TECHNICAL UNIVERSITY OF MUNICH

Department of Civil, Geo and Environmental Engineering

Master's Thesis

The Effects of the Built Environment on Bicycle Route Choice

Author: Julianie Charmeil
Master: Transportation Systems

Supervisors: Dr. Ana Moreno, Qin Zhang
Chair: Professorship of Modeling Spatial Mobility

Submission Date: December 3, 2019

Declaration of Authorship

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

Munich, November 28, 2019

Julianie Charmeil

Abstract

Cycling is an environmentally friendly means of transport that reduces pollution, congestion, and noise. However, there is a lack of knowledge about the behavior of bicyclists. In this study, a bicycle route choice model is estimated in the city of Amiens, located in the north of France. GPS data of bicycle trips collected during the European cycling challenge in 2016 and 2017 are used for this purpose. To understand why a route is preferred, other possible route alternatives need to be enumerated for each recorded trip. The main originality of this thesis lies in the choice set generation method. The choice set is composed of the observed routes and several other paths, such as the shortest path and the path maximizing the cycling infrastructures. The different routes are then characterized in terms of cycling infrastructures, road type, land-use, and topography by using different sources of open data. Finally, a multinomial logit model is estimated with 2,362 trips and includes a path-size factor to correct for route overlap. The results indicate that cyclists prefer routes with bicycle facilities and fewer traffic. Routes with few intersections and along water or green areas are especially attractive. The findings can be used for a bicycle traffic assignment and can help transport planners to develop bicycle-friendly cities.

Acknowledgements

I would like to extend my deepest gratitude to Dr Ana Moreno and Qin Zhang who mentored me and supervised my work. Thanks for their trust, their precious guidance and their kindness. My thanks also go to the professorship of Professor Moeckel for allowing me to carry out this thesis and working on this topic. I am especially grateful to the municipality of Amiens Métropole who kindly agreed to share bicycle GPS data with me. It has been a great pleasure to discover the very nice city of Amiens.

Cette master thesis marque la fin de six années d'études. Elles auront été si riches en apprentissage, en rencontres et en épanouissement personnel. Je tiens à remercier tout d'abord mes parents pour la totale confiance qu'ils m'ont toujours accordée et qu'ils continuent de m'accorder malgré mes aspirations pas toujours très conventionnelles. Papa, Maman, merci pour votre dévouement, votre bienveillance et pour cette immense richesse culturelle dont vous nous avez fait le cadeau. Ariane, Timothée, les mots me manquent pour décrire le lien qui nous unit, je suis submergée par votre amour. Mes amis de l'ENSTA, de Stan ou d'ailleurs, merci pour tous ces moments de joies, merci de m'avoir tant fait grandir, de si bien me comprendre.

Finally, I would like to thank my dearest friends of Munich. You have confirmed to me how enriching diversity can be, I am so grateful I met you and looking forward to seeing you again whether in China, Taiwan, United Arab Emirates, Germany or France!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Recognizing the environmental and social benefits of cycling, many countries and cities are actively pursuing a strategy to increase daily bicycle-usage. Cycling can help to reduce noise and air pollution, while avoiding congestion. It also provides many benefits for human health. However, there is a lack of knowledge about the behavior of bicyclists, and particularly about route choice. The traditional transportation models and route planners were firstly created for cars and then adapted to other modes, and travel time is often considered as the only significant parameter for selecting a route. For bicyclists, additional parameters play a crucial role, such as the perceived safety, the cycling infrastructures, the topography or the land-use. Moreover, bicyclists are more willing to take a minor detour to follow a more pleasant route.

However, this behavior is currently neglected by transport planners, primarily due to a lack of data. Thus, there is a significant need to better understand how bicyclists choose their route. This will help improve the bicycle traffic assignment of current transport models. The results could be useful for transport planners to analyze existing cycling conditions and to evaluate likely impacts of future projects, which is essential for the development of a more efficient and sustainable transportation system.

1.2 Research gaps

The emergence of GPS technologies has offered new opportunities to understand real behaviors of cyclists. The first bicycle model using GPS data was estimated by Menghini et al. (2010) in Zurich. Since then, several other authors have developed bicycle route choice models in other cities. However, most of the time the considered attributes are limited to trip length, traffic volume, grade and the presence of bicycle facilities. Few studies have explored the influence of other land-use attributes such as green and water areas, landmarks, or number of amenities.

Moreover, estimating a route choice model requires knowing the