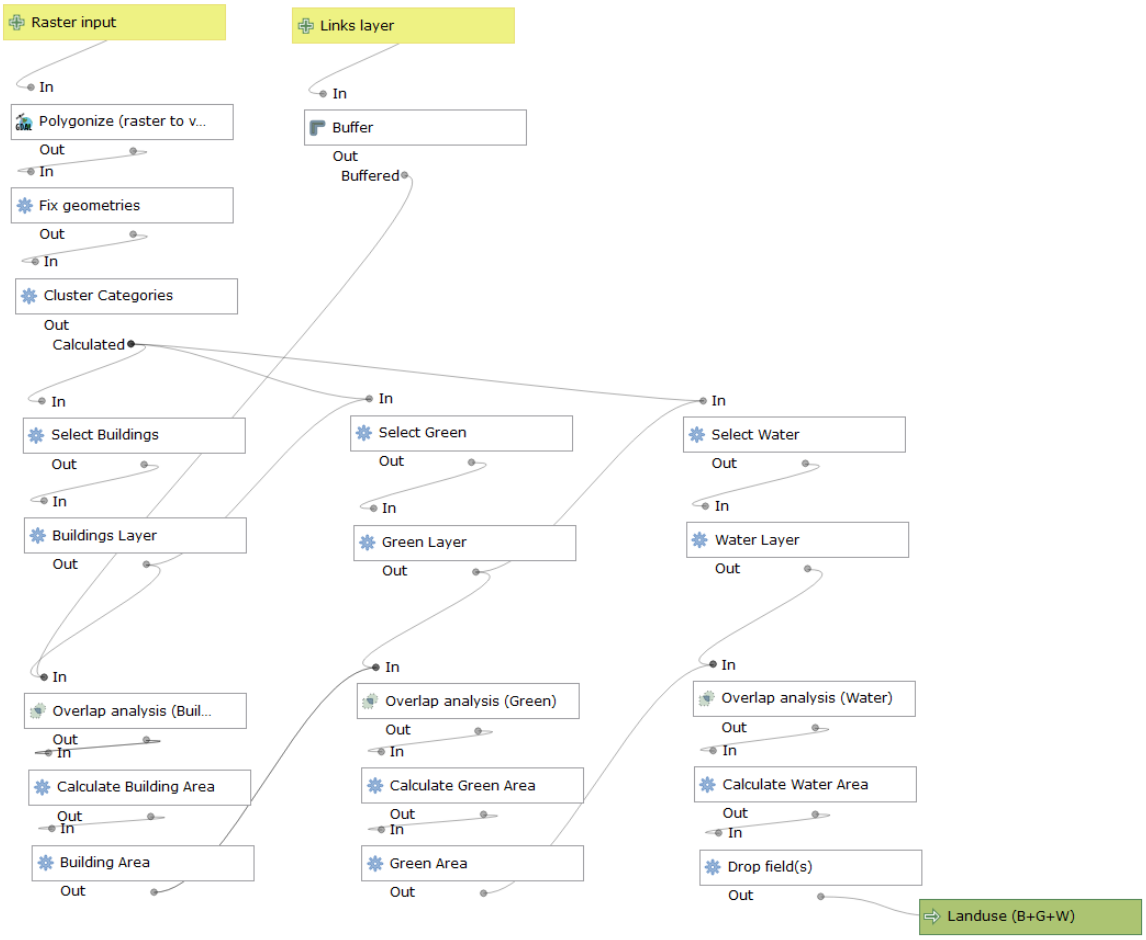


Figure 10. Overview of process of gathering land use data on link-level in QGIS.

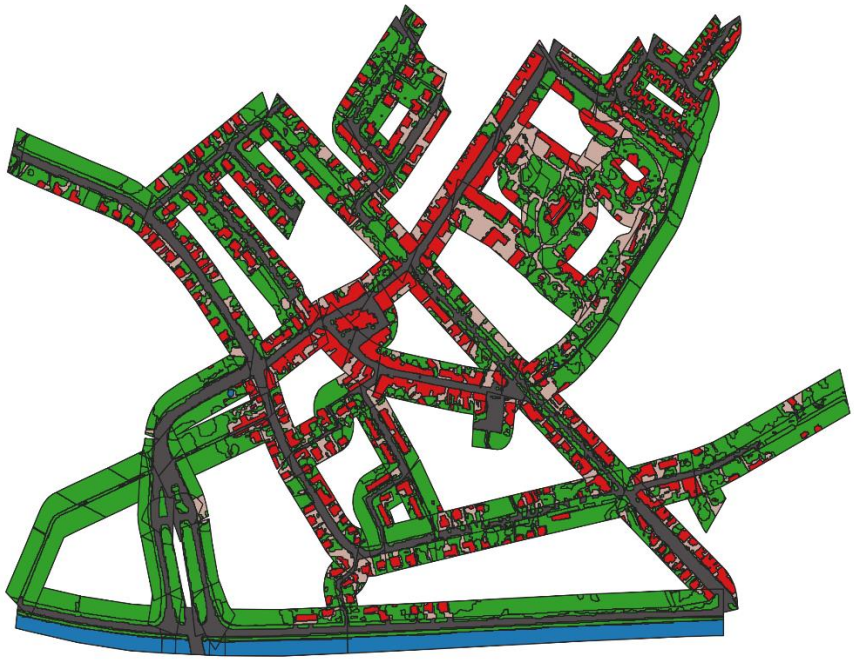


Figure 11. Impression of land use data within 30m link buffers of a small part of the network.

4.4 DATA PROCESSING: GPS DATA

The original raw GPS data including both the Ghent and Leuven regions consists of 3.7×10^6 GPS observations for 42 participants. The registered data logged per observation are: device/user id, coordinates (latitude, longitude, altitude), accuracy in meters, number of satellites, heading degree, speed and the time. The observations were not clustered into trips. The dataset was already filtered on missing data.

The participants are all based in Ghent and Leuven. Nonetheless, the GPS traces are quite widespread across Belgium, Netherlands and France (see Figure 12). After having a closer look at the Flanders region (Figure 13), many arguable recreational trips can be observed. These are the very long trips and tours made by a handful of participants. Some of these traces seem to be following highways and have relatively large distances between consecutive observations, indicating high speeds. These could be traces of car trips with the bike mounted on the car and the GPS tracker on. These very long (car) trips and tours need to be identified and filtered out of the dataset, as these trips are considered to be of recreational purpose. This study aims to only consider route choice behavior for non-recreational (utilitarian) trips.

After having a closer look into the speeds and distances of the trips, some trips have considerably low average speeds which could be identified as trips made walking next to the E-bike. These are not of use for this study as well.

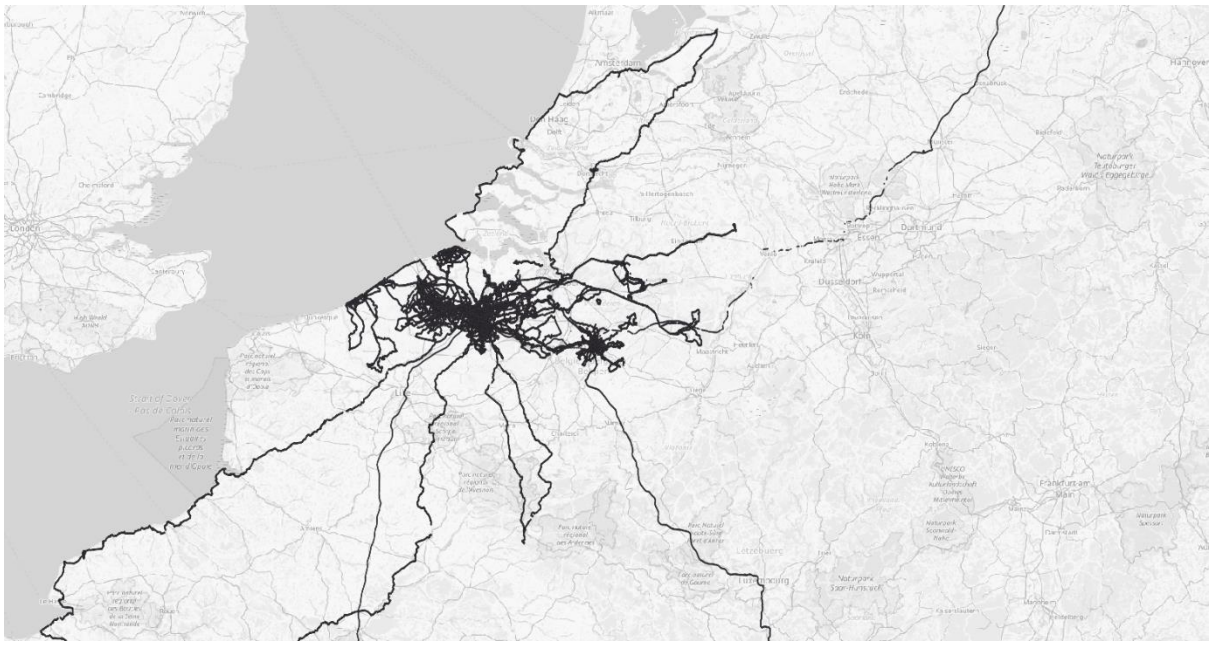


Figure 12. Raw GPS traces overview.

The region considered in this research is initially bounded by squares of 30×30 km^2 around the city of Ghent (see Figure 13). This region is chosen considering that a sufficient large number of utilitarian trips falls within the region and that a bigger network would result in much bigger computational effort needed for route generation and estimating the RL route choice model.

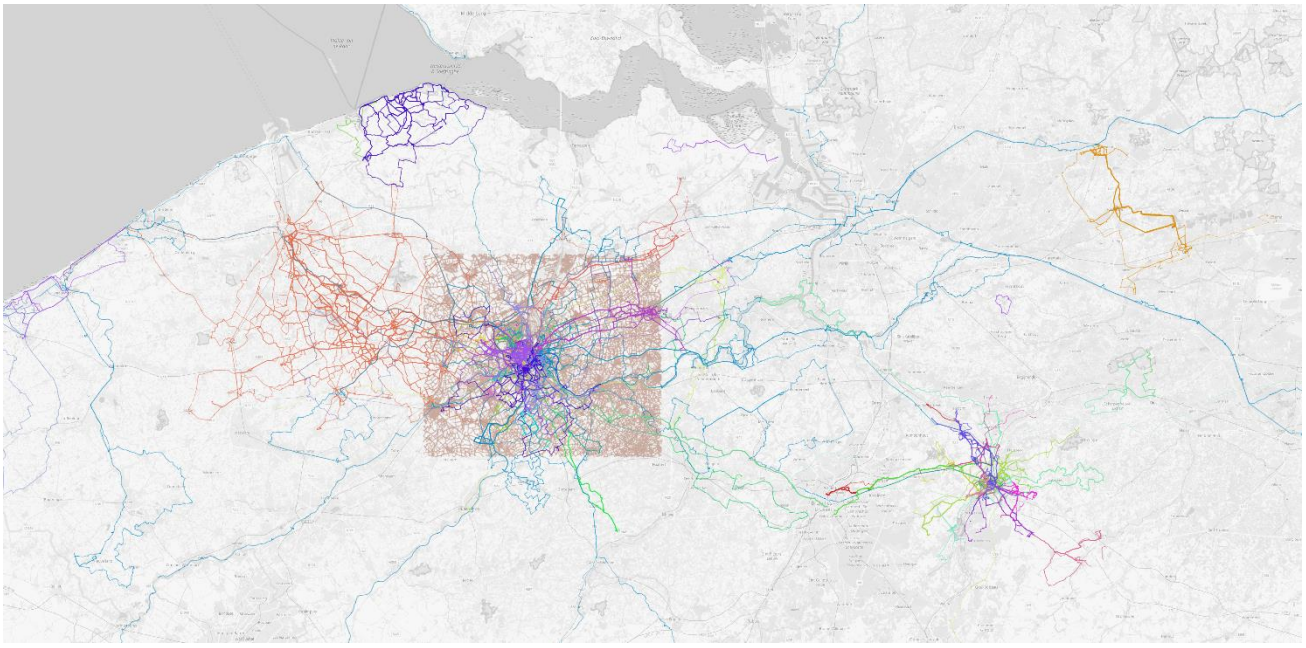


Figure 13. Raw GPS traces overview Flanders (color per device id) with the considered network around Ghent (bronze).

It is evident that some processing is needed to filter out the data which is of interest for this research. The overall procedure for filtering out the data is described below and the details are discussed in the subsections.

The GPS data processing procedure is visualized in Figure 14. The coordinates of the raw GPS data are first transformed into the Lambert 72 (local coordinate scheme) coordinates. This allows for conveniently measuring the distance between points in meters.

The trips first need to be identified, after which unwanted trips can be filtered out. This trip determination can be done based on the user id and time difference between observations (section 4.4.1).

Some trips include some GPS observation points that are much further way from other observations within the trip (a.k.a. random jumps) and require substantial accelerations to reach the consecutive observation locations. These points are identified as errors, which could be caused by satellite or receiver issues. Observations points within trips with errors are removed (section 4.4.3). This error filtering is intentionally done after trip identification, such that trips are not misidentified.

As discussed earlier, trips that are made by car, recreational, short, by walking or outside of the considered area can be filtered out (sections 4.4.2-4.4.7). Since the GPS error filter is based on accelerations, car trips are filtered out before the error filter.

Since the trips are widespread, filtering trips outside the considered area is expected to have the biggest impact and is therefore left as last trip filtering step. The order of the other trip filters does not matter as much, since there will be no cutting of trips.

Very short trips (length< 500m and duration< 3 min) will not add much extra value while estimating the route choice model and are therefore removed (section 4.4.5). Lastly, the traces of each trip are simplified such that map-matching can happen computationally faster (section 4.4.8).

The results and impacts of each filter are discussed (section 4.4.9).

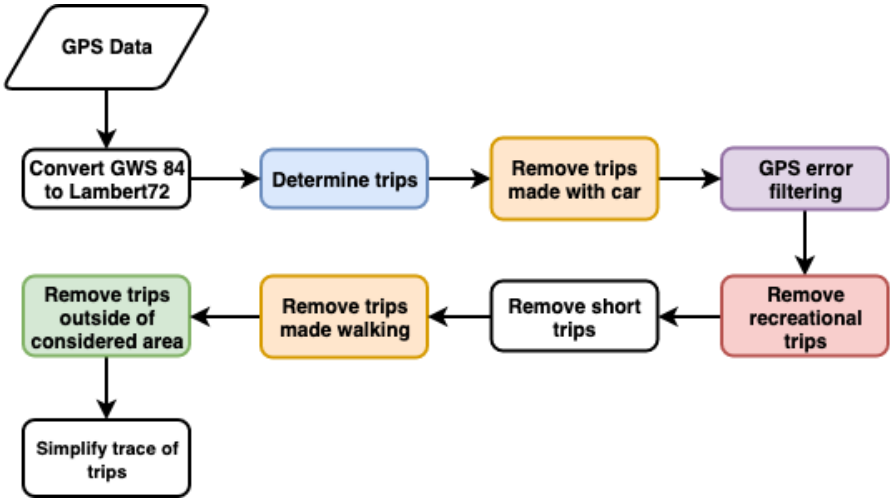


Figure 14. Overview of GPS data processing procedure.

4.4.1 Trip identification

Trips are identified based on the user id and the time difference between consecutive observations (Figure 16). Whenever the e-bike is turned off, the GPS device switches off and starts logging the position when the e-bike is turned on again. Whenever this time difference between these two observations $dt = t_i - t_{i-1}$ is large, it is assumed that a participant started a new trip. The threshold value for t_{dwell} in previous researches vary between 2, 3, 5 and even 10 minutes (Gong et al., 2014). Noteworthy is that this previous research on trip identification is mainly applied on car GPS data. The threshold values for dwell time t_{dwell} of 2 and 3 minutes are arguably too short for identifying trips on a bike in this case. A large time difference between observations can also be caused due to signal loss when driving into a tunnel for instance or due to filtered-out observations with GPS errors. With lower speeds, the time difference for cyclists would be bigger in case the signal is lost due to a tunnel than for cars. It is important that the trip identification process is not too strict, such that splitting of recreational trips or trips with errors in the observations into smaller trips is minimized. On the other hand, it should not be too relaxed such that trips with a quick activity in between (e.g. quick stop at the supermarket) are combined as a tour. These trips will be mistaken for tours and thus be filtered out in the recreational filter.

Figure 15 gives an indication of how the time differences between consecutive observations are distributed over the dataset. Most of these time differences lie below 10 seconds, as expected from GPS data on time intervals of 5 seconds. Arguably a new trip starts after a gap of 15 minutes between consecutive observations, as it is long enough for an activity in between trips to happen (e.g. shopping). Time differences below 2 minutes could be due to GPS connection failures. The 2-5 minutes interval has a very small share of the pie and thus any t_{dwell} within that region is not likely.

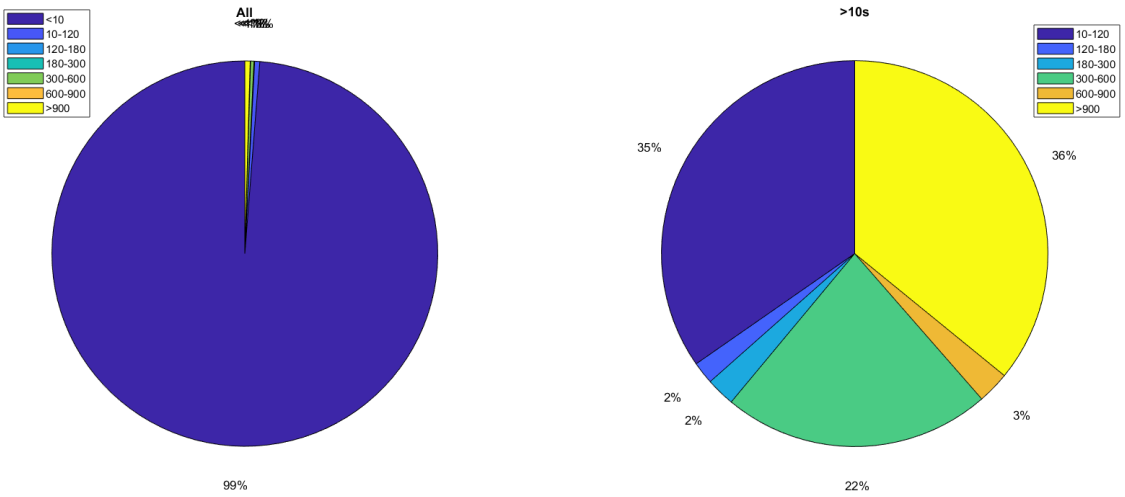


Figure 15. Pie chart of time difference between 2 consecutive observations in seconds with-(l) and without (r) the [0-10] seconds interval.

A visual validation check is done for 10 random samples of 10 sets of 3 consecutive trips. If a new trip started within $dt = t_{dwell} + 20\%$ and would not form a tour when combining the trip with the previous one, the trip would then be identified incorrectly. Thus, indicating that a higher t_{dwell} is needed. Also, if an observation within a trip would have a time difference of $dt > t_{dwell} - 20\%$, then that trip should have been split into 2 trips. Thus, indicating that a lower t_{dwell} is needed.

The visual validation check resulted in 93% of trips correctly determined for the value of $t_{dwell} = 5\text{min}$.

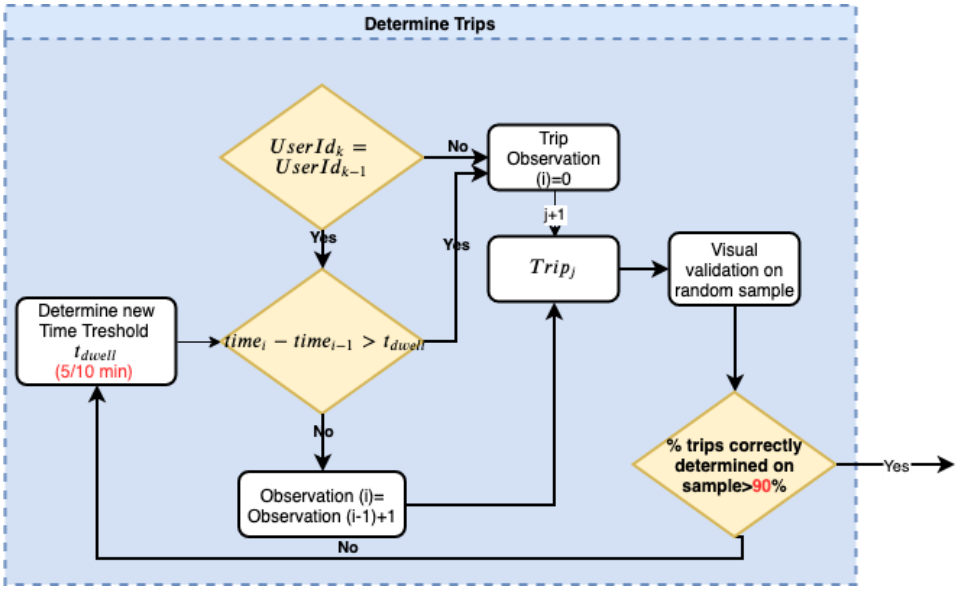


Figure 16. Trip identification procedure.

Additionally, heading information of each GPS point can be used for improving trip determination. This information was available but not used in this study. A big change in heading could also indicate a trip destination.

4.4.2 Remove car trips

Some of the trips seem to have quite high average speeds, even for speed-bikes that can go up to 45 km/h (Figure 17). Some of these trips seem to be made on highways, confirming that they are indeed trips made by car. Notice the distribution has two peaks, most likely indicating the shorter trips on e-bikes with a top speed around 25 km/h and the longer trips made on the so-called speed-bikes with a top speed of 45 km/h.

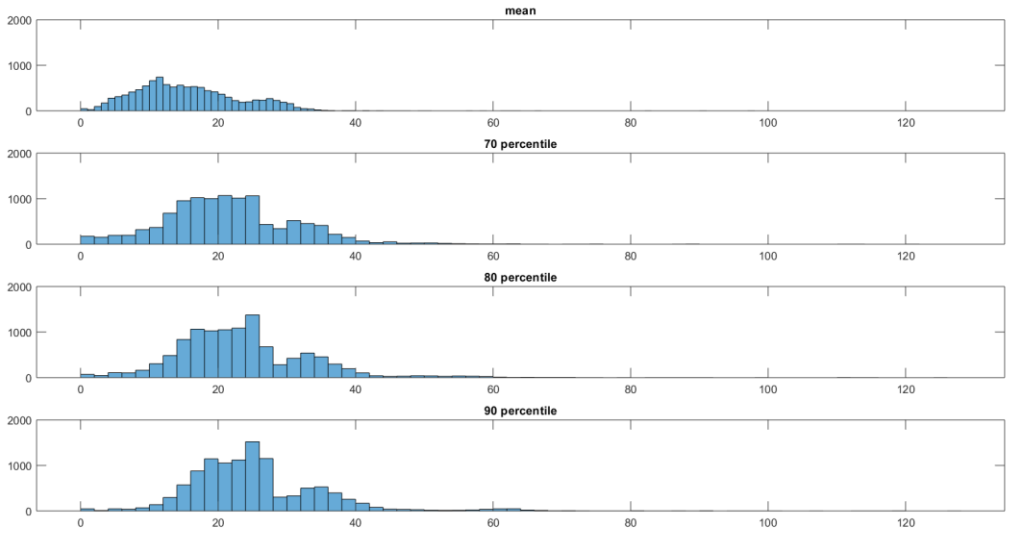


Figure 17. Distribution of average-(a), 70-,80- and 90-percentile (b-d) speed over all trips

To determine whether a trip was made by car, the '80-percentile' speed value was used. This speed indicator allows for less compensation of high speeds with low speeds than the 'average' speed indicator and also allows for outliers in GPS observations. When this speed is above 45 km/h, the trip is classified as a car trip (Figure 18).

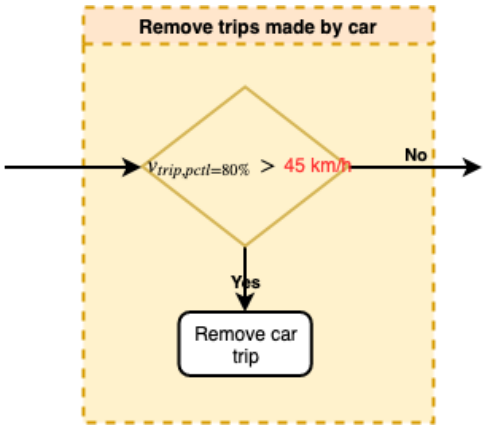


Figure 18. Procedure of removing trips made by car.

Not all car trips are expected to be removed with this algorithm. It is difficult to determine different trip-modes within the city centers based on speeds alone, since the speeds are difficult to distinguish from those of e-bikes. This portion of trips made by car within the city center is expected to be very small and regarded as insignificant.

4.4.3 GPS Error filtering

All observation points within the dataset were made using 4-12 satellites. The registered accuracies of the observations are mostly within 2 meters. There are many outliers in GPS observation points noticed related to random jumps or satellite/receiver issues, where for instance a single observation point within a trip is found on the north pole. These outliers are determined by calculating the acceleration needed between two consecutive GPS observations within a trip (Figure 19). If this calculated acceleration is found to be unrealistic for an e-bike, the observation point is removed.

The 'maximum' acceleration value found (approx. 1.2 m/s^2) can be justified by accelerations found by Dukulis et al. (2013). They found that the most powerful e-bike that they tested, had a maximum acceleration of approximately 1.3 m/s^2. A conservative safety factor of 2 is multiplied by this maximum acceleration, to cope with GPS inaccuracies, increased accelerations due to slopes, etc.

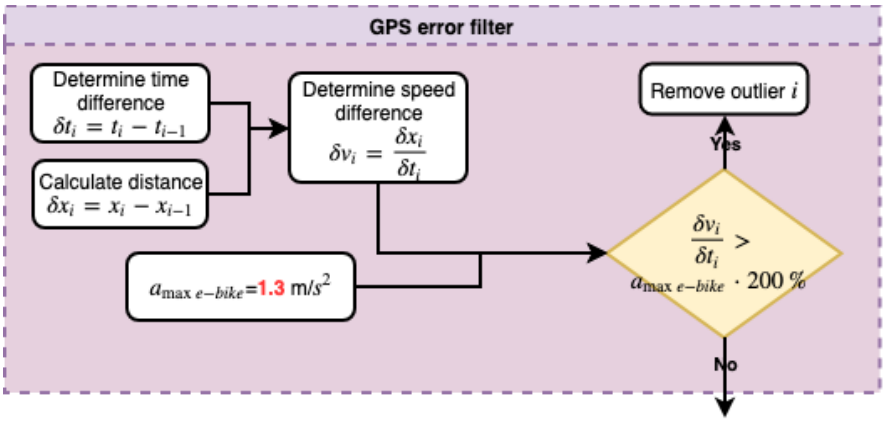


Figure 19. Procedure GPS error filter.

It's noteworthy that most of these outliers are found in the initial stages of the trips, which might indicate that the outliers are mostly related to starting of the GPS device (a.k.a. cold start issues).

4.4.4 Remove recreational trips

Cyclists on trips with a recreational purpose will yield other tradeoffs between route attributes than those trips with a utilitarian purpose. Since the interest of this research lies in utilitarian trips, these recreational trips are removed as much as possible from the dataset.

Recreational trips are often found to be very long trips and tours. These trips are thus filtered out as can be seen in Figure 20. The threshold value for the duration of a recreational trip is chosen $dur_{trip} > 90$min. Most trips have a duration shorter than 90 minutes and even the longer utilitarian trips are expected to be well below 90 minutes.

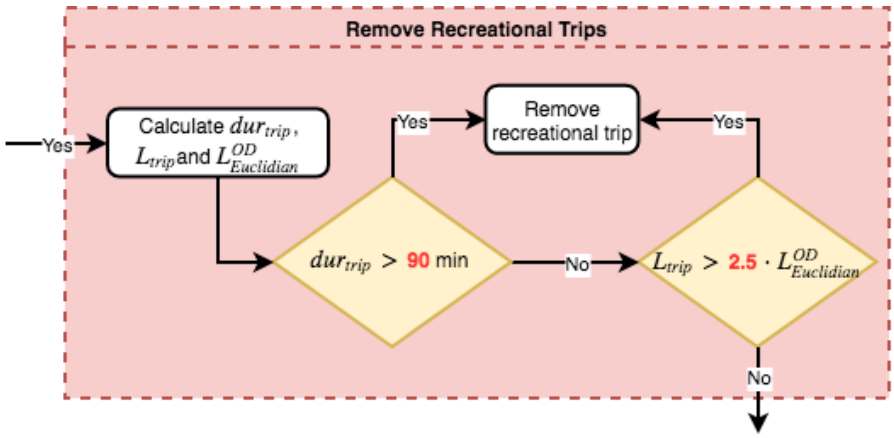


Figure 20. Procedure of removing recreational trips.

The tours are identified by comparing the length of the trip to the Euclidian distance between the origin and destination. When the difference is large enough, the trip is considered a tour. These trips can be classified as two clusters of trips (utilitarian and recreational). A scatter plot of the two distances against each other indicates that there are 2 clusters of trips (Figure 21). The threshold value for the distance difference is set to 2.5. This value is found by comparing the share of trips identified as tours for values between 1 and 10 (Figure 22). Using the elbow method, the value 2.5 is found to be the most suitable. After 2.5 there is relatively little change in the trip share.

Arguably, the clusters in Figure 21 are not clearly distinguishable based on a simple straight line. Perhaps a curved line which would tolerate more deviation of the trip distance for short trips would be provide a better division. This an interesting topic to look more detailed into for further research.

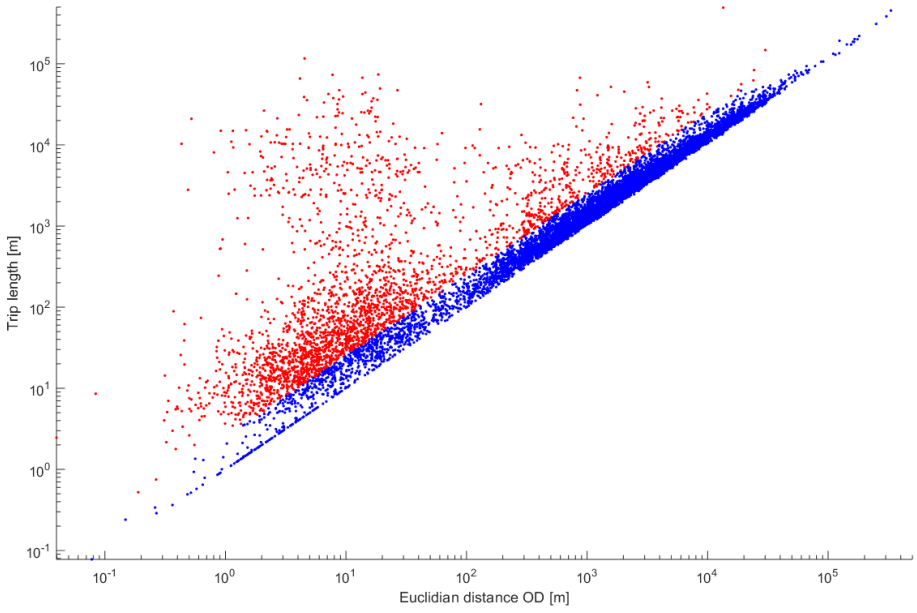


Figure 21. Scatterplot Euclidian distance OD vs Trip length with blue dots indicating trips below a threshold value of 2.5.

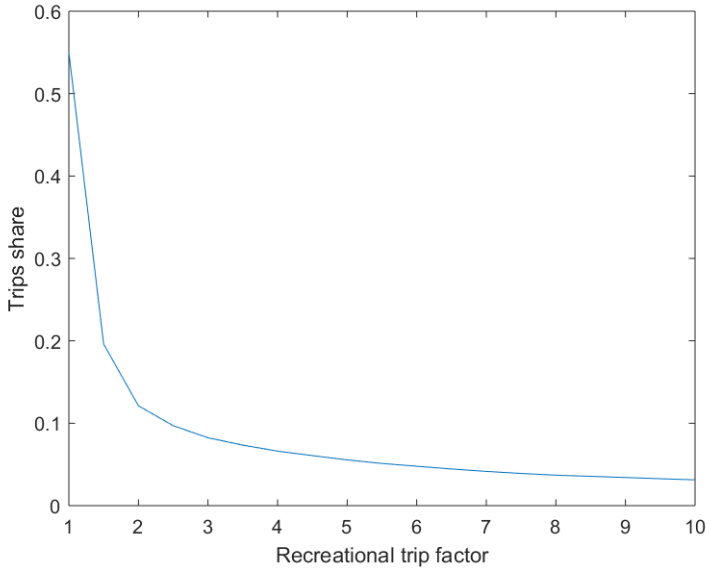


Figure 22. Trip share of recreational trips for different recreational trip/ tour factor.

4.4.5 Remove short trips

Short trips are of little importance for estimating route choice models, since there would be a small amount of route alternatives. This would have limited impact on the tradeoffs between the different route attributes. Trips with a distance lower than 300 m and with a duration shorter than 3 min are discarded.

4.4.6 Remove walking trips

Some trips even have low average speeds ($v_{trip,avg} < 6$km/h), most likely indicating walking trips while the bike is still on (Figure 17). These are also to be discarded for this research.

To determine whether a trip was made by walking, the '80-percentile' speed value was used similarly to the car-trip filter. This speed indicator allows for less compensation of high speeds with low speeds than the 'average' speed indicator and allows for outliers in GPS observations. When this speed is below 6km/h, the trip is classified as a walking trip.

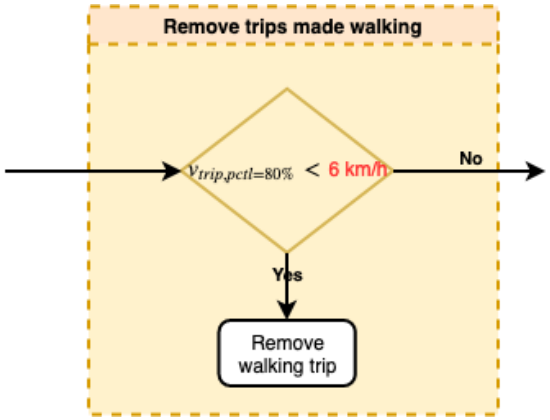


Figure 23. Procedure of removing trips made by care or walking.

4.4.7 Remove trips outside of considered area

As mentioned earlier, the initial considered area is a square of 30×30 km^2 around the city of Ghent. Trips entirely or partially outside of the considered area are filtered out (Figure 24). A small sensitivity analysis of

the size of the region showed that an expansion to the east of 5km of the Ghent region shows a significant improvement in the number of trips included.

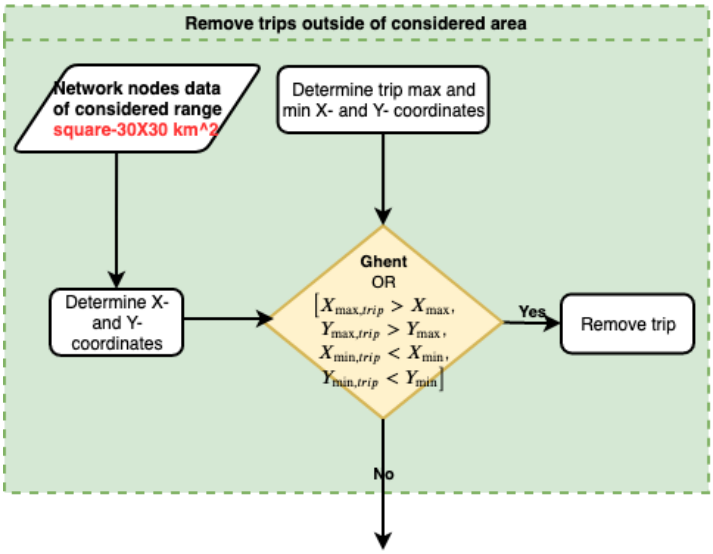


Figure 24. Procedure of removing trips outside of the considered area.

4.4.8 Simplify traces

The traces are reduced according to the 'Douglas-Peucker' polyline simplification algorithm. The reduction would improve the computation time of the map-matching procedure, while still being able to match the trips correctly. This algorithm removes unnecessary points on a path based on point to edge distances within a specified tolerance.

The algorithm is illustrated in Figure 25. It starts by creating a single edge between the origin and destination points. The point to edge distance for all intermediate points is calculated. The point with the largest edge distance and is larger than a specified tolerance, will be included in the simplified path. This process will continue until all points are within the specified tolerance edge distance. The tolerance used is the squared distance of 50 m^2, resulting in an absolute distance of 7m from the edge. This is roughly the width of a 2-lane road and thus any change of direction on an intersection or curve along a road, would be included. Points on a straight line would be removed.

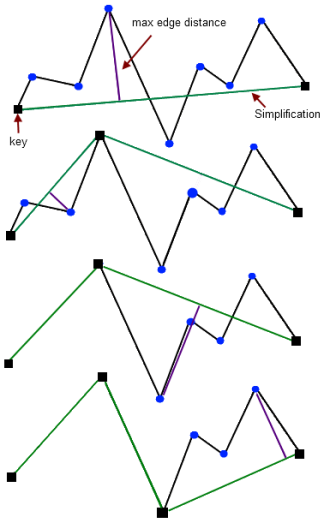


Figure 25. Douglas-Peucker polyline simplification algorithm illustrated.
Source: Psimple website. Retrieved from: http://psimpl.sourceforge.net/douglas-peucker.html

The impact of the simplification process of a small part of the trips in Ghent can be seen in the comparison given in Figure 26 , with each trip assigned to a different color. The traces are simplified substantially by reducing the amount of observation points by 81%. The traces are still recognizable after simplification.

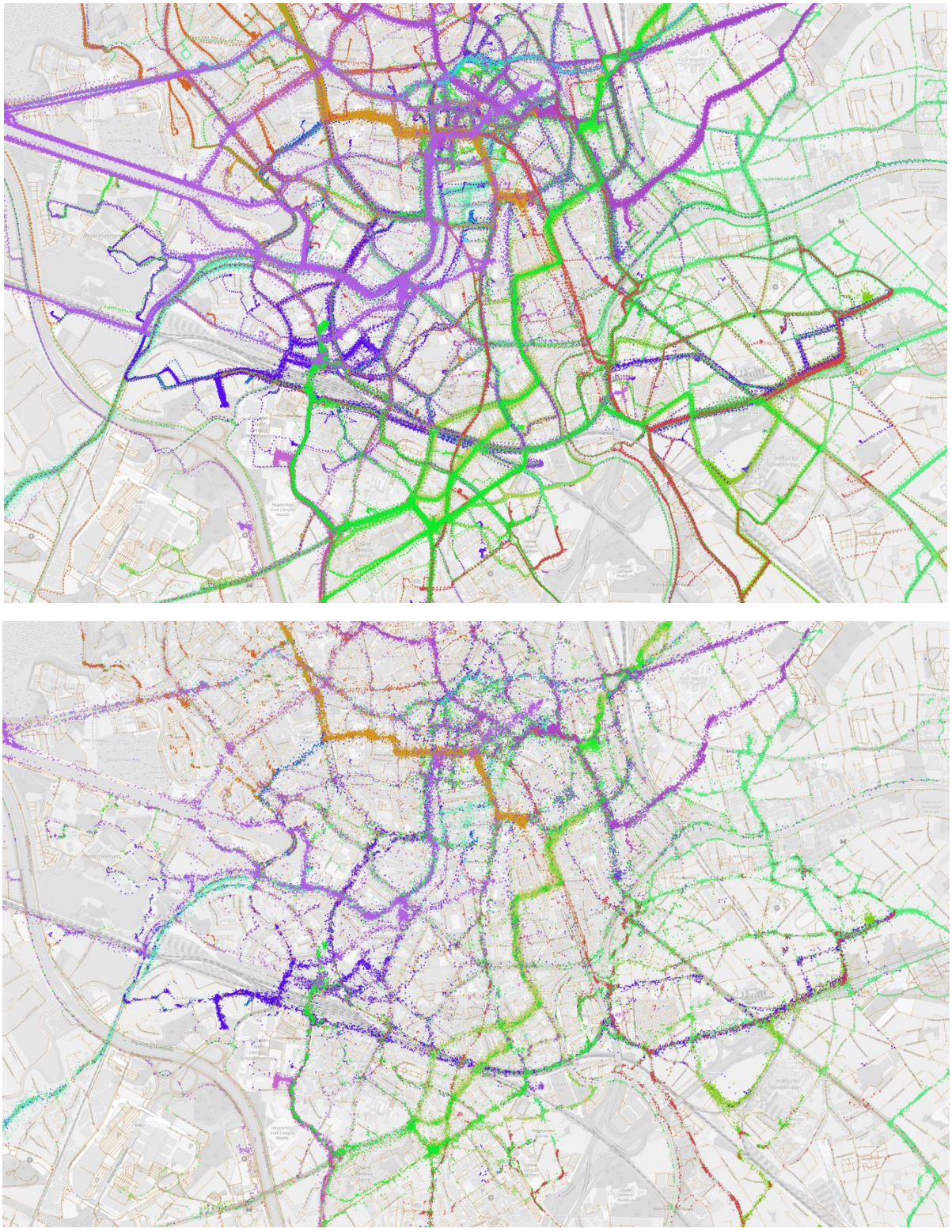


Figure 26. Sample GPS traces within Ghent before -and after simplification.

4.4.9 Results processing GPS data

An overview of the resulting number of observation points and trips after each step is given in Table 5. Around 10% of trips found to be of recreational purpose. Almost half of the identified trips consisted of less than 4 observation points. The filter with the biggest impact is as expected that of removing trips outside of the considered area. The trace simplifying algorithm removes most of the observation points.

Table 5. Overview of impact of each filtering procedure.

	# observation points	***# trips***	***Filtered out obs.***	***Filtered out obs. start***	***Filtered out trips***	***Filtered out trips start***
Start # trips (t_dwell=5 min)	3694746	28042				
Remove trips made with car	3586662	27713	2.9%	2.9%	1.2%	1.2%
Remove GPS outliers	3547729	27713	1.1%	4.0%	0.0%	1.2%
Remove recreational trips						
- trips>90 min	3191792	27477	10.0%	13.6%	0.9%	2.0%
- L trip> 2.5 L od Eucl	2863639	24835	10.3%	22.5%	9.6%	11.4%
Remove short trips						
-trips<16 s	2850774	12220	0.4%	22.8%	50.8%	56.4%
- L < 300m	2814790	11067	1.3%	23.8%	9.4%	60.5%
- T<3min	2807073	10746	0.3%	24.0%	2.9%	61.7%
Remove trips made walking	2786070	10529	0.7%	24.6%	2.0%	62.5%
Remove trips outside of considered area						
30km square	1811447	6572	35.0%	51.0%	38.4%	76.9%
add 5km Ghent East	1664360	6989	40.3%	55.0%	33.6%	75.1%
Simplify trace (point-edge tolerance 50 m^2)	153223	6898	90.8%	95.9%	0.0%	68.4%

An overview of the resulting GPS traces in the Ghent region is given in Figure 27, in which each participant is attributed a unique color.

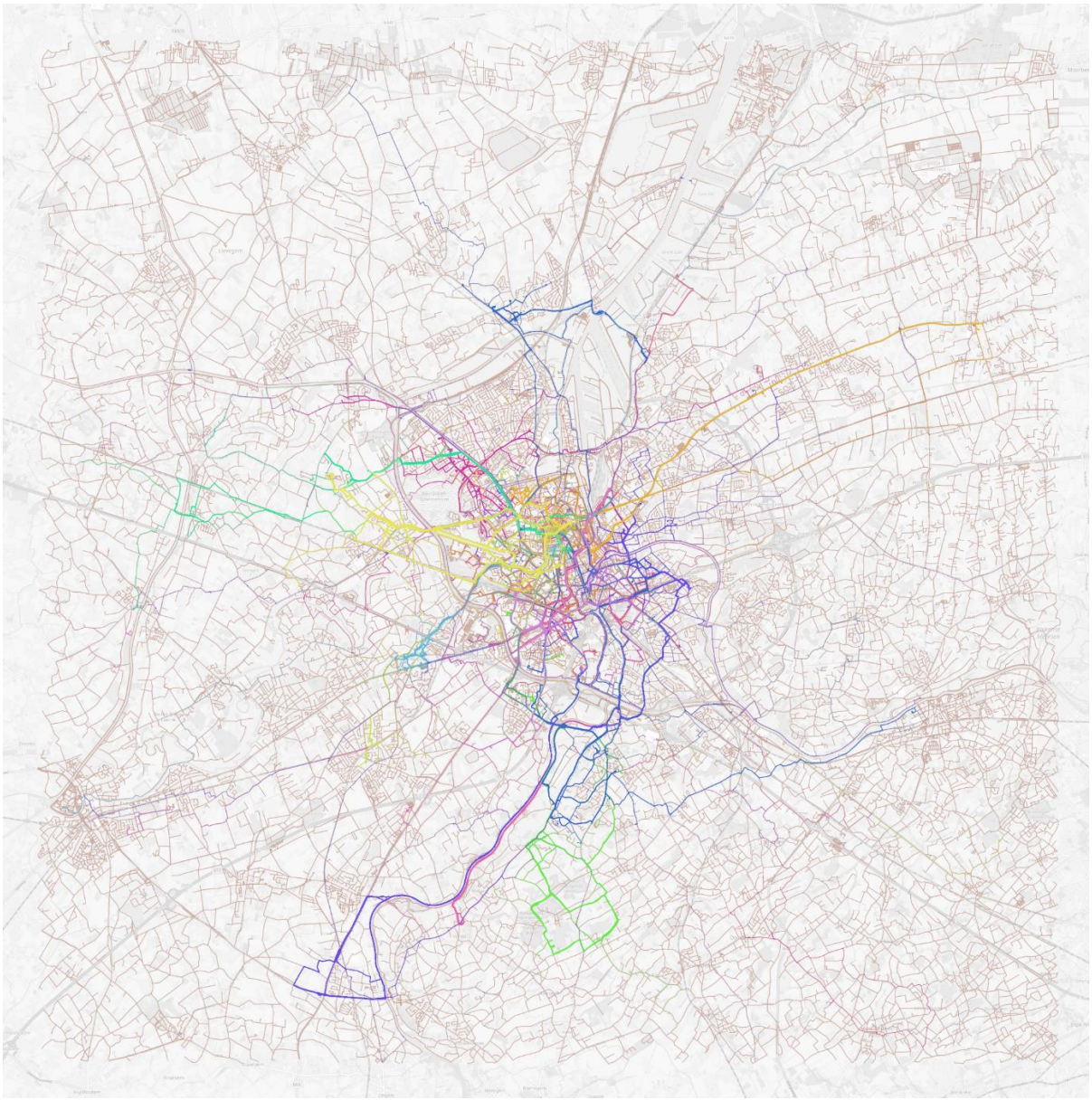


Figure 27. Overview of the resulting GPS data in the Ghent region.

4.5 MAP MATCHING

Now that both the network - and GPS data have been processed, the sequence of observed GPS positions are aligned to the road network on the digital map. In this case, the GPS data is matched to the OpenStreeetMap (OSM) network. This makes it possible to identify the processed network characteristics of a trip and compare the routes.

The map-matching algorithm used, is based on that of Lou et al. (2009) called 'ST-Matching for low sampling-rate GPS trajectories'. The method includes both a spatial and temporal analysis to find the best matching path sequence out of a candidate graph. The spatial analysis considers both the distance between an observation node (GPS observation) and candidate road links, and the shortest path between consecutive candidate nodes. The temporal analysis involves comparing the actual average speed between observation points, to speed limits of the candidate road links. For map-matching of this bicycle GPS data, the temporal analysis is discarded, since road speed limits generally don't apply to the cyclists.

An overview of the different processes of the algorithm used, is given in Figure 28.

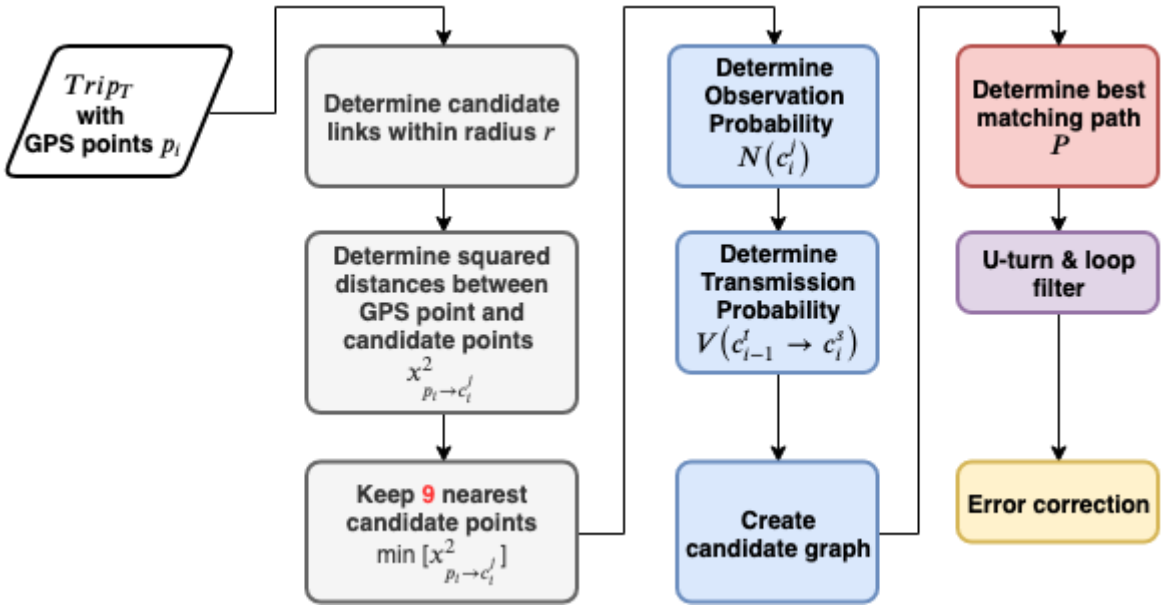


Figure 28. Overview of Map Matching method.

The algorithm first gathers the candidate road segments from the network based on a radius r around each GPS observation p_i . Then, the nearest points on these candidate road segments to the GPS observation p_i are considered candidate nodes c_i^j. Based on the squared distance $\mathrm{x}^2_{\mathrm{p_i -> c_i^j}}$ to each candidate node, the 9 nearest candidate points c_i^j are considered for the spatial analysis. The spatial analysis involves calculating the observation probability and the transmission probability.

The observation probability $N(c_i^j)$ gives the likelihood that a GPS observation p_i matches a candidate point c_i^j based on the distance between the points $\mathrm{x}^2_{\mathrm{p_i -> c_i^j}}$. The error in the GPS observation is assumed to be normally distributed $N(0, \sigma^2)$, with a zero mean and a standard deviation of σ =20 meters. The observation probability is defined as follows:

$$N(c_i^j) = \frac{1}{\sqrt{2\pi}\sigma} . e^{\frac{-x^2_{p_i -> c_i^j}}{2\sigma^2}} \tag{19}$$

The transmission probability $\mathrm{V}(\mathrm{c_{i-1}^t -> c_i^s})$ gives the likelihood that the observation path between two consecutive observation points $p_{i-1} \to p_i$ follows the shortest path between their respective candidate points $c_{i-1}^t \to c_i^s$. The transmission probability is defined as follows:

$$\mathrm{V}(c_{\mathrm{i-1}}^{\mathrm{t}} -> c_{\mathrm{i}}^{\mathrm{s}}) = \frac{d_{p_i -> p_{i+1}}}{w_{c_{i-1}^t -> c_i^s}} \tag{20}$$

Where $d_{p_i -> p_{i+1}}$ is the Euclidian distance between the observation points p_{i-1} and p_i, and $w_{c_{i-1}^t -> c_i^s}$ the length of the shortest path from c_{i-1}^t to c_i^s.

The spatial analysis function is thus defined as the product of the two different probabilities, as follows

$$F_s(c_{\mathrm{i-1}}^{\mathrm{t}} -> c_{\mathrm{i}}^{\mathrm{s}}) = N(c_i^j) \times \mathrm{V}(c_{\mathrm{i-1}}^{\mathrm{t}} -> c_{\mathrm{i}}^{\mathrm{s}}) \ for\ 2 \le i \le n \tag{21}$$

given that a path contains n observations.

After the spatial analyses a candidate graph $G'_T(V'_T, E'_T)$ for each trip T can be created, with V'_T being the candidate points for each GPS observation and E'_T the set of edges representing the shortest paths between two neighboring candidate points (see Figure 29). Each node has an observation probability $N(c_i^j)$ and each edge a transmission probability $\mathrm{V}(\mathrm{c}_{\mathrm{i}-1}^{\mathrm{t}} -> \mathrm{c}_{\mathrm{i}}^{\mathrm{s}})$.

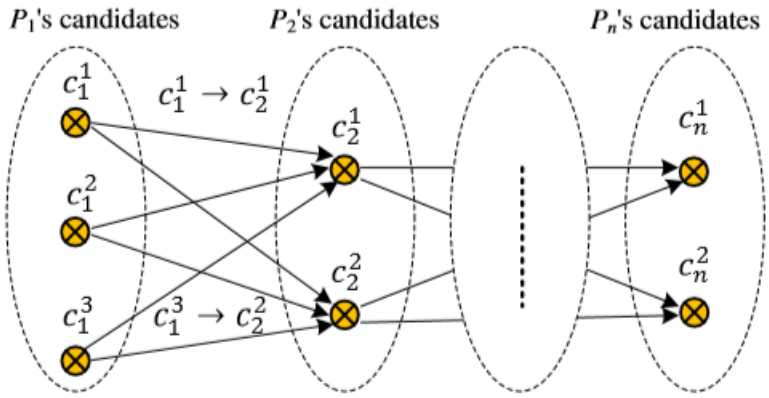


Figure 29. Candidate Graph $G'_T(V'_T, E'_T)$.
Source: Lou et al. (2009).

A path in the candidate graph is a candidate path sequence, which is denoted as $P_c = c_1^{S_1} \to c_2^{S_2} \to \cdots \to c_n^{S_n}$. Each candidate path sequence P_c gets an overall score based on the spatial analysis function as follows $F(P_c) = \sum_{i=2}^{n} F_s(c_{i-1}^{s_{i-1}} \to c_i^{s_i})$. The candidate path with the highest overall score is considered the best matching path P for the trip T and is formally defined as

$$P = \arg\max_{P_c} F(P_c), \qquad \forall P_c \in G'_T(V'_T, E'_T) \tag{22}$$

Due to errors in the GPS data and network data, trips might be map-matched incorrectly. After visual checks of the map-matched trips, there seem to be U-turns and loops occurring. There are also deviations at the start or end of trips, due to missing driveways in the network. U-turns and loops in the map-matched trips are removed, by cutting out that part of a path in-between each node that is visited more than once.

The algorithm also includes an 'observation probability threshold factor' f_{opt}, which allows to skip trying to match observations with observation probabilities lower than this threshold. Different values for the assumed standard deviation of the GPS error σ and the observation probability threshold factor f_{opt} have been used to check the performance of the map matching method. The performance of the procedure is checked visually and by applying a performance indicator $PI_{MM,T}$ based on the trip length.

The average absolute deviation of the length of the GPS trajectory $L_{GPS,T}$ from the length of the map-matched trajectory $L_{MM,T}$ for a sample of trips with sample size n, is considered as a performance indicator $PI_{MM,T}$ for the map matching procedure. It is defined as follows:

$$PI_{MM,T} = \frac{\sum_{T=1}^{n} \left| \frac{L_{GPS,T}}{L_{MM,T}} - 1 \right|}{n} \tag{23}$$

A limited amount of values for both variables have been tested for the same sample of trips, using both the visual check and $PI_{MM,T}$. Tested values are $\sigma = [10, 20, 30]$m and $f_{opt} = [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$. The best

combination of values is found to be σ=20m and f_{opt}=10^{-2}. Noteworthy is that a deviation of the same size has more impact on $PI_{MM,T}$ for a short trip than a long trip.

Some errors were still found in the map-matched data in terms of skipped links between two nodes on a path. This seemed to occur for 1 out of the 200 observations at specific locations. The map-matching algorithm failed to assign a node for these observations for some reason. Perhaps observation points in between the nodes were too far off any network nodes due to a GPS error (caused by a location specific disconnection from satellites).

A workaround for this problem is added by determining the shortest path between these disconnected consecutive nodes on a map-matched path. The nodes found on the shortest path are added to the map-matched path.

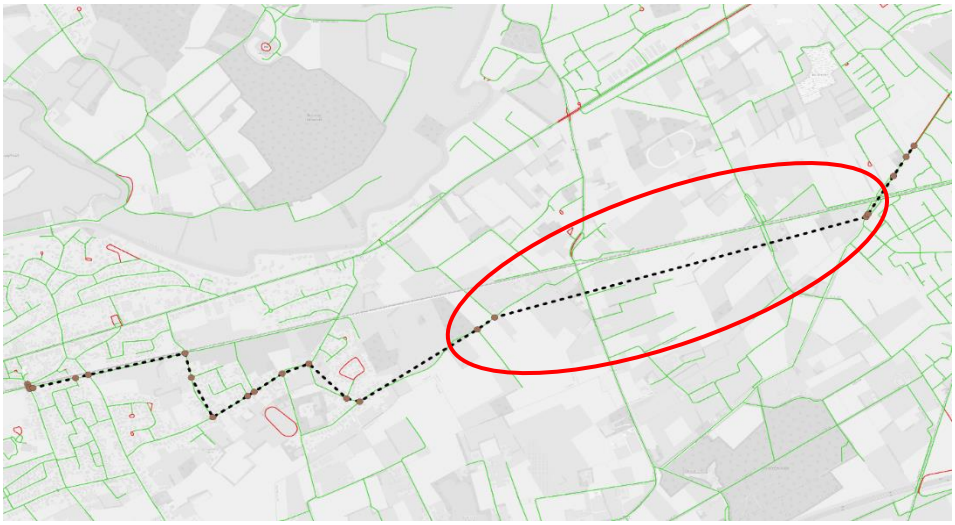


Figure 30. Example of skipped links error in the map-matched data.

4.6 DESCRIPTIVE STATISTICS OF THE FILTERED DATA

The filtered GPS data contains many trips per person collected over a period of up to 95 weeks. Naturally many of the routes made by the same person are repeated. Since studying these panel effects are not within the scope of this project, the panel data is reduced by removing non-unique trips made by the same person. After the GPS traces were map-matched, they could be compared to each other in terms of the sequence of nodes on the path. Trips that are made by the same person and are completely identical in terms of their path, are removed such that the trip only occurs only once. As a result, the number of trips in the Ghent region was reduced from 6989 to 6897 unique trips.

An overview of the density distributions of the resulting trip characteristics in terms of average trip distance, -duration and -speed is given in Figure 31.

Three main types of trips can be distinguished based on the three bumps seen in the trip distance and -duration distributions. These three types of trips can be called short (shorter than 4km and less than 16 minutes), medium (between 4-12 km and 16-35 minutes) and long trips (between 12-28 km and 35-60 min). Most of the trips are surprisingly of the short type. With E-bikes at their disposal, it was expected that the participants would make more of the medium and long trips.

The average speed distribution is more widespread however, two clear bumps can be noticed around 12 km/h and 25 km/h. Perhaps these two speeds are linked to trips made within the city center and outside of the city center respectively.

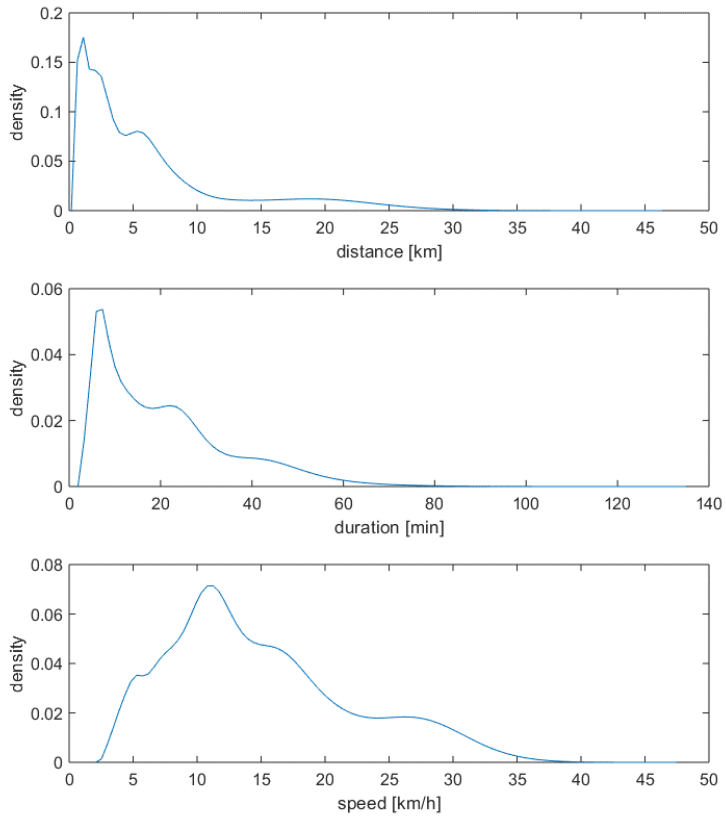


Figure 31. Density distributions of the trip characteristics.

Some more trip statistics for each participant are displayed in Figure 32. The *number of weeks over which data was collected* in weeks is determined based on the difference between the timestamps of the last and first trip recorded. The *average number of weekly trips* is determined by dividing the total number of trips by the number of weeks the person participated in the research. The *average trip distance* is taken over all the participant's trips. And finally, the *average speed over all trips* is also included.

The duration of participation as well as the number of weekly trips varies a lot between the participants. Most participants made around 10 trips per week, which corresponds with home-work and work-home trips on 5 workdays. Some participants have a low average number of weekly trips. It seems like these participants barely made any trips on their E-bike. Perhaps they used their E-bike more for recreational trips, which were filtered out of the dataset. The average speeds are mostly around 10-15 km/h range, which could perhaps be related to most participants making trips taking place mostly within the city center.

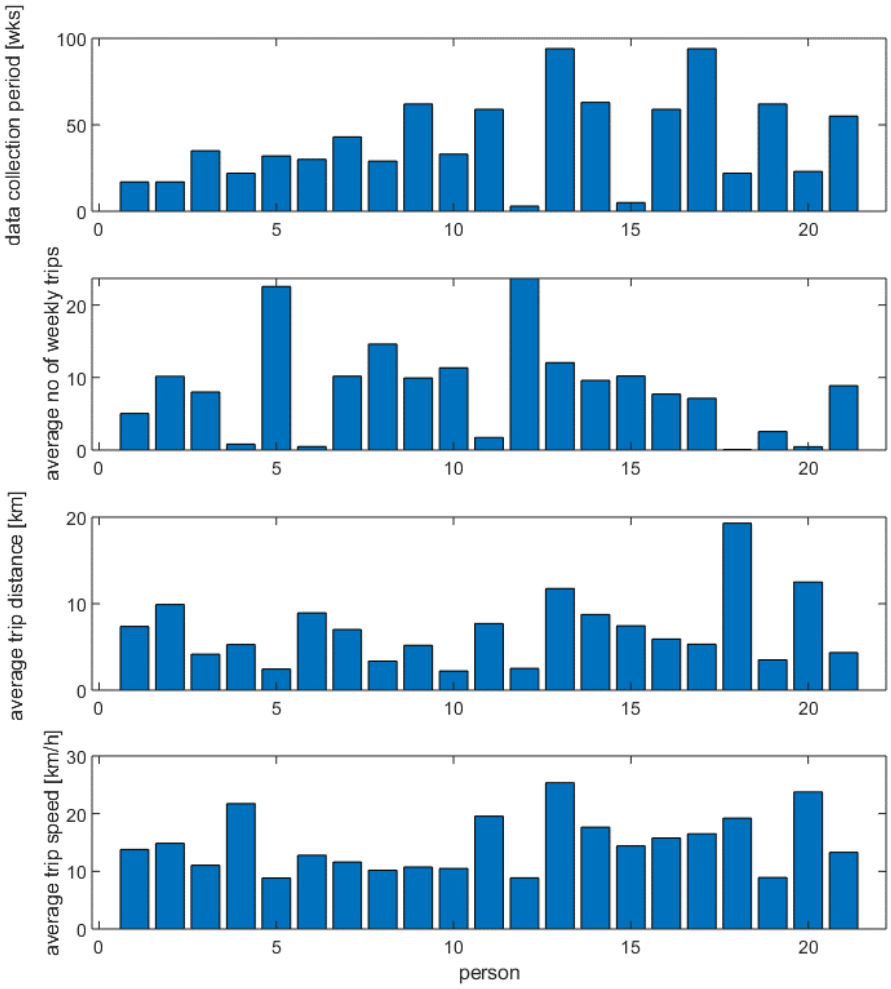


Figure 32. Trip statistics for each participant.

An overview of the considered network data related to the observed trips is given in Table 6. The descriptions of the attributes are discussed in section 3.4.1. Roughly 13% of the trips take place on dedicated bikeways on average. However, cycling infrastructure information in the OSM network data is lacking in terms of cycling lanes and shared cycling roads. Some bikeways also seem to be unrightfully classified as track/ path which now fall under the 'other roads' category. On average, most of a trip takes place on small roads. One-way restrictions are considered as cycling in the opposite direction of a one-way street except if that street is classified as a residential street, as cyclists are allowed to cycle in both directions. The distances cycled in the wrong way of one-way restricted streets is very small compared to the route distances. On average one out of two trips encounters a roundabout and every trip encounters a traffic light on average as well. Trips include almost the same amount of left turns as right turns on average. There are almost 47 intersections crossed per observed trip on average, including all types of controlled and uncontrolled intersections, with or without right of way and with roads of the different categories. Information on stops signs or right of way is unfortunately not included in the OSM dataset. Links with significant up-slopes are barely used by the E-bikers, which is not surprising since the topology of the region is more or less flat.

Table 6. Descriptive statistics of each considered attribute for the observations.

	Unit	Mean	Median	St. dev.
Length	*km*	6.97	5.01	6.31
PS	-	0.52	0.54	0.25
Wrong way	*km*	0.03	0.00	0.06
Cycleway	*km*	0.91	0.31	1.47
Large road	*km*	2.03	1.30	2.22
Small road	*km*	2.62	2.08	2.11
Other road	*km*	1.40	0.22	2.95
Green	*%. km*	2.92	1.36	3.39
Water	*%. km*	0.59	0.20	0.96
Building density	*%. km*	1.10	0.99	0.72
Scenic (G+W based)*	*km*	5.52	2.98	5.79
Roundabouts	*#*	0.56	0.00	0.96
Traffic Lights	*#*	1.25	1.00	1.65
Left turns	*#*	4.77	4.00	3.82
Right turns	*#*	4.93	4.00	3.71
Intersection cross (no turn)	*#*	46.53	41.00	32.38
Up-slope 2-4%	*km*	0.18	0.11	0.20
Up-slope 4-6%	*km*	0.02	0.00	0.04
Up-slope >6%	*km*	0.02	0.00	0.04

** determined in section 6.1*

The geographical spread of some of the attributes are visualized in Appendices B, C, D and E. The traffic lights are spread across the larger roads (primary and secondary roads) with the majority located within the city center (see Appendix B). The roundabouts are less clustered in the city center than the traffic lights and are mostly located in the sub-urban region outside the city center (see Appendix B).

The considered region is 'flat' in general, with hills in the southeast part as well as some significant elevation in the city center (see Appendix C). The calculated slopes between the start- and end nodes of each link are visualized in Appendix D. The links with significant slopes are located within the city center and in the southeast hill region. It is noteworthy that these links in the city center are short, while those in the southeast are longer.

The land use map is given in Appendix E. The urban area in terms of the city center is densely built, while the rural area is filled with greenery. Most of the greenery in the rural area consists of grass- and agricultural fields and forests.

5 RECURSIVE LOGIT VS PATH SIZE LOGIT

The findings of the RL model, as well as a general comparison of the PSL model and RL model are discussed in this chapter.

Disclaimer!

Initially the focus of this thesis project was on analyzing cyclist's route choice behavior with the state-of-the-art Recursive Logit (RL) model and attempting to compare the RL model with the conventional PSL model. Looking at the formulation of the RL model (discussed in section 2.1.4), it is clearly much more complex than the PSL model (discussed in section 2.1.3). Several attempts were made to get the RL model to converge consistently for different initial parameter values without success. Because of the complexity and lack of experience with this model, the exact source of the problem could not be identified within this project. The decision was eventually made to not go further with this model and move on to the PSL model, which is proven to work well. During the implementation of the PSL model, some errors in the map-matched paths were discovered. However, the RL model was not tested with the corrected data.

The procedure and outcomes of the tests should however still be of added value for further research, as well as a discussion on potential aspects to further investigate.

The RL algorithm required some additional data processing as the RL model is based on turn-based data (discussed in section 5.1). Several attempts were made to get the RL model to converge without success. The results of the attempts are given in section 5.2 and further discussed in section 5.3. A quantitative comparison between the RL model and the PSL model could not be made. Nonetheless, a qualitative comparison is made based on the experience gained of implementing both models in section 5.4.

5.1 ADDITIONAL DATA PROCESSING

Additional data processing is required for this model, since the model makes use of turn-based (link-to-link) information instead of node-to-node based information. A simple illustration of the turn-based data format is given in Figure 33. The map-matched trip data node-to-node data had to be converted, additional dummy links (without any attribute value) had to be included at the origins and destinations (indicated by links 1 and 6 in Figure 33) and link-based attribute data had to be converted into turn-based data.

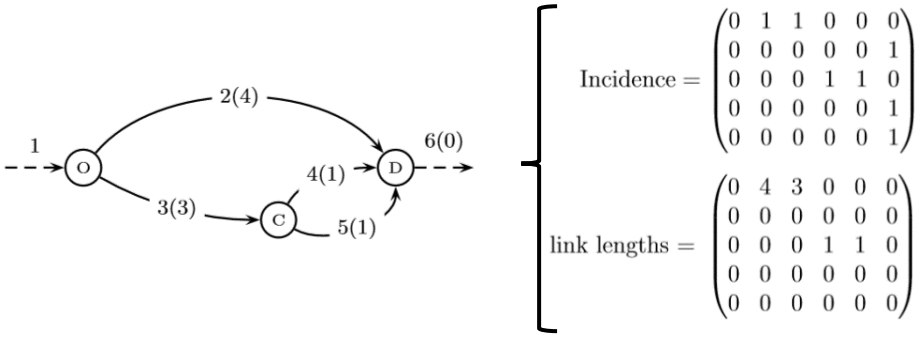


Figure 33. Simple three paths example for data input format of the RL model. Source: Fosgerau et al. (2013).

5.2 GETTING THE MODEL TO CONVERGE

Attempts were made to get the RL model to converge consistently, without success.

In order to keep the test simple, a subset of the data was used to test the model and the link distance was the only attribute included in the model. A test network within approximately a region of 7x8 km^2 in the center

of Ghent consisting of 11400 links and 8167 nodes, was defined for this test only. Observations that completely or partially take place within this predefined region were cut-off at the borders of the region and used as trip data input for the model. By doing this, 2733 trips were identified.

It seems that the model is very sensitive to the initial parameter values. Several tests were conducted on the test data including multiple network sizes, number of observations, non-simplified vs network simplified (as discussed in section 4.3) and different initial parameter values β_{L_0}.

Given in Table 7 are the results for the test dataset using different initial parameter values for the link length (β_{L_0}) based on the 2733 observations. The other columns represent: the number of iterations, final Log-likelihood (final LL), final parameter estimate for the link length ($\beta_{L_{\text{final}}}$), standard deviation of the final parameter estimate and the stopping argument. For some observations, the link probabilities could not be determined initially, and the model would stop optimizing. Therefore, the model was adapted such that it could skip these observations for that iteration. The total amount of observations skipped over all iterations is given in the last column of Table 7.

The model was not able to optimize with positive initial parameter values $\beta_{L_0} \geq 0$ and aborted after one iteration. The model seems to converge at the same point for initial parameter values of -0.15 and -0.1, although the stopping argument indicates the optimization was not successful. Note that for these two initial parameter values, the model hardly skipped any observation. For the other tested initial parameter values, the model was not able to converge.

Table 7. Results test RL with link length as only attribute.

β_{L_0}	# iterations	Final LL	$\beta_{L_{final}}$	St. dev.*	Stopping argument	# observations skipped
-0.40	71	39.90	-0.39	NaN	TOO SMALL STEP	110107
-0.30	64	45.05	-0.30	NaN	TOO SMALL STEP	78380
-0.25	58	47.01	-0.23	NaN	TOO SMALL STEP	51665
-0.20	65	45.78	-0.16	NaN	TOO SMALL STEP	32661
-0.15	52	38.17	-0.07	1.37^{E}-05	TOO SMALL STEP	7
-0.10	50	38.17	-0.07	1.35^{E}-05	TOO SMALL STEP	0
-0.05	70	12.35	-1.03	3.16^{E}-07	TOO SMALL STEP	168779
0.00	71	12.68	-0.99	8.32^{E}-08	TOO SMALL STEP	169165

** could not be determined for some initial parameter values*

The final parameter estimates $\beta_{L_{final}}$ for the first four tested initial parameter values did not differ much from their initial values $\beta_{L_0} = [-0.40, -0.30, -0.25, -0.20]$. Meanwhile, the final parameter estimates $\beta_{L_{final}}$ for the small initial parameter values $\beta_{L_0} = [-0.05, -0]$ lie around $\beta_{L_{final}} \approx -1$. For the latter, the algorithm might be taking a too big initial step (an initial step size of 1), making it search for a local optimum somewhere outside the global optimum. The initial step size for the optimizer was varied between 1 and 0.1, which however led to similar results as those presented in Table 7.

5.3 DISCUSSION: RL MODEL

As mentioned in the disclaimer, some errors were found in the map-matched observations after the RL model was set aside for this project. These errors involved some links being skipped by the map-matching algorithm. This happened for 1 out of the 200 observations at certain locations. This was afterwards fixed during the implementation of the PSL model and included in the map-matching algorithm. The RL model was not tested using the corrected map-matched observations, thus the results might vary from those shown in the previous

section. The number of skipped observations is expected to decrease and probably lead to more consistent results for the different initial parameter values. Nonetheless, the discussion is expanded on the results obtained using the data containing the errors.

The value function (equation (10)) does not appear to have a solution for all the parameter values β. Fosgerau et al. (2013) admit that the value function can only be solved if matrix $(\boldsymbol{I} - \boldsymbol{M})$ is invertible and therefore making the search for a suitable initial parameter value difficult. With $\boldsymbol{I}$ being the identity matrix and $\boldsymbol{M}$ the incidence matrix with instantaneous utilities $\boldsymbol{M} = \delta(a|k)e^{\frac{1}{\mu}v_n(a|k)}$. They state that the invertibility of $(\boldsymbol{I} - \boldsymbol{M})$ depends on the balance between the number of alternative paths between OD pairs and the size of the instantaneous utilities $\frac{1}{\mu}v(a|k)$.

If there are many alternative paths, the expected utility to the destination increases. If the instantaneous utility is close to zero, the expected utility might become positive due to the random terms being positive and larger than the small negative deterministic utility. This positive expected utility could result in routes with infinite number of loops on small links. Thus, these two mentioned reasons may lead to $(\boldsymbol{I} - \boldsymbol{M})$ being ill-conditioned or even singular.

The OSM network data used in this project was very detailed and included many short links (<10m) that might have made it impossible to solve the value function. These short links were removed by simplifying the network but did not improve the result of the RL model (the results shown in Table 7 are based on the simplified network data). Perhaps the threshold for short links used (<10m) is too small and thus positive expected utilities still occur. Larger threshold values could be tested in the simplification process.

Another point of discussion for the RL model as proposed by Fosgerau et al. (2013), is that the direct optimization methods as they proposed only work from small to medium size networks. They classify the network that they used with 7459 links as medium sized. The test network used in this project is already much bigger (11400 links) and might also be a possible reason why the model did not converge. Fosgerau et al. (2013) refer to iterative solution methods for large networks. These have not been investigated in this research.

5.4 QUALITATIVE COMPARISON RL - AND PSL MODELS

The major advantage that the RL model has over the PSL model is that it does not require generation of route alternatives. It considers all links within the network and thus prevents any bias in the choice set composition.

However, the RL model is more difficult to use and less user friendly compared to the PSL model. The fact that the errors in the map-matched paths were found only after applying the simple PSL model, underlines its complexity- and lower level of user friendliness of the RL model. The model acts as a 'black box' which requires input that is already validated and free from errors. A big downside of the RL model compared to the PSL is that the model is very sensitive to the initial parameter values. The RL model cannot be optimized for certain initial parameter values and in some cases seems to find a local optimum for the parameter estimates for different values of initial parameter values. Due to the sensitivity of the RL model, calibration of the initial parameter estimates is needed. The many trials needed for calibration of each initial parameter values add up for each new attribute added to the model. Also, compared to the PSL model, additional data processing is required to convert node-to-node data to link-to-link data and dummy links need to be added at origins and destinations. This link-to-link formulation is arguably less intuitive to use, since the traffic network is represented as node-to-node data by default.

The tested RL model (without LS overlap attribute) runs relatively fast when compared with the PSL model and BFS-LE route generation combined. The RL model took around 3-5 minutes for the test network within

7x8 km^2 and 2733 observed routes. Only route generation with the BFS-LE would take more than an hour of computation for the test network. If the link size overlap factor LS would be included in the RL model, the computation time is expected to increase drastically as is proven in previous studies on the RL model (Fosgerau et al., 2013 and Zimmermann et al., 2017).

The PSL model can be extended such that the model takes correlation of choices into account. This becomes the mixed logit model. It is unknown if the RL model can be extended such that it can take these correlations of choices (i.e. panel effects) into account.

6 ROUTE GENERATION

The generated route set, which forms the basis of the PSL route choice model, is discussed in this chapter. The following question is answered in this chapter:

> To what extent does a multi-attribute cost function in the route generation improve the choice set compared to the distance-only method?

The performance of different route sets generated with the Breadth First Search on Link Elimination (BFS-LE) method using different multi-attribute cost functions is compared on a small dataset as mentioned in section 3.3.1. The attributes considered are the road categories- and land use attributes, in line with Halldórsdóttir et al. (2014). First, a dummy variable is created from the calculated land use fractions of the area of the link buffers (section 6.1). Different cost functions were tested with different attributes and different weight permutations for each type of link (section 6.2). Descriptive statistics of the resulting route set are presented as well (section 6.3).

6.1 LAND USE- SCENIC THRESHOLD

As mentioned in section 3.3.1, in terms of the land use attribute, the links are classified as either scenic or non-scenic. Arguably, whether a road is classified as scenic is rather subjective, but nonetheless an attempt is made to classify the links.

This classification is done based on threshold values for both green- as well as water area within a 30m buffer on both sides of the links. If a link has a higher percentage of either green- or water area within its buffer than the respective threshold values for green- and water area, then this link is classified to be a scenic link. The threshold value for green area is set to 20% and that of water area to 2%. These values are determined based on visual validation.

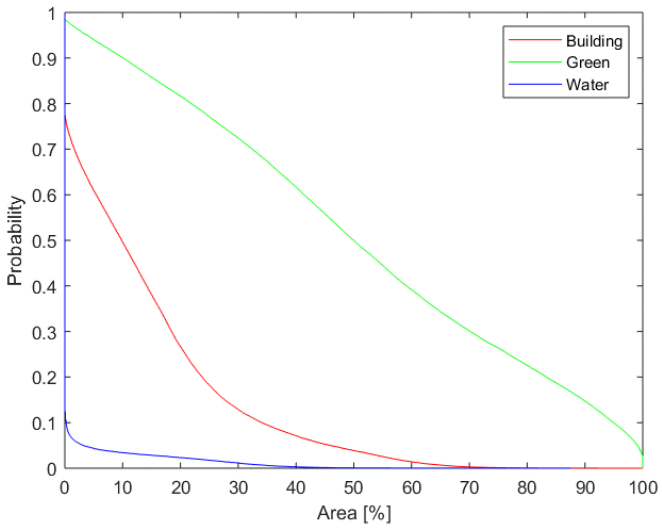


Figure 34. Cumulative distributions of land use area on links.

Figure 35 gives an impression of buffers of links for different green are percentages. In general, for links with green area values below 20%, green scenery is not seen from urban roads. The green areas within the buffer lie behind buildings that are in between the road and greenery (Figure 35- top). This is not considered to be scenic, whereas for percentages above 20%, greenery is in direct contact with the road (Figure 35- bottom).

Figure 35. Highlighted link buffers in an urban region based on green area percentages between 15-20 % (top) and 20+% (bottom).

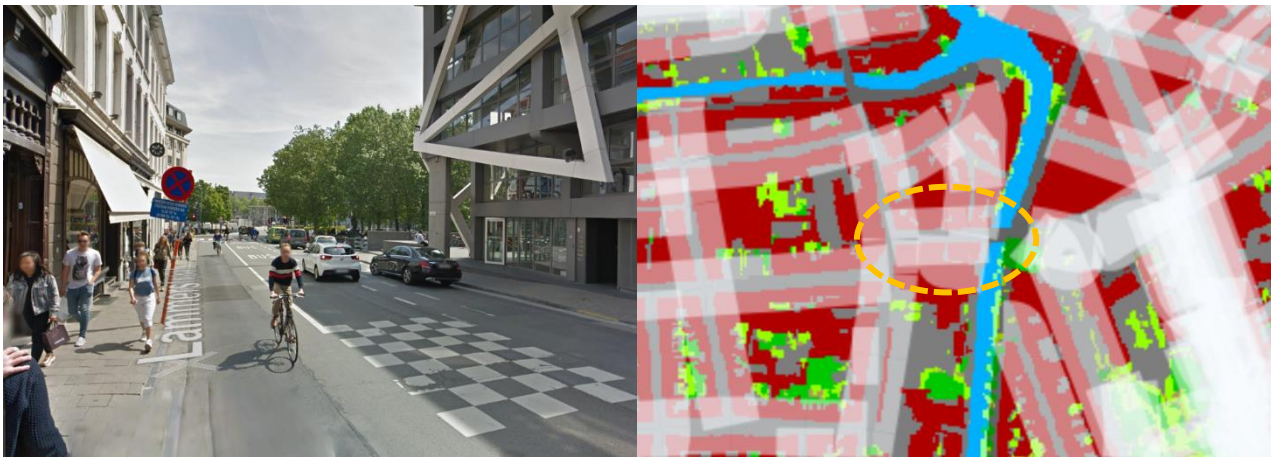


Figure 36. Impression of an arguably non-scenic link with water area percentage of <2% in its buffer. Source: Google Maps.

A threshold value for water area is set to avoid that links, that have a very small overlap with a canal or lake, are considered scenic. The cumulative distributions of the percentages of area of buildings, green and water within the direct environment of the links are displayed in Figure 34. Using the elbow method on the cumulative distribution of links with different percentages of water area within their buffer area, it is decided that 2% is the optimal value of dividing links in two clusters based on their percentage of water area. Based

on random visual checks using street views in Google Maps, many of these have water near one of their ends. Arguably, these links should not be classified as scenic. See Figure 36 for an impression of one of such links with water area <2% within its buffer.

6.2 CALIBRATION OF COST FUNCTIONS

A small random set of 50 unique observations is taken within a large part of the network for calibrating the weight parameters of the cost functions. The descriptive statistics are given in Table 8.

Table 8. Descriptive statistics of the 50 unique random observations.

	Unit	Mean	Median	St. dev.
Length	*km*	4.29	3.69	2.98
PS	-	0.52	0.54	0.25
Wrong way	*km*	0.59	0.49	0.55
Cycleway	*km*	0.58	0.18	0.75
Large road	*km*	1.41	0.76	1.53
Small road	*km*	2.06	1.58	1.48
Other road	*km*	0.24	0.11	0.32
Green	*%. km*	1.30	1.05	1.13
Water	*%. km*	0.25	0.08	0.32
Building density	*%. km*	0.99	0.87	0.65
Scenic (G+W based)*	*km*	2.88	2.17	2.43
Roundabouts	*#*	0.60	0.00	1.20
Traffic Lights	*#*	0.94	0.00	1.35
Left turns	*#*	4.44	3.50	3.49
Right turns	*#*	4.26	3.50	3.16
Intersection cross (no turn)	*#*	38.08	35.00	24.86
Up-slope 2-4%	*km*	0.17	0.11	0.18
Up-slope 4-6%	*km*	0.02	0.00	0.04
Up-slope >6%	*km*	0.01	0.00	0.02

** determined in section 6.1*

The time to run the route generation code for one cost function takes on average around 16 minutes. The predefined discrete values for the weights β_{scenic}, β_{Large} and β_{Small} are chosen to be [0.5, 0.7, 1, 1.5, 2]. This set of values allows for a relative preference or -aversion towards a certain type of link and is in the same order of magnitude as those found significant by Halldórsdóttir et al. (2014). The weight value for the large roads $\beta_{Large} = 1$ indicates that there is no difference in the perceived cost for large roads compared to that of roads with 'other roads' as category. The weight value for scenic roads $\beta_{scenic} = 0.5$ indicates that scenic roads are perceived two times shorter than non-scenic roads. The weight $\beta_{Cycleway}$ was chosen out of [0.5, 0.7, 1], since separate cycle lanes are expected to reduce the perceived distance over pedestrian paths and service roads (Other roads).

For almost all possible permutations of these parameter values for the two cost functions (as explained in section 3.3.1), routes are generated, and the performance is measured. In Table 9 the resulting consistency index CI, reproduction rate for different overlap threshold values RR and average path size factor over all choice sets PS (avg/CS) are given for different weight values for the scenic cost function.

The different cost functions do not seem to perform much different and thus an additional indicator is determined to give somewhat insight into the generated routes. A cost impact I_{cost} measure is introduced,

which measures the impact of the chosen weights on the generated route set as well. The definition is as follows:

$$I_{cost} = \frac{1}{R} \times \sum_{r=1}^{R} \frac{C_r}{L_r} \quad (24)$$

For which C_r and L_r are the cost and length for a generated route r respectively. The impact of the cost is determined by taking the average cost vs length rate over all generated routes R. If $I_{cost} = 1$, then there are probably no links of type k included for which a weight $\beta_k \neq 1$ is included in the cost function.

Table 9. Performance of scenic cost functions.

	CI	RR (for different overlap threshold values)								PS (avg/ CS)		I_{cost}
β_{scenic}		30%	40%	50%	60%	70%	80%	90%	100%	Mean	St. dev	
1	52.6%	70%	56%	50%	44%	28%	28%	18%	8%	0.177	0.057	1.00
0.5	53.3%	72%	60%	50%	46%	28%	26%	18%	10%	0.176	0.056	0.65
0.7	51.0%	70%	54%	46%	36%	24%	24%	16%	8%	0.174	0.053	0.79
1.5	50.2%	72%	56%	44%	38%	24%	24%	16%	6%	0.172	0.053	1.33
2	52.0%	76%	58%	48%	38%	26%	20%	14%	6%	0.176	0.056	1.65

In Table 10 the same is done for six of the 75 tested permutations of weights for the road category cost function. The first three permutations had the highest CI values and the other three the lowest CI values.

Table 10. Performance of road category cost functions.

	CI	RR (for different overlap threshold values)								PS (avg/ CS)		I_{cost}
Pm		30%	40%	50%	60%	70%	80%	90%	100%	Mean	St. dev	
a	54.1%	72%	64%	58%	46%	34%	24%	14%	6%	0.178	0.049	0.98
b	53.4%	74%	60%	52%	40%	30%	26%	18%	10%	0.179	0.053	0.69
c	52.7%	70%	58%	50%	38%	34%	26%	18%	8%	0.179	0.054	1.44
d	49.6%	70%	54%	46%	40%	28%	14%	14%	4%	0.183	0.074	1.19
e	49.5%	70%	54%	46%	36%	28%	16%	14%	4%	0.182	0.072	1.13
f	50.1%	70%	56%	46%	38%	26%	22%	16%	6%	0.175	0.058	1.13

	$\beta_{Cycleway}$ - β_{Large} - β_{Small}
Permutation	***a: 0.5 – 2 – 0.7***
	b: 0.5 – 1 – 0.5
	c: 1 – 2 – 1.5
	d: 1 – 0.5 – 2
	e: 0.5 – 0.5 – 2
	f: 0.5 – 1 – 1.5

The different cost functions did not make a substantial difference in the performance of the route sets compared to the distance-only cost function indicated in Table 9 (β_{scenic}=1), even though the cost impact $Icost$ changed substantially. The mean and standard deviation values for the distribution of the consistency index CI are 51.7% and 1.1% respectively. The mean and standard deviation values for the distribution of the average path size factor for a choice set $PS\ (avg/CS)$ are 0.178 and 0.003 respectively. The differences seem to be caused by the randomness included in the method rather than the different weights provided. This randomness is inherent to the method since routes are generated by determining a new least cost path by eliminating a random network link on the previous least cost path. Generating a route set with the same set of weights, leads to a slightly different routeset and thus a slightly different performance.

Only a small part of the routes was generated completely. Almost all route sets for all OD pairs where of the same size. The route generation method could almost always generate the desired number of routes of 20.

On average, 98% of the maximum of 20 routes to be generated per OD pair was generated. The similarity of the performances is, because the generated routes are found to be similar (see Figure 37 for some examples).

Nonetheless, a joint cost function was made by combining the cost functions of both the attributes with the most promising parameter values in terms of CI such that:

$$C_a = [0.5 \cdot Cycleway_a \cdot Length_a] + [2 \cdot Large_a \cdot Length_a] + [0.7 \cdot Small_a \cdot Length_a] + [Other_a \cdot Length_a] + [0.5 \cdot Scenic_a \cdot Length_a] + [Nonscenic_a \cdot Length_a] \quad (25)$$

This combined cost function does not perform substantially better than the conventional distance-only cost function (see Table 11). A choice for the best function between the tested multi-attribute cost functions and different weight sets is difficult to make, since they are similar in terms of performance. The observed route choice behavior could not be sufficiently be replicated with these attributes included in the cost function. This emphasizes the need for a route choice model when generating route alternatives in practice. Adding other attributes to the cost function such as cycle lanes, right of way, turns, elevation, land-use and car traffic volumes are expected to improve the performance of the route generation method.

Given this finding that the tested multi-attribute cost functions do not improve the route sets, the distance-only cost function is used to generate the routes for the big dataset. The conventional distance-only cost function does not perform much different but is often used and accepted in previous research.

Table 11. Performance of the final route set based on the combined cost function.

CI	RR (for different overlap threshold values)								PS (avg/ CS)		I_{cost}
	30%	40%	50%	60%	70%	80%	90%	100%	Mean	St. dev	
52.1%	70.0%	60.0%	48.0%	40%	30%	26%	16%	8%	0.176	0.054	1.66

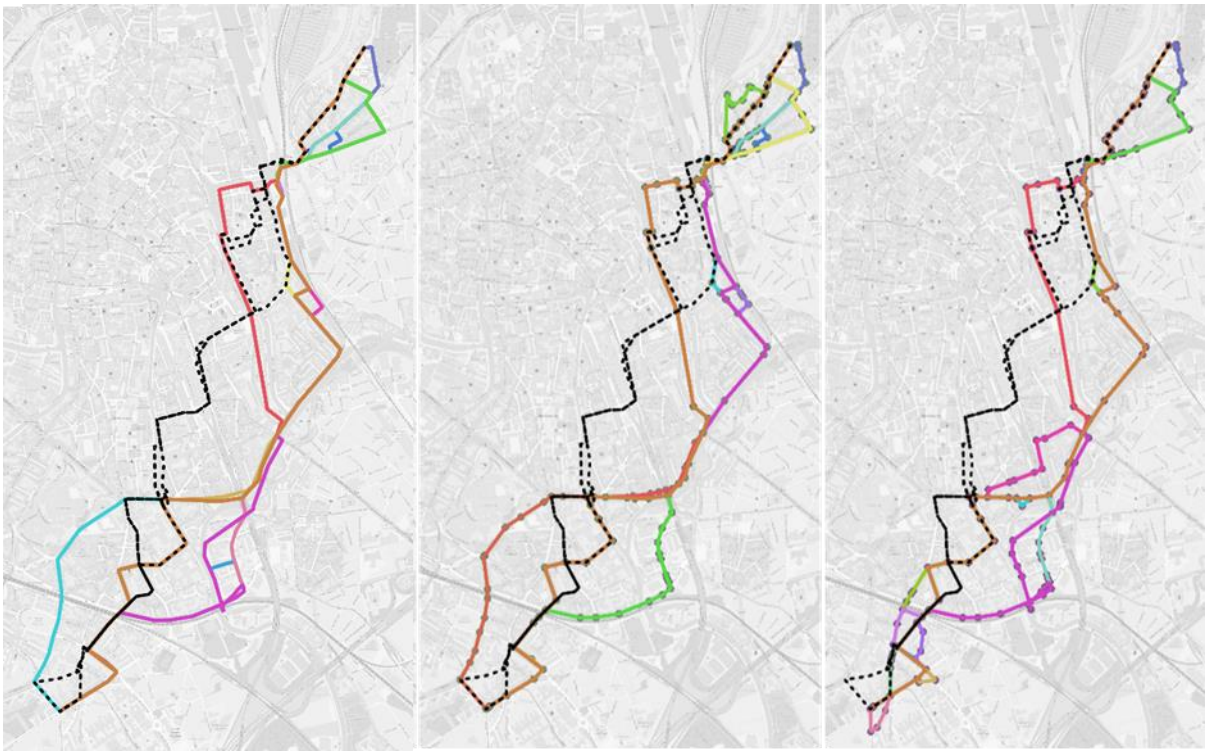


Figure 37. Comparison examples generated route sets for the same OD pair for distance-only - (l), road category- (m) and combined cost functions (r) with three observed routes for the OD pair in black dashed lines and the different generated routes in different colors.

6.3 FINAL ROUTE SET

The performance of the final route set based on the distance-only cost function is displayed in Table 12. The routeset seems to perform as well as the small set used for the calibration of the cost functions. The routes seem to be more unique on average, indicated by the higher value for the average path size factor for a choice set $PS\ (avg/CS)$.

Table 12. Performance of the final route set based on the distance-only cost function.

CI	RR (for different overlap threshold values)								PS (avg/ CS)		I_{cost}
	30%	40%	50%	60%	70%	80%	90%	100%	Mean	St. dev	
52.9%	77%	63%	47%	36%	28%	22%	17%	13%	0.184	0.068	1.00

An overview of the route set is given in terms descriptive statistics of the explanatory variables in Table 13. The generated routes are compared to the observed routes as well. The differences in mean and standard deviation between the generated routes and the observed routes are given in the last columns as well.

Table 13. Descriptive statistics per considered attribute of the generated routes compared to the observed routes.

		OBSERVATIONS			*GENERATED ROUTES*			*Diff. Mean*	*Diff. St. dev.*
	Unit	*Mean*	*Median*	*St. dev.*	*Mean*	*Median*	*St. dev.*		
Length	*km*	6.97	5.01	6.31	7.55	5.70	6.78	8%	8%
PS	-	0.52	0.54	0.25	0.16	0.12	0.12	-69%	-50%
Wrong way	*km*	0.03	0.00	0.06	0.03	0.00	0.06	-7%	-2%
Cycleway	*km*	0.91	0.31	1.47	1.14	0.43	1.60	25%	9%
Large road	*km*	2.03	1.30	2.22	1.48	1.21	1.34	-27%	-40%
Small road	*km*	2.62	2.08	2.11	2.14	1.64	1.91	-18%	-9%
Other road	*km*	1.40	0.22	2.95	2.79	0.63	4.99	99%	69%
Green	*%. km*	2.92	1.36	3.39	3.67	1.56	4.31	26%	27%
Water	*%. km*	0.59	0.20	0.96	0.89	0.38	1.40	52%	46%
Building density	*%. km*	1.10	0.99	0.72	0.81	0.67	0.53	-26%	-26%
Scenic (G+W based)	*km*	5.52	2.98	5.79	6.49	3.55	6.84	18%	18%
Roundabouts	*#*	0.56	0.00	0.96	0.43	0.00	0.69	-23%	-28%
Traffic Lights	*#*	1.25	1.00	1.65	0.74	0.00	0.97	-41%	-41%
Left turns	*#*	4.77	4.00	3.82	4.60	4.00	3.10	-3%	-19%
Right turns	*#*	4.93	4.00	3.71	4.63	4.00	2.91	-6%	-22%
Intersection cross (no turn)	*#*	46.53	41.00	32.38	27.04	26.00	15.82	-42%	-51%
Up-slope 2-4%	*km*	0.18	0.11	0.20	0.13	0.08	0.15	-26%	-23%
Up-slope 4-6%	*km*	0.02	0.00	0.04	0.02	0.00	0.05	0%	24%
Up-slope >6%	*km*	0.02	0.00	0.04	0.01	0.00	0.04	-15%	20%

The generated routes are similar in terms of length, wrong way distance, and number of turns. However, they differ quite significantly with respect to the other variables. All the differences in these exploratory variables between the routes should give some indication on why not all the observed routes are reproduced.

The generated routes have a much lower PS factor on average compared to the observed routes, indicating that the generated routes are far less unique in their choice sets and differ significantly from the observed routes. The generated routes seem to make use of other roads much more and less of large and small roads.

The other roads include unpaved paths through woods as well, which are not used by the E-bikers included in this study. The generated routes make use of links with more greenery, more water and less buildings in their direct environment. There are also less traffic lights and roundabouts included on the generated routes. A big difference in the number of intersection crossings (no turns) is also noticed. Perhaps, the observed cyclists cross many intersections where they have priority over crossing traffic and prefer to cross those intersections instead of making turns on other intersections where they must stop. Information about right of way (or stop signs) was unfortunately not available and hence excluded from this research. In terms of the proportion of up-slopes, the generated routes are somewhat similar to the observed routes.

The correlation coefficients between each attribute pair within the routeset have been determined (Table 14), such that an overview can be given of the independence of the attributes. If attributes are correlated, then the utility function with the combination of these corelated attributes could be avoided.

Table 14. Attribute correlation matrix.

	Length	log(PS)	Wrong way	Cycleway	Large road	Small road	Other road	Green	Water	Building density	Scenic	Roundabouts	Traffic Lights	Left turns	Right turns	No turn	Up-slope 2-4%	Up-slope 4-6%	Up-slope >6%
Length	1.00																		
log(PS)	-0.26	1.00																	
Wrong way	-0.16	0.01	1.00																
Cycleway	0.10	-0.10	0.01	1.00															
Large road	-0.26	0.04	0.07	-0.13	1.00														
Small road	-0.49	0.22	0.11	-0.28	-0.33	1.00													
Other road	0.63	-0.19	-0.17	-0.23	-0.42	-0.54	1.00												
Green	0.57	-0.10	-0.22	0.07	-0.33	-0.37	0.58	1.00											
Water	0.48	-0.20	-0.09	0.04	-0.29	-0.36	0.56	0.11	1.00										
Building density	-0.64	0.19	0.24	-0.23	0.20	0.58	-0.59	-0.87	-0.33	1.00									
Scenic	0.53	-0.13	-0.22	0.20	-0.31	-0.37	0.49	0.85	0.37	-0.90	1.00								
Roundabouts	-0.06	0.05	0.00	-0.03	0.15	-0.05	-0.06	0.02	-0.05	-0.07	0.06	1.00							
Traffic Lights	-0.17	0.03	0.06	-0.02	0.36	-0.11	-0.18	-0.26	-0.17	0.18	-0.26	-0.04	1.00						
Left turns	-0.45	0.15	0.09	-0.20	-0.06	0.45	-0.27	-0.46	-0.18	0.54	-0.46	0.01	0.04	1.00					
Right turns	-0.48	0.16	0.08	-0.17	-0.04	0.45	-0.30	-0.48	-0.20	0.55	-0.48	0.03	0.08	0.76	1.00				
No turn	-0.62	0.26	0.22	-0.13	0.25	0.43	-0.54	-0.71	-0.28	0.75	-0.65	0.03	0.19	0.26	0.29	1.00			
Up-slope 2-4%	-0.24	0.08	0.05	-0.11	0.11	0.17	-0.19	-0.34	-0.02	0.34	-0.32	-0.01	0.09	0.27	0.28	0.32	1.00		
Up-slope 4-6%	-0.13	0.02	0.07	-0.03	0.00	0.18	-0.16	-0.19	-0.04	0.24	-0.21	-0.04	0.00	0.15	0.15	0.20	0.14	1.00	
Up-slope >6%	-0.05	0.02	-0.03	0.02	-0.04	0.10	-0.07	-0.03	0.03	0.06	-0.04	0.05	-0.03	0.08	0.07	0.03	0.14	0.22	1.00

The length attribute has rather strong correlations with many of the other attributes, as expected. Almost all attributes are normalized in terms of the route length. The larger the proportion of links with a specific attribute, the stronger the correlation. Longer routes contain fewer small roads and more other roads, as well as less buildings and intersections. The longer trips mostly take place outside of the city center, along long scenic roads, paths and unknown roads.

A very strong negative correlation is found between the green and building density attributes. This is expected as well, since a larger built area leaves less space for greenery. The scenic attribute is based on the green- and water area, which justifies the strong positive correlation with those attributes and strong negative correlation with building density.

The No turn attribute is negatively correlated with the green attribute and thus positively correlated with the building density, which is most likely a result of long rural roads along greenery. These rural roads have less intersections than the roads in the dense urban network. Left turns and right turns are positively correlated. A higher proportion of turns in one direction seems to be associated with an increase in the proportion of turns in the other direction. The left- and right turn attributes do not seem to be significantly correlated with the no-turn attribute. Surprisingly, the other intersection-based attributes (roundabouts, traffic lights, turns)

have very less correlations. Traffic lights are positively correlated with large roads, as most traffic lights are spread along these roads (see Appendix B).
The up-slope attributes have very little correlations with all other attributes. Perhaps that is because there are very less roads on the routes with these up-slopes and thus only make up a very small proportion of the route. The majority of the up-slopes can be found on the shorter routes in the dense built areas (see Appendix D) and thus justifies the negative correlation with length and positive correlation with building density.

7 RESULTS OF THE PATH SIZE LOGIT MODEL

The results of the PSL route choice model are discussed in this chapter. The following question will be answered in this chapter:

> What attributes do E-bikers find important in their route choice and to what extent?

Different permutations of the attributes mentioned in Table 3 are tested in the PSL model. These attributes include: the route length, PS factor, different road categories, turns, one-way restrictions, land use (green-, building- and water area nearby the link as well as a scenic attribute combining the previous attributes into dummy values), intersection control (roundabout and traffic lights) and up-slopes. The final model is presented (section 7.1) and the parameter estimates are discussed (section 7.2). In order to compare the relative importance of the different attributes, each attribute is attempted to convert into a distance value. This is done by taking the trade-off rates between the parameters (section 7.3).

7.1 MODEL ESTIMATES

In total over 70 models with different permutations of the attributes were tested. The utility function of the PSL model is expanded step by step. First, the most relevant attributes are added separately to the length and PS factor to get a 'feeling' of the impact of an attribute on the model fit. Then, different permutations of the attributes are tested, while keeping track of the model fit and the significance of the estimated parameters. Combinations of attributes that are highly correlated are avoided in the utility function of the choice model. For instance: combinations of green and building density, no turn with green and building density, scenic and green or building density. The model fit in terms of the final loglikelihood (Final LL) and $\bar{\rho}^2$ are logged for each tested model. The fit of the models is compared using the likelihood ratio statistic LRS (eq. (13)). The parameter estimates for the best performing model in terms of model fit are given in Table 15. A full overview of the results of all the tested models can be found here[5].

The estimated parameters were used to predict the choices of the validation data set. The obtained loglikelihood was -2531 and the hit-ratio was 63%.

[5] Link: shorturl.at/pqJKW

Table 15. Parameter estimates with their respective standard deviations and t-values for the chosen PSL model.

Attribute	***Unit***	***Parameter Value***	***St. dev.***	***t-value***
Length	km	-0.65	0.03	-20.20
Ln (PS)		2.42	0.03	69.50
Cycleway	*	3.92	0.31	12.50
Large road	*	3.24	0.28	11.50
Small road	*	4.53	0.27	16.90
Left turns	per km	-0.23	0.04	-5.32
Right turns	per km	-0.20	0.04	-4.80
Roundabouts	per km	0.56	0.15	3.82
Traffic Lights	per km	0.54	0.10	5.26
Wrong way	*	-5.47	1.42	-3.85
Green area	**	-4.70	0.40	-11.90
Up-slope 2-4%	*	10.40	0.86	12.00
Up-slope 4-6%	*	16.40	2.13	7.71
0 LL			-12772	
Final LL			-5863	
LRS (init. model)			13817	
$\bar{\rho}^2$			0.54	

** proportion of route length*
*** weighted proportion of route length*

7.2 DISCUSSION OF PARAMETER VALUES

The computation time of estimating such a model in PandasBiogeme (Bierlaire, 2018) took roughly 3 minutes. All included attributes are found to be highly significant (>0.1% significance level). The other tested parameters that were not found to be significant and do not improve the model fit are: water area, building area and up-slopes >6%. The parameter estimates for each attribute including these insignificant ones are discussed in this section.

As expected, the parameter value for the *length* is negative, indicating that in general these cyclists dislike long distances. Also consistent with expectations is that these cyclists are willing to extend the trip distance in order to ride on cycleways, indicated by the positive parameter value.

Less expected is the preference for *large roads*. Note that many large roads (primary, secondary and tertiary roads) include bike lanes, which are not specified in the dataset. Riding on these bike lanes is expected to increase the utility (Hood et al., 2011, Casello & Usyukov, 2014, Skov-Petersen et al., 2018). Surprisingly, there seems to be an even bigger preference for riding on *small roads* than riding on *cycleways*. The larger preference for small roads over large roads can be related to the arguably lower traffic volume on the small roads. The positive sign on these three parameters related to road categories indicate that there is a dislike for the other roads (pedestrian/ path/ tracks, service and unknown), which is used as reference for road category in this model. This dislike for other roads is also found in other models where one of the three road category attributes was excluded instead.

As expected there seems to be a dislike for *turning left or right* at intersections. The previous studies (indicated in appendix A) which include the turn attributes all found that turns are associated with a disutility as well. The dislike for turning in general for e-bike users most likely has to do with them rather cruising on high speeds

without stopping and interacting with other road users. Turning left results in a higher disutility than turning right as expected, since it requires crossing the road and considering other road users. Turning right does not require crossing the road and requires less consideration of other road users.

These cyclists seem to prefer *roundabouts* on their route. Prato et al. (2018) also found that roundabouts have a positive impact on the route choice. They relate this preference to the right of way that cyclists have on roundabouts over car traffic in Copenhagen. However, in Ghent this does not seem to be the case. It was expected that roundabouts would have a negative impact caused by to the interaction required with other traffic when making a left turn on an intersection with a roundabout. The cycleways are often not connected at the roundabout and thus requiring leaving the cycleway, riding on the road in between cars and rejoining the cycleway on the other side. An example of such a situation is illustrated below. The preference for roundabout found is more likely to be related to the fact that the generated alternative routes in the choice set make less use of the large roads than the observed route. Since almost all the roundabouts are on intersections between these large roads.

Figure 38. Illustration of the path (black dotted arrow between red cycleways) that cyclists make to turn left on a roundabout (top) and an image of a cyclist making that crossing maneuver on that same roundabout. Source: Google Maps.

Traffic lights are found to have a positive effect as well. It would be interesting to consider the combination of a type of turn and the traffic volume with traffic lights as interaction attributes, as this preference could then be better justified. Broach et al. (2012) have found that cyclists in Portland have a disutility for turning left or going straight at a signalized intersection. Cyclists in Portland can turn right on cycleways and -lanes without having to wait, as is the case for cyclists in Ghent. It is thus expected that these right turns at signalized intersections with this dataset of Ghent might be insignificant, but left turns and straight crossings of these

intersections might indeed be preferred over crossing the same busy intersection without traffic lights. The observed routes in this study might consist of many more of such crossings or left turns at busy signalized intersections than the generated route alternative and thus justifying the positive impact found for traffic lights.

An aversion for cycling in the *wrong way* in one-way streets is also found, as expected. Cycling in the opposite direction is not allowed on most of the roads in the Ghent region, except for one-way restricted residential streets. The alternative routes are generated such that cycling in the opposite direction of one-way residential streets is allowed. This cycling in the opposite direction perhaps causes some discomfort caused by extra mental effort needed to cycle safely through traffic coming in the opposite direction.

As for land use, only the *green area* attribute is found to be significant. Surprisingly, routes along greenery are avoided by these cyclists. Ghanayim & Bekhor (2018) found scenic areas to have a positive impact. Prato et al. (2018) also found that cycling through green scenic areas reduces the perceived distance, but only in combination with high temperatures. In all other circumstances tested by Prato et al. (2018), green scenery is found to increase the perceived distance such as is the case in this study. After visually checking some random observed routes, it seems that the generated routes include many pedestrian paths along lots of greenery, which are not cycled on by these cyclists. This finding is backed by the large distances on other roads for the generated routes indicated in Table 13. Another possibility is that routes through woods and parks lack proper lighting, which makes the cyclist feel more unsafe and therefore dark paths through green are avoided at night. Interaction with streetlights and daylight with the land-use attributes would be interesting to look at.

The other land use attributes tested (building– and water area) are not found significant. This might suggest that the E- bikers included in this study are indifferent to roads along dense built areas and roads nearby canals and lakes. A point of discussion is of course the definition of the attributes. The water area is defined as the average fraction of buffer area around streets covered with water. This buffer area is chosen to be within 30m distance of the link. It is possible for a water body to be in sight for a cyclist on a road but not within the 30m distance of the link and therefore the link is not classified as one with water within its buffer. This might especially be the case around parks and rural areas, where there are empty fields between the roads and canals, rivers or lakes. The other way around might also be possible, that there might be some water within the buffer area of a link but not in direct sight of the cyclist due to separation of buildings. As mentioned in section 6.1, the green area attribute also includes greenery within the buffer that is not directly visible by the cyclists, thus exaggerating the perceived greenery by the cyclist. Models with different attribute permutations with the scenic attribute mentioned in section 6.1 were tested, but the model fit did not improve significantly (based on the Likelihood Ratio Test explained in section 2.1.5).

The biggest surprise in these findings is that of the preference for *up-slopes*. Up-slopes of 2-4% and 4-6% are found to have a large positive impact on the route choice. Up-slopes of 6% are not found significant however. It would be reasonable that E-bikers experience less disutility compared to normal cyclists or even that E-bikers are indifferent towards up-slopes since they need to put in less effort in climbing uphill than conventional cyclists. However, such a large positive parameter estimate was not expected. The large estimate could be related to links with up-slopes that are inevitable for certain OD-pairs within the city center without having to make large detours. Many short links with big slopes are located within the city center of Ghent (see Appendix D). It could also be possible that the up-slope attribute is strongly correlated to attributes that are not considered, such as unpaved roads, and therefore the up-slope attributes serve as proxies for the other attributes. Even though the parameter estimates for up-slopes are large, the impact on the utility function is not much bigger than other attributes since the attribute values for the proportion of links with up-slopes are very small. The observed routes have on average 0.18 km of up-slopes between 2-4% while the average route length is around 6 km (see Table 13). A noteworthy discussion point is that the up-slope data is based on elevation difference between link start- and end nodes and thus any elevations in between are ignored. These

elevations in between the link nodes might also be relevant such as in the case of bridges and tunnels. A link might appear to be 'flat' based on this data, but in reality, it contains a huge up-slope which the cyclists consider in their route choice.

It is noteworthy that there are substantial positive correlation values found between the estimated parameters for the road categories, ranging between 0.67 and 0.83.

The signs of the resulting parameter estimates seem to be related to the respective differences in average attribute value between the observed and generated routes given in Table 13. The large- and small roads, roundabouts, traffic light and up-slope parameter estimates are all positive, while the generated routes on average have lower attribute values for each of these attributes compared to the average of the observed routes.

The validity of the resulting model might be questioned, since it is not able to predict choices much better than a random model given the hit-ratio of 63%. The route generation method not being able to reproduce the observed behavior most of the time resulted in these big differences between the generated routes and the observed routes and thus impacting the outcome of the route choice model.

7.3 DISTANCE TRADE-OFFS

The marginal rates of substitution between each estimated parameter with the distance are given in

Table 16. Each rate $r_{x/L}$ gives an indication of the distance value for a unit change of the attribute variable x, and is determined as follows:

$$r_{x/L} = \frac{\beta_x}{\beta_L} \tag{26}$$

In which β_L is the parameter estimate for the distance L and β_x is the parameter estimate for an attribute x.

Note that the final parameter estimates are the average preference over all individuals. These rates thus indicate the average distance trade-off as well. There are two types of attributes considered: the proportion-based attributes (such as: cycleway, large roads, small road, wrong way, green area and the up-slope attributes) and the units-per-km based attributes (such as: turns, roundabouts and traffic lights attributes). Attributes of the same type can be easily compared in terms of their relative importance w.r.t. the distance trade-off, while it is not so straight forward to compare the different types of attributes.

A 100% increase in the proportion of cycleways on a route is thus equivalent to a 6km reduction in distance given that all other attributes stay the same. Or vice versa, a reduction of the route distance of 1km is equivalent to an increase in the proportion of cycleways of 16.7%. These E-bikers are thus willing to travel longer on routes with a higher proportion of cycleways, given that all other attributes of the routes are equal. They are willing to travel on even longer routes with higher proportions of small roads and much longer routes with higher proportions of up-slopes. The E-bikers perceive the route distance to increase with higher proportions of wrong way streets and streets along green areas.

In terms of the unit-per-km based attributes, left and right turns are associated with an increase in perceived travel distance, while roundabouts and traffic lights have the opposite and larger effect.

Table 16. Distance trade-off rates of each parameter.

Attribute	***Unit***	***Ratio*** $r_{x/L}$
Cycleway	$km/\Delta prop$	-6.00
Large road	$km/\Delta prop$	-4.96
Small road	$km/\Delta prop$	-6.94
Left turns	$\frac{km}{turn/km}$	0.35
Right turns	$\frac{km}{turn/km}$	0.30
Roundabouts	$\frac{km}{unit/km}$	-0.86
Traffic Lights	$\frac{km}{unit/km}$	-0.83
Wrong way	$km/\Delta prop$	8.38
Green area	$km/\Delta\ \mathrm{w}.prop$ *	7.20
Up-slope 2-4%	$km/\Delta prop$	-15.93
Up-slope 4-6%	$km/\Delta prop$	-25.11

$\boldsymbol{\Delta prop}$*= unit change in proportion*

** kilometers per unit change of weighted proportion*

8 CONCLUSIONS AND RECOMMENDATIONS

The research is concluded by providing an answer to the research question based on the results (section 8.1). An overall discussion on the findings and limitations is given (section 8.2) and recommendations are provided for future work (section 8.3) and practice (section 8.4) as well.

8.1 CONCLUSIONS

This research aimed at analyzing route choice behavior of E-bikers, in terms of which factors play a role and to what extent. The research is performed based on E-bikers route data in Ghent. This research included three interesting developments as well, in terms of multi-attribute cost functions as input for route generation, including land use attributes to cyclist's route choice models and comparing the most promising path-based method with the recent Recursive Logit model.

The sub-research questions are addressed one by one, followed by the main question.

Which route choice model should be used based on literature?

Based on the literature study, the following three main types of route choice models exist based on their model structure: Logit-, Generalized Extreme Value- (GEV) and Non-GEV structures. Logit models are models that keep the simple MNL structure and can be extended to cope with correlation of the route alternatives. The GEV model structures deal with correlations in the stochastic part of the utility function, have a tree structure and relate the network topology to the specific coefficients, but they do not consider taste variation or correlation over time of unobserved factors. Non-GEV model structures also allow random taste variation and correlation in unobserved factors over time. These models do not present closed-form expression for the choice probabilities. Their estimation is based on simulation and thus more complex than the other models.

For this research, heterogeneity of preferences and correlation of unobserved factors are neglected. For estimating the general route preferences of E-bikers, the logit structured models should suffice. The simple Logit structured and well-established Path Sized Logit (PSL) model is used in this research. An attempt is made to estimate the relatively new and promising logit structured Recursive Logit (RL) model as well. Several attempts were made to get the RL model to converge consistently for different initial parameter values without success. Because of the complexity and lack of experience with this model, the exact source of the problem could not be identified within this project. However, a qualitative comparison is still made. The PSL model is thus the route choice model chosen for this study.

Which route generation method should be used based on literature?

An overview is made on the different types of *route generation methods*. Four types of route generation methods are distinguished: Deterministic shortest path-based, Stochastic Shortest path-based, Constrained Enumeration methods and Probabilistic methods. Most of the route generation methods are of the Deterministic shortest path-based type. The cyclist's route choice behavior studies apply a wide variety of route generation methods with the majority being the Labeling -, Breadth First Search on Link Elimination (BFS-LE) - and Doubly stochastic generation function (DSGF) methods. The first two methods are of the Deterministic shortest path-based type while the DSGF is of the Stochastic shortest path-based type.

Halldórsdóttir et al., 2014 compared, from what they found in literature, the three most promising methods for car route generation in the cycling context. The three methods they compared are: Breadth First Search on Link Elimination (BFS-LE), Doubly stochastic generation function (DSGF) and Branch & Bound (B&B). They

applied multi- attribute cost functions by adding different weights to the distance of links with different road types, cycle lanes and land use attributes. They state that multi-attribute cost functions increase the performance of the route generation methods in terms of consistency with observed behavior and heterogeneity of the choice set. However, they do not compare the performance achieved by using the multi-attribute cost functions with that of the basic distance-only function for the same dataset.

The BFS-LE and DSGF method had similar performance in terms of consistency. The DSGF had a slightly better heterogeneity in terms of route choice set than the BFS-LE method, but the BFS-LE was by far the most computationally efficient of the three. The BFS-LE method is not the most applied method in the cyclist's route choice studies, which arguably could be related to the fact that none of the studies considered multi-attribute cost functions except for Prato et al. (2018). Due to the great performance and efficiency of the BFS-LE method when using multi-attribute cost functions, it is the route generation method of choice for the PSL route choice model in this research.

Which relevant attributes should be included in the route choice model based on literature?

The performed literature study, on studies about cyclist's route choice analysis using GPS data, lead to a list of *attributes* that were found to be of significance on cyclist's route choice behavior. None of the studies included E-bikers route choice behavior, but the same attributes are expected to be of relevance for E-bikers. The attributes are grouped into the following clusters: infrastructure related-, socio-demographic, environmental and land use attributes. Based on these studies, a list of attributes with an indication of their relative impacts on the route choice behavior is made.

A selection has been made from the list of attributes mentioned above. From the list of infrastructure related attributes, all attributes are considered in this research except for cycle lanes, traffic volume, car speed limits, bridges, number of lanes, surface quality (paved/ not) and traffic safety (number of crash incidents). Data for these attributes was not available. Socio-demographic attributes are not considered since the data consists of a small set of 21 people and are therefore not expected to be of significance. Environmental attributes like weather, temperature, wind and daylight were not considered, as they are expected to be of little to no significance based on the findings of previous studies and are expected to have more impact on the mode choice than the route choice. Land use attributes are included.

The list of attributes considered for this study is made based on the most included and significant attributes in the related studies and the availability of the data. These are: Distance, Cycleways, Large roads, Small roads, Roundabouts, Traffic lights, Left turns, Right turns, Wrong way, Green area, Water area, Building area, Up-slopes 2-4%, Up-slopes 4-6% and Up-slopes >6%.

To what extent does a multi-attribute cost function in the route generation improve the choice set compared to the distance-only method?

Different *multi-attribute cost functions* are considered in the route generation using the BFS-LE method. Halldórsdóttir et al. (2014) included attributes related to road types, cycle lanes and land use attributes. They were able to generate routes that were highly consistent with the observed routes. In line with them, the attributes considered are 'Road category' and 'Land use'. Links are put into mutual exclusive categories for each attribute. Links are put into mutual exclusive categories for each attribute. Calibration of the weights for each category within an attribute is done on a limited set of attributes, a small part of the data that can represent the whole data set and with limited predefined discrete weight values. Different cost functions, with each one attribute as well as a combination of attributes, are defined and their performances compared. None

of the multi-attribute cost functions are found to perform substantially better than the conventional distance-only cost function. A choice for the best function between the tested multi-attribute cost functions and different weight sets is difficult to make, since they are similar in terms of performance. Given this finding that the tested multi-attribute cost functions do not improve the route sets, the distance-only cost function is used to generate the routes for the big dataset. The conventional distance-only cost function does not perform much different but is often used and accepted in previous research.

What attributes do E-bikers find important in their route choice and to what extent?

Different permutations of the considered attributes are tested in the Path Size Logit model. These attributes include: the route length, path size factor, one-way restrictions, different road categories (cycleways, large-, small-, other roads), land use (green-, building- and water area nearby the link as well as a scenic attribute combining the previous attributes into dummy values), intersection control (roundabout and traffic lights), turns (left, right, none) and up-slopes (2-4%, 4-6% and >6%).

The attributes other than the distance that are found to be of importance (in terms of their parameter value) for E-bikers in Ghent ranging from most important to least important are: the proportion of up-slopes within 4-6%, up-slopes 2-4%, one-way restricted roads on the route, roads with greenery in the direct environment, small roads, dedicated cycleways and large roads. The number of roundabouts, traffic lights, left turns and right turns are also of importance.

However, the validity of the resulting model might be questioned, since the generated routes are not consistent with the observed routes. The route generation method not being able to reproduce the observed behavior most of the time resulted in these big differences between the generated routes and the observed routes and thus impacting the outcome of the route choice model.

How do E-bikers choose their route?

Based on the findings of the route choice analysis of E-biker's route choice observations in Ghent, a statement can be made on how E-bikers choose their route. However, keep in mind that the model may be invalid due to the poorly generated route alternatives.

E-bikers do not necessarily take the shortest path to their destinations but consider multiple route attributes when choosing their route. They prefer to ride on roads that are small or are cycleways and even large roads over paths shared with pedestrians (pedestrian paths, service roads such as parking lots, bus stops and alleys). They have an aversion towards riding in the opposite direction of one-way restricted roads and roads with lots of greenery in the direct environment. At intersections they prefer controlled intersections over uncontrolled ones with a slight preference of roundabouts over traffic lights and avoid making turns (especially left turns). They are found to be indifferent towards the amount of buildings and water in their direct environment, as well as high up-slopes.

8.2 DISCUSSION

The issues and limitations of the three different aspects of this research (data gathering and processing, route generation and choice model) are separately discussed in this section.

Data gathering and processing

The OSM network includes only cycleways as cycling infrastructure. Cycling lane information on roads is however not included in OSM network data. Inclusion of this information could improve the route generation with multi-attribute cost functions on route generation as shown by Halldórsdóttir et al. (2014). It could also increase the route choice model fit by explaining the preference for large roads (where cycle lanes are on) better.

Some links with different road categories were combined when there is no real intersection separating them by the OSMnx tool. Simplifying the categories of these combined links with diverse categories led to a loss of detail which makes the network less realistic. For example: a bikeway connected with a short pedestrian stair path are combined to a single bikeway link. A cyclist in real life, would probably not consider the cycleway, because of the stairs at the end of the cycleway. But in the model the cyclist could make use of the combined link as if there is no stair.

Route Generation

The performance of the generated route sets in this project are lower compared to what Halldórsdóttir et al. (2014) have reached in terms of consistency index and reproduction rate. Most likely this is related to the difference in data. Several reasons might include:

- This dataset of E-bike trips includes longer trips which are proven to be more difficult to reproduce than shorter ones. It is noteworthy that Halldórsdóttir et al. (2014) showed that reproducing the observed trip becomes more difficult the longer the trip is.
- It is noteworthy that Halldórsdóttir et al. (2014) don't compare the performance achieved by using the multi-attribute cost functions with that of the basic distance-only function for the same dataset. The dataset might be such that the performance of the later may already be much higher than other studies, because the data is inherently already easy to reproduce by the method.
- E- bikers are probably less sensitive to distance and therefore might accept detours in exchange for other attributes not captured in cost function. This makes it much harder to reproduce these routes. For short routes this is less of a problem.

While the BFS-LE method does produce many different routes, it does not always yield realistic routes that people would consider. Generated routes often make a detour around the removed link and follow the original shortest path. This has to do with its principle of randomizing the network by removing random links from the network. An interesting approach would be to generate different sub-route sets using different cost functions and combining them into one route set with more behavioral consistency and more heterogeneity. The different cost functions would represent the different preferences of different people.

Route generation based on a distance-only cost function provides a biased routeset for which the distance is the most important attribute. A routeset generated on a single multi-attribute cost function would arguably result in a routeset that is less biased towards distance. A routeset generated by sampling routes generated from different multi-attribute cost functions would arguably decrease the bias even more by incorporating a more diverse routeset.

The observed route choice behavior could not be sufficiently be replicated with these attributes included in the cost function. This emphasizes the need for a route choice model when generating route alternatives in

practice. Adding other attributes to the cost function that are found significant in the route choice model such as cycle lanes, right of way, turns, elevation and car traffic volumes could improve the performance of the route generation method, but requires more parameters to be calibrated.

Choice model

The validity of the resulting model might be questioned, since it is not able to predict choices much better than a random model given the hit-ratio of 63%. The route generation method not being able to reproduce the observed behavior most of the time resulted in these big differences between the generated routes and the observed routes and thus impacting the outcome of the route choice model. Calibrating the weights for the multi-attribute cost functions for route generation and afterwards estimating the preferences seems like a 'Chicken-Egg' dilemma. Nonetheless, there is some iteration needed between the two, in order to capture the choice behavior in the route generation already. This dilemma is less present in the link-based RL model, since no route generation is needed. This iteration has not been applied in this study but is highly recommended.

The GPS dataset contains multiple observations per person and thus can be classified as panel data. The estimated model assumes that the choices made for each observation is independent of the previous choices and thus the results might be skewed. Ideally, the preferences of a person should be kept constant over all the choices that that person makes. This requires extending the PSL model to a Mixed Logit model with the same path size penalty attribute.

There may be a selection bias in terms of the small set of mostly experienced cyclist's only in the dataset. This may affect the results since experienced cyclists are arguably less sensitive to some attributes like weather, turns, elevation, etc.

8.3 RECOMMENDATIONS FOR FUTURE RESEARCH

Some interesting attributes yet to be included in future work include: Light poles, Mini roundabouts, Trip purpose (even tough not explicitly given by participants, it can be inferred from the most commonly used OD's per participant.), Bike type (in terms of speed limit of E-bike), Cycle lane, Panel effects, Car Traffic volume. Additional information on traffic priority in terms of stop signs could be also an interesting addition. Crossing an intersection without having priority is hypothesized to result in a lower utility than crossing while giving priority to motorized traffic. Interaction effects are also not investigated. However, some recommendations for interaction effects to check include: Gender with each other attribute, Peak/ non-peak hour trip with each attribute, Turns with junction type (Roundabout/ TL/ Give way) and Car traffic volume with intersections. Socio-demographics are not included in this study, since a small group of 21 participants is included. The inclusion of these in the choice model would however still be interesting to look at. Even though many studies mention the possible relevance of socio-demographic attributes, almost none of the RP cyclist's route choice behavior studies have found these attributes significant in the models. Hood et al. (2011) have found gender and cycling frequency to be significant as interaction terms.

The RL model requires the data to be correctly processed beforehand. The model is complex and does not provide clear feedback about errors. It is advised to validate the data first with a simple conventional model like the PSL model and to be extra aware of units, since the model cannot cope with link utilities close to zero. Perhaps the problem of the RL model not being able to optimize for certain parameter values can be mitigated by the following:

- Non-positive error distribution for error terms in utility function can be used, such that a solution can be found irrespective of the size of the instantaneous utilities.

- Test larger threshold values for short links in the simplification process. Currently this threshold value is set to (<10m), but the expected utility might still become positive for large enough random components in the utility function.

Hood et al. (2010) suggest that the effect of the panel data on the outcome of the choice model could be mitigated even more without having to increase the complexity of the model. They notice that the number of observations per individual in the dataset differs a lot. So, they add weights to each observation in the likelihood function in terms of the inverse of the number of observations for the individual. This results in everyone having a more equal impact on the choice model.

In order to get a routeset with a higher consistency with long observed routes, the observed routes could be split into multiple sections for which route generation is done separately and later combined. Or alternatively make routing through certain 'anchor' points along routes more attractive as is done by Manley et al. (2015).

A route set performance indicator that indicates not only the consistency with observed routes in terms of distance of overlapping links, but also in terms of overlapping characteristics like cycleways or land use would be interesting to look at as well. Ghanayim & Bekhor (2018) have proposed such a generalized overlap indicator.

8.4 RECOMMENDATIONS FOR PRACTICE

The parameter estimates can be used cautiously as input for the cost functions for E-bikers for route generation in traffic models or navigation software. But it is strongly advised to do so after a feedback loop between the route generation and choice model is applied.

In terms of data collection, it is advised to let participants of route choice data collection study state their trip purpose in future in terms of utilitarian and recreational trips. Or at least let them mention the locations of their home, work and one or two regular activity locations such that the trip purpose can be inferred for the majority of the trips of that person. Also, open source network data of OSM is proven to be very useful for this type of research. Enriching this open source data set with even more attributes is recommended such that future research and projects can easily make use of high-quality data.

In terms of infrastructure investments and policy measures for the city of Ghent, E-bikers in Ghent have a big aversion for cycling in the wrong direction. Adequate route alternatives should be provided for these cyclists when measures concerning one-way restrictions are to be implemented. With respect to intersection control, E-bikers in Ghent have a slight preference for roundabouts over traffic lights. Investments in controlled intersections could be justified using these preferences in terms of perceived distances.

LITERATURE

Astegiano, P., Tampère, C. M., & Beckx, C, Mayeres, I., Himpe, W. (2017). Electric cycling in Flanders; Empirical research into the functional use of the e-bike. *Steunpunt Verkeersveiligheid*

Bierlaire, M. (2018). PandasBiogeme: a short introduction. Technical report TRANSP-OR 181219. *Transport and Mobility Laboratory, ENAC, EPFL*.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126-139.

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: Policy and Practice*, *46*(10), 1730-1740.

Casello, J., & Usyukov, V. (2014). Modeling cyclists' route choice based on GPS data. *Transportation Research Record: Journal of the Transportation Research Board*, (2430), 155-161.

Dill, J., & Gliebe, J. (2008). Understanding and measuring bicycling behavior: A focus on travel time and route choice.

Dukulis, I., Berjoza, D., & Jesko, Z. (2013). Investigation of electric bicycle acceleration characteristics. *Jelgava*, 23, 327-331.

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

Ghanayim, M., & Bekhor, S. (2018). Modelling bicycle route choice using data from a GPS-assisted household survey. *European Journal of Transport & Infrastructure Research*, *18*(2).

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences*, *138*, 557-565.

Halldórsdóttir, K., Rieser-Schüssler, N., Axhausen, K. W., Nielsen, O. A., & Prato, C. G. (2014). Efficiency of choice set generation techniques 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*, *3*(1), 63-75.

Khatri, R., Cherry, C. R., Nambisan, S. S., & Han, L. D. (2016). Modeling route choice of utilitarian bikeshare users with GPS data. *Transportation Research Record: Journal of the Transportation Research Board*, (2587), 141-149.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., & Huang, Y. (2009). Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 352-361). ACM.

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

Mai, T. (2016). A method of integrating correlation structures for a generalized recursive route choice model. *Transportation Research Part B: Methodological*, *93*, 146-161.

Mai, T., Bastin, F., & Frejinger, E. (2017). On the similarities between random regret minimization and mother logit: The case of recursive route choice models. *Journal of choice modelling*, 23, 21-33.

Manley, E. J., Addison, J. D., & Cheng, T. (2015). Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *Journal of transport geography*, *43*, 123-139.

McFadden, D. (1978). Modelling the choice of residential location. *Spatial Interaction Theory and Residential Location* (Karlquist A. Ed., pp. 75-96).

Menghini, G., Carrasco, N., Schüssler, N., & Axhausen, K. W. (2010). Route choice of cyclists in Zurich. *Transportation research part A: policy and practice*, *44*(9), 754-765.

Meyer de Freitas, L. (2018) A recursive logit multimodal route choice model. Master thesis. Institute for Transport Planning and Systems. ETH Zurich, Zurich.

Oehrlein, J., Förster, A., Schunck, D., Dehbi, Y., Roscher, R., & Haunert, J. H. (2018). Inferring Routing Preferences of Bicyclists from Sparse Sets of Trajectories. *arXiv preprint arXiv:1806.09158*.

Oortwijn, J. (2019, August 30). E-Bike Sales Skyrockets Across Europe. Retrieved October 13, 2019, from: https://bike-eu.com

Pan-European Programme. (2014). *Fourth high-level meeting on transport, health and environment.* Paris: World Health Organisation & United Nations. Retrieved from Paris Declaration: http://www.unece.org/fileadmin/DAM/thepep/documents/Déclaration_de_Paris_EN.pdf

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

Prato, C. G., & Bekhor, S. (2007). Modeling route choice behavior: How relevant is the composition of choice set?. *Transportation Research Record, 2003*(1), 64-73.

Prato, C. G., Halldórsdóttir, K., & Nielsen, O. A. (2018). Evaluation of land-use and transport network effects on cyclists' route choices in the Copenhagen Region in value-of-distance space. *International Journal of Sustainable Transportation*, 1-12.

Rupi, F., & Schweizer, J. (2018). Evaluating cyclist patterns using GPS data from smartphones. *IET Intelligent Transport Systems*, *12*(4), 279-285.

Schuessler, N., & Axhausen, K. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, (2105), 28-36.

Skov-Petersen, H., Barkow, B., Lundhede, T., & Jacobsen, J. B. (2018). How do cyclists make their way?-A GPS-based revealed preference study in Copenhagen. *International Journal of Geographical Information Science*, 1-16.

Ton, D., Cats, O., Duives, D., & Hoogendoorn, S. (2017). How Do People Cycle in Amsterdam, Netherlands? Estimating Cyclists' Route Choice Determinants with GPS Data from an Urban Area. *Transportation Research Record: Journal of the Transportation Research Board*, (2662), 75-82.

Ton, D., Duives, D., Cats, O., & Hoogendoorn, S. (2018). Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. *Travel Behaviour and Society*, *13*, 105-117.

van Overdijk, R., Van der Waerden, P., & Borgers, A. (2017). The Influence of Comfort and Travel Time on Cyclists' Route Choice Decisions (No. 17-01284).

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

APPENDICES

A. OVERVIEW OF CYCLIST'S ROUTE CHOICE STUDIES USING GPS DATA

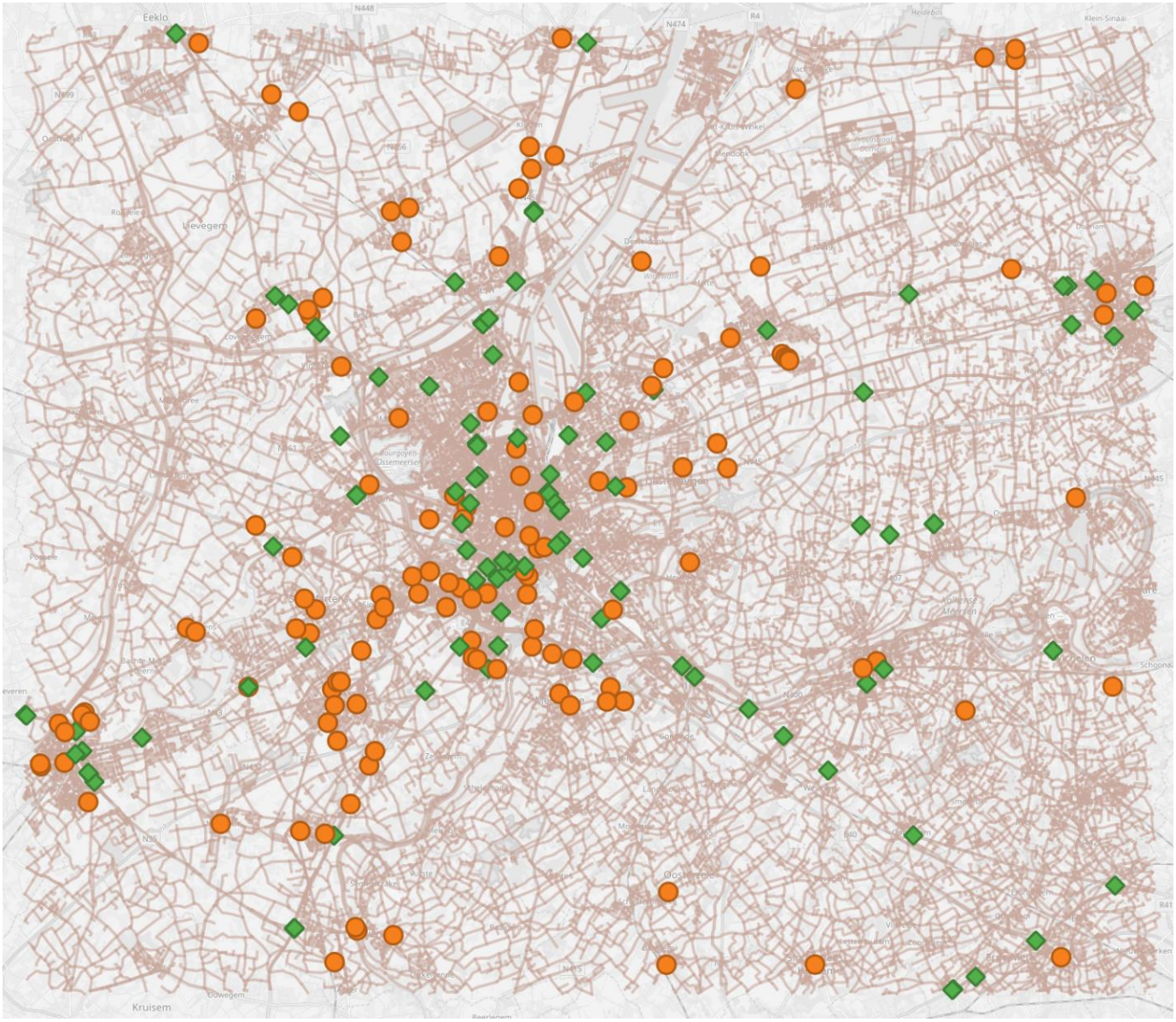
	Dill & Gliebe (2008)	*Menghini et al. (2010)*	*Hood et al. (2011)*	*Broach et al. (2012)*	*Casello & Usyukov (2014)*	*Khatri et al. (2016)*	*Ton et al (2017)*	*Zimmermann et al. (2017)*	*Ghanayim & Bekhor (2018)*	*Prato et al. (2018)*	*Ton et al. (2018)*	*Skov-Petersen et al. (2018)*
City	Portland	Zurich	San Francisco	Portland	Waterloo	Phoenix	Amsterdam	Eugene	Tel Aviv	Copenhagen	Amsterdam	Copenhagen
Data type	GPS+ survey	GPS	GPS + survey	GPS+ survey	GPS	GPS	GPS	GPS + survey	GPS + survey	GPS + survey	GPS+ survey	GPS + survey
# participants	164	2435	366	164		1866		103	221	291		183
# trips	1599	73493	2777	1449	724	9101	3045	648	545	3384	2819	1267
Network size (L x N)		24680 x 8686	33575 x 10234	88000 x 66000				42384 x 16352	127053 x 92670	361053 x 268762	25135 x 7628	64866 x ...
Choice set generation		BFS-LE	DSGF	Calibrated Route Labeling	A priori selection	Labelling	Data driven approach	None	LE & LP & Sim	DSGF	DDPI, BFSLE, Labelling	Labelling
Route Choice Model	None (Survey)	PSL	PSL	PSL	MNL	PSL	PSL	RL, RL-LS, NRL	MNL, C-logit, PSL	G-ML	PSL	CL with PS
Distance trade-off considered		y	y	y		y				y		
INFRASTRUCTURE												
Length	--	-	--	--	-	--	--	--	---	--	-	--
Bike facility	+	+	++	+	++	++	0	+	++	+	+/ ++	+++
Junction control (TL / Roundabout)	-	(+)		-		+		0		-+/+		+
Elevation (up slope)	-	---	-	---	-			---		---		
Car traffic volume	-		0	---	0	(-)		-	--			
Car speed limit			0		-	0		0		0		
Turns			(-)	-		(-)		-		--		(-)
Surface quality										+		
# intersections							-	0	(+)	-	(-)	
Oneway restrictions			---	0		-		0		---		
Bridge/ tunnel				+				--- /++		-- /++		
# lanes (>2)			0							-		
Crash incidents						0						
CYCLIST												

Gender	Y	~	0	0	~	~		0	0	~	0
Age	Y	~		0	~	~		0	0	~	0
Income	Y	~		0	~						0
Physical condition		~									
Cycling experience	Y	~	0	0	~				0		
Bike type											
Cultural background						~					
TRIP											
Purpose	Y		y	Y					y		
Departure time					y	y			y		
ENVIRONMENT (NATURAL)											
Weather	Y			Y		0			0		
Daylight						0			0		
Wind	Y			Y					0		
Temperature	Y			Y							
ENVIRONMENT (BUILT)											
Land use											
Scenery								+	+		0
Building density								++	+		---
Social safety			y								
Overlap attribute Ln(ps)		+	+	+	+	-	(-)	+ / (-)	--	-/+	+

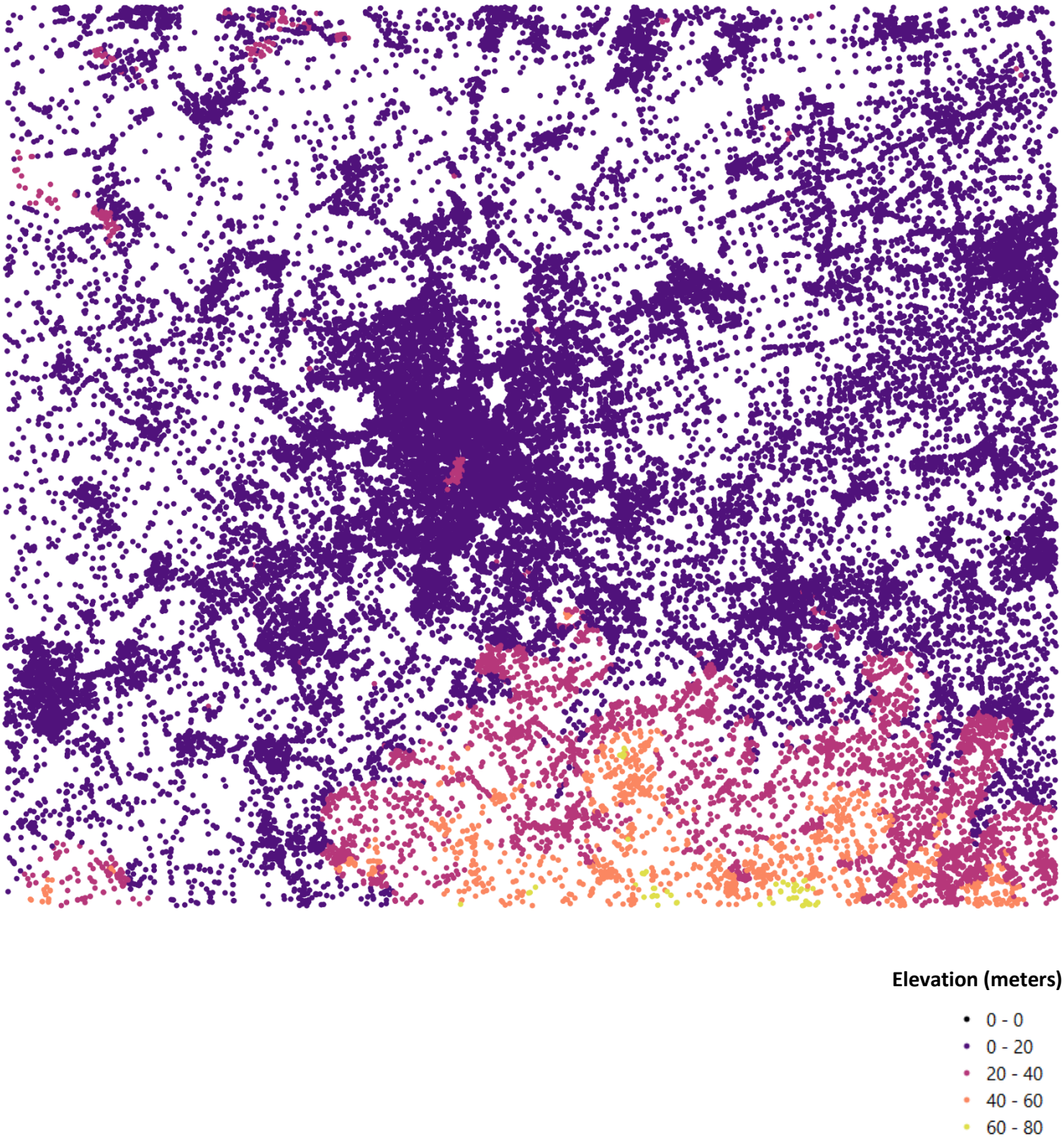
***+** Positive impact; **-** Negative impact; **+++** or **---** Very positive or negative impact; **(+)** or **(-)** Very small impact; **0** Not found significant; **Y** included; **~** not available but recommended*

Disclaimer: the relative magnitude of parameter estimates are dependent on the study's context, attribute description and – unit.

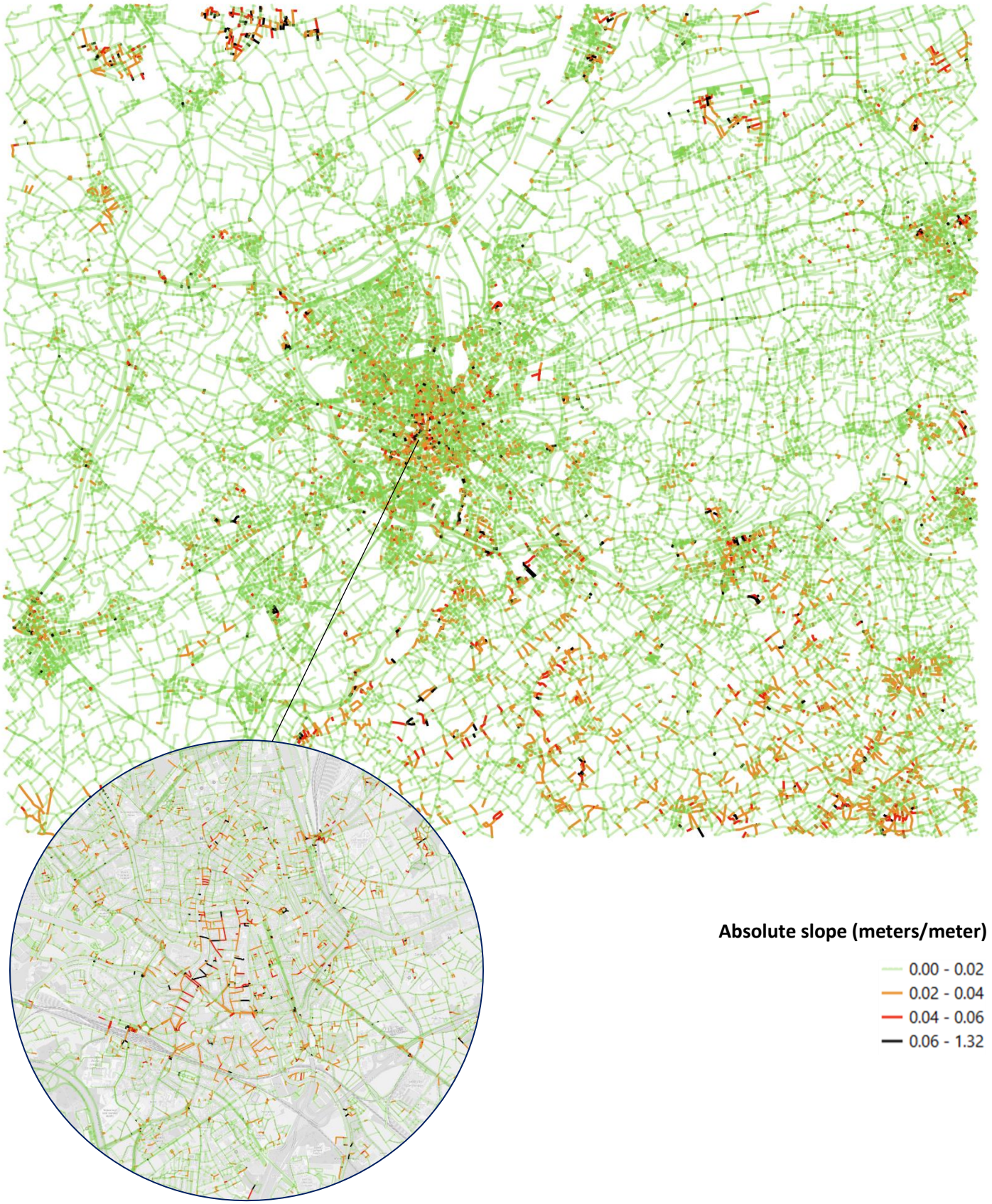
B. TRAFFIC LIGHT AND ROUNDABOUTS WITHIN REGION



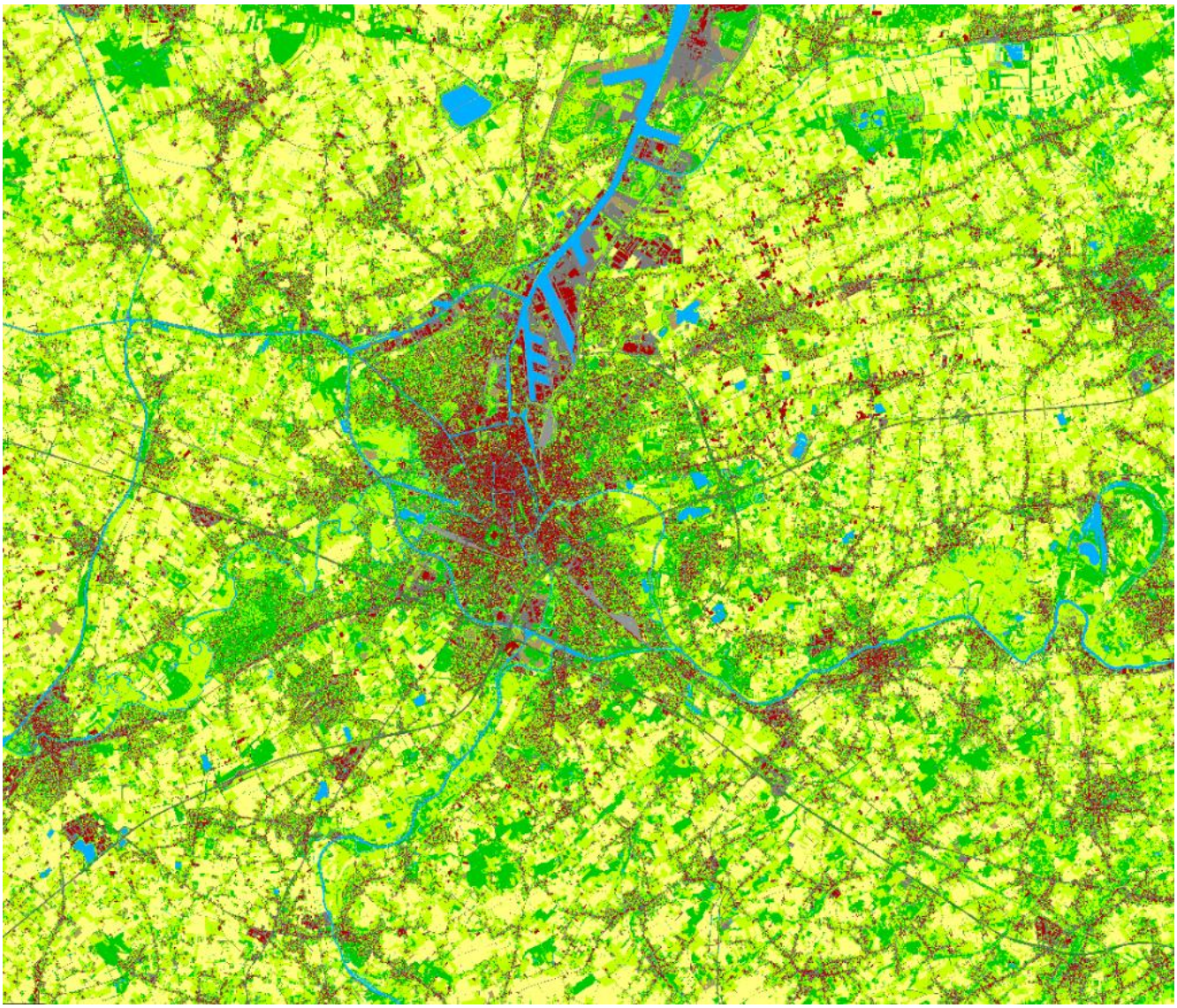
C. Node elevation map of Ghent and surroundings



D. LINK SLOPES MAP OF GHENT AND SURROUNDINGS



E. LAND USE MAP OF GHENT AND SURROUNDINGS OBTAINED FROM GRB



F. MATLAB CODE: NETWORK SIMPLIFICATION

Load

```
clear all

% uiopen('C:\Users\suraj\Documents\Data GPS+codes [PC]\1. GraphGen\Landuse (B+G+W).csv',1)
%             % change edgesG.osmid to 'categorical' data format before importing!
% edgesG=LanduseBGW; clear LanduseBGW;
% uiopen('C:\Users\suraj\Documents\Data GPS+codes [PC]\1. GraphGen\Elevation data\nodesG.csv',1)
% edgesG.uniqueid=[1:height(edgesG)]';
% % save networkraw.mat edgesG nodesG
load networkraw.mat
```

Conversion from spherical coordinates to Lambert 72

```
[nodesG.xco,nodesG.yco]=lambert72(nodesG.lat,nodesG.lon);
```

New numbering and remove unwanted variables

```
nodesG.No=(1:height(nodesG))'; nodesG.elevation=[];nodesG.ref=[];
edgesG.area=[];edgesG.width=[];edgesG.ref=[];edgesG.key=[];edgesG.name=[];

[loc1 edgesG.fromNode]=ismember(edgesG.from,nodesG.osmid);[loc1 edgesG.toNode]=ismember(edgesG.to,nodesG.osmid);
clear loc1
```

Simplify road categorization

get categories and the frequency of each category

```
        cat=table(unique(edgesG.highway)); cat.Properties.VariableNames{'Var1'}='raw';
        for i=1:height(cat)
           cat.count(i)=sum(edgesG.highway==cat.raw(i));
        end
        clear i
    % (export to excel and re-categorize manually)
    % import new categories (see: '/Network/'Road categories.xlx'')
         load ./Network/road_categories.mat categories cat
edgesG.Properties.VariableNames{'highway'}='cat_old';
for i=1:height(edgesG)
    edgesG.highway(i)=cat.new(edgesG.cat_old(i)==cat.raw);
end
edgesG.cat_old=[];
clear i cat categories

    % remove links with unwanted categories ('Remove!')
    edgesG(edgesG.highway=='Remove!',:)=[];
        % remove nodes not included in links
        include(:,1)=ismember(nodesG.No,edgesG.fromNode);
        include(:,2)=ismember(nodesG.No,edgesG.toNode);
        include(:,3)=include(:,1)+include(:,2);
        nodesG(include(:,3)==0,:)=[];
        % update numbering
```

```
        nodesG.No=(1:height(nodesG))';
        [loc1 edgesG.fromNode]=ismember(edgesG.from,nodesG.osmid);[loc1
edgesG.toNode]=ismember(edgesG.to,nodesG.osmid);
        clear loc1 include
% link numbering
edgesG.No=(1:height(edgesG))';
```

Simplify roundabouts (and add traffic lights)

```
links=edgesG;
nodes=nodesG;

% identify traffic signals
nodes.TL=(nodes.highway=="traffic_signals");

% identify roundabouts
roundabout_links=links(links.junction=='roundabout',:);

    % common nodes+ neighbours (x3)
for i=1:height(roundabout_links)
    s=roundabout_links.fromNode(i);
    e=roundabout_links.toNode(i);
    f=find(roundabout_links.fromNode==s | roundabout_links.toNode==e| roundabout_links.fromNode==e |
roundabout_links.toNode==s);%find neighbours
    rabouts{1,i}=roundabout_links.No(f);
end
clear i e s f
rabouts=uniquecell(rabouts);

for x=1:3
        for cl=1:length(rabouts) % find clusters with common links

            index=[];clu=[];
            for cli=1:length(rabouts{cl})
                index(cli,:)=cellfun(@max,cellfun(@(x) x==rabouts{cl}(cli,1),rabouts,'UniformOutput', 0));
            end
            clst=sum(index,1)>0; clst(cl)=0; % get index of cells with common links
            if sum(clst)==0 % skip if no neighbours
                continue
            end

            p=1;
            for i=1:length(clst)
                if clst(i)==1
                    clu(p:p+length(rabouts{i})-1,1)=rabouts{i};
                    p=p+length(rabouts{i});
                end
            end
            rabouts{cl}=union(rabouts{cl},clu,'sorted');
        end
       clear cl cli index clst i clu p
       rabouts=uniquecell(rabouts);
    end
    clear x

% combine nodes to single midpoint
```

```
for cl=1:length(rabouts)
    % get unique nodes for each link in cluster
    nf=[]; nt=[];
    for l=1:length(rabouts{1,cl})
        nf(l,1)=roundabout_links.fromNode(rabouts{cl}(l)==roundabout_links.No);
        nt(l,1)=roundabout_links.toNode(rabouts{cl}(l)==roundabout_links.No);
    end
    roundabout_nodes{1,cl}=unique([nf;nt]);

    % get coordinates of nodes in cluster
    for n=1:size(roundabout_nodes{1,cl},1)
        roundabout_nodes{1,cl}(n,2)=nodes.xco(roundabout_nodes{1,cl}(n,1)==nodes.No);
        roundabout_nodes{1,cl}(n,3)=nodes.yco(roundabout_nodes{1,cl}(n,1)==nodes.No);
    end

    % determine midpoint of nodes in cluster
    roundabout_nodes{2,cl}(1,1)=roundabout_nodes{1,cl}(1,1); % first node id is used as midpoint
    roundabout_nodes{2,cl}(1,2)=mean(roundabout_nodes{1,cl}(:,2));% x coor
    roundabout_nodes{2,cl}(1,3)=mean(roundabout_nodes{1,cl}(:,3));% y coor

    % determine change in distance for each node to midpoint
%     for n=1:size(roundabout_nodes{1,cl},1)
%         roundabout_nodes{1,cl}(n,4)=pdist
%     end

end
clear l cl nf nt n

% remove combined nodes
nodes_old=nodes;
nodes.removed=zeros(height(nodes),1);
for nc=1:length(roundabout_nodes)
   nodes.removed(ismember(nodes.No,roundabout_nodes{1,nc}(2:end,1)))=1;
   % update location of midpoint node
   nodes.xco(ismember(nodes.No,roundabout_nodes{1,nc}(1,1)))=roundabout_nodes{2,nc}(1,2);
   nodes.yco(ismember(nodes.No,roundabout_nodes{1,nc}(1,1)))=roundabout_nodes{2,nc}(1,3);

   % add junction info to midpoint node
   nodes.roundabout(ismember(nodes.No,roundabout_nodes{1,nc}(1,1)))=1;
   % add traffic light info to midpoint node
   if sum(nodes.TL(ismember(nodes.No,roundabout_nodes{1,nc}(2:end,1))))>0 % check if one of the nodes had
traffic lights
        nodes.TL(ismember(nodes.No,roundabout_nodes{1,nc}(1,1)))=1;
   end
end
clear nc
nodes(nodes.removed==1,:)=[];nodes.removed=[];

% update links
links_old=links;
for cl=1:length(rabouts)
    links(ismember(links.No,rabouts{cl}),:)=[];% remove clustered links
end
clear cl

for nc=1:length(roundabout_nodes) % change nodes of links to combined cluster node id
   links.fromNode(ismember(links.fromNode,roundabout_nodes{1,nc}(2:end,1)))=roundabout_nodes{1,nc}(1,1);
```

```
    links.toNode(ismember(links.toNode,roundabout_nodes{1,nc}(2:end,1)))=roundabout_nodes{1,nc}(1,1);
end
clear nc
```

Simplify clusters of short links

```
dist=10;
links_s10=links(find(links.length<=dist),:); % very short links
links_s=links(find(links.length<=20),:); % short links

% find clusters of short links
for i=1:height(links_s)
    s=links_s.fromNode(i);
    e=links_s.toNode(i);
    f=find(links_s.fromNode==s | links_s.toNode==e| links_s.fromNode==e | links_s.toNode==s);%find short
neighours
    cluster_links{1,i}=links_s.No(f);
end

clear i e s f
cluster_links=uniquecell(cluster_links);


    % also neighbours of neighbours [3] iterations
    for x=1:3
        for cl=1:length(cluster_links)
            % find clusters with common links
            index=[];clu=[];
            for cli=1:length(cluster_links{cl})
                index(cli,:)=cellfun(@max,cellfun(@(x)
x==cluster_links{cl}(cli,1),cluster_links,'UniformOutput', 0));
            end
            clst=sum(index,1)>0; clst(cl)=0; % get index of cells with common links
            if sum(clst)==0 % skip if no short neighbours
                continue
            end

            p=1;
            for i=1:length(clst)
                if clst(i)==1
                    clu(p:p+length(cluster_links{i})-1,1)=cluster_links{i};
                    p=p+length(cluster_links{i});
                end
            end
            cluster_links{cl}=union(cluster_links{cl},clu,'sorted');
        end
       clear cl cli index clst i clu p
       cluster_links=uniquecell(cluster_links);
    end
    clear x

    % Discard clusters without very short links (length<=10m)
    a=1;
    for cl=1:length(cluster_links)
        if sum(ismember(cluster_links{cl},links_s10.No))==0
            x(a)=cl;
```

```
            a=a+1;
        end
    end
    cluster_links(x)=[];
    clear a x cl links_s10

% combine common nodes of link clusters
for cl=1:length(cluster_links)
    % get unique nodes for each link in cluster
    nf=[]; nt=[];
    for l=1:length(cluster_links{1,cl})
        nf(l,1)=links_s.fromNode(cluster_links{cl}(l)==links_s.No);
        nt(l,1)=links_s.toNode(cluster_links{cl}(l)==links_s.No);
    end
    cluster_nodes{1,cl}=unique([nf;nt]);

    % get coordinates of nodes in cluster
    for n=1:size(cluster_nodes{1,cl},1)
        cluster_nodes{1,cl}(n,2)=nodes.xco(cluster_nodes{1,cl}(n,1)==nodes.No);
        cluster_nodes{1,cl}(n,3)=nodes.yco(cluster_nodes{1,cl}(n,1)==nodes.No);
    end

    % determine midpoint of nodes in cluster
    cluster_nodes{2,cl}(1,1)=cluster_nodes{1,cl}(1,1); % first node id is used as midpoint
    cluster_nodes{2,cl}(1,2)=mean(cluster_nodes{1,cl}(:,2));% x coor
    cluster_nodes{2,cl}(1,3)=mean(cluster_nodes{1,cl}(:,3));% y coor

    % determine change in distance for each node to midpoint
%     for n=1:size(cluster_nodes{1,cl},1)
%         cluster_nodes{1,cl}(n,4)=pdist
%     end

end
clear l cl nf nt n

% remove combined nodes
nodes_old=nodes;
nodes.removed=zeros(height(nodes),1);
for nc=1:length(cluster_nodes)
   nodes.removed(ismember(nodes.No,cluster_nodes{1,nc}(2:end,1)))=1;
   % update location of midpoint node
   nodes.xco(ismember(nodes.No,cluster_nodes{1,nc}(1,1)))=cluster_nodes{2,nc}(1,2);
   nodes.yco(ismember(nodes.No,cluster_nodes{1,nc}(1,1)))=cluster_nodes{2,nc}(1,3);
   % add roundabout info to midpoint node if included
   if sum(nodes.roundabout(ismember(nodes.No,cluster_nodes{1,nc}(2:end,1))))>0
        nodes.roundabout(ismember(nodes.No,cluster_nodes{1,nc}(1,1)))=1;
   end
   % add traffic light info to midpoint node if included
   if sum(nodes.TL(ismember(nodes.No,cluster_nodes{1,nc}(2:end,1))))>0
        nodes.TL(ismember(nodes.No,cluster_nodes{1,nc}(1,1)))=1;
   end
end
clear nc
nodes(nodes.removed==1,:)=[];nodes.removed=[];

% update links
links_old=links;
```

```
for cl=1:length(cluster_links)
    links(ismember(links.No,cluster_links{cl}),:)=[];% remove clustered short links
end
clear cl

for nc=1:length(cluster_nodes) % change nodes of links to combined cluster node id
   links.fromNode(ismember(links.fromNode,cluster_nodes{1,nc}(2:end,1)))=cluster_nodes{1,nc}(1,1);
   links.toNode(ismember(links.toNode,cluster_nodes{1,nc}(2:end,1)))=cluster_nodes{1,nc}(1,1);
end
clear nc
```

Remove parralel links

find links with same start- and end nodes/ start_a=end_b or vice versa

```
while 1 % loop untill break statement is true
    cluster_links_prl={};
    p=1;
    for i=1:height(links)
        s=links.fromNode(i);
        e=links.toNode(i);
        f=find(links.fromNode==s & links.toNode==e | links.fromNode==e & links.toNode==s);
        if length(f)>1
            cluster_links_prl{1,p}=links.No(f);
            p=p+1;
        end
    end
    clear i e s f p
    if ~isempty(cluster_links_prl) % if parallel link clusters exist
        cluster_links_prl=uniquecell(cluster_links_prl);
    else
        break % if no parallel link clusters exist, stop while loop
    end


    % keep cycleway/ combine oneway/ keep shortest
    prl_info={};
    for i=1:length(cluster_links_prl)
        for ii=1:length(cluster_links_prl{i})
            prl_info{1,i}(ii,1)=links.highway(links.No==cluster_links_prl{i}(ii,1));
            prl_info{2,i}(ii,1)=links.length(links.No==cluster_links_prl{i}(ii,1));
            prl_info{3,i}(ii,1)=links.oneway(links.No==cluster_links_prl{i}(ii,1));
        end

        if sum(prl_info{1,i}=='Cycleway')==1 % check if cycleway
            links(ismember(links.No,cluster_links_prl{i}(prl_info{1,i}~='Cycleway')),:)=[];% remove non-cycleway

        elseif sum(prl_info{3,i}=='TRUE')>0  % if at least one one-way--> combine into a two-way street
            if sum(prl_info{2,i}~=min(prl_info{2,i}))==0 %if links have equal length, choose first one
                links(ismember(links.No,cluster_links_prl{i}(2)),:)=[];
                links.oneway(links.No==cluster_links_prl{i}(1))='FALSE'; % make two-way
            else
                links(ismember(links.No,cluster_links_prl{i}(prl_info{2,i}~=min(prl_info{2,i}))),:)=[];% keep
shortest link
                links.oneway(links.No==cluster_links_prl{i}(prl_info{2,i}==min(prl_info{2,i})))='FALSE'; % make
two-way
```

```
            end     % what if same direction one-ways?? ignore?


        elseif  sum(prl_info{3,i}=='TRUE')==0 || sum(prl_info{1,i}=='Cycleway')>1 % check if multiple cycleways
            if sum(prl_info{2,i}~=min(prl_info{2,i}))==0 %if links have equal length, choose first one
                links(ismember(links.No,cluster_links_prl{i}(2:end)),:)=[];
            else
                links(ismember(links.No,cluster_links_prl{i}(prl_info{2,i}~=min(prl_info{2,i}))),:)=[];% keep
shortest link
            end
        end
    end
end
clear i ii
```

Remove looping links (s=e)

```
    links(links.fromNode==links.toNode,:)=[];
% update link numbering
    links.No=[1:height(links)]';
```

Update nodes numbering and link nodes

update osmid's links

```
[a b]=ismember(links.fromNode,nodes.No);    links.from=nodes.osmid(b);
[a b]=ismember(links.toNode,nodes.No);      links.to=nodes.osmid(b);
clear a b
% update node id's in link table
[loc1 links.fromNode]=ismember(links.from,nodes.osmid);[loc1 links.toNode]=ismember(links.to,nodes.osmid);
clear loc1

% update node numbering
nodes.No=(1:height(nodes))';
```

Add edge angles

```
edgesG=links;
nodesG=nodes;

var={'xco','yco'};
for i=1:height(edgesG)
    s=edgesG.fromNode(i);        e=edgesG.toNode(i);
    s_co=nodesG(find(nodesG.No==s),var); e_co=nodesG(find(nodesG.No==e),var);
    dx=e_co.xco-s_co.xco;        dy=e_co.yco-s_co.yco;
    edgesG.angle(i)=atan2d(dy,dx);
end

clear var s e i s_co e_co dx dy
```

save

only Ghent for now!

```
save network.mat nodesG edgesG
```

Visualize in Qgis

```
nodesG1=nodesG; edgesG1=edgesG;
load networkraw.mat
incl=table;
incl.uniqueid=edgesG.uniqueid;
incl.included=ismember(edgesG.uniqueid,edgesG1.uniqueid);
writetable(incl,'./Network/edges_included.csv')
```

Update network grid

```
grid_network();% function
```

Published with MATLAB® R2019a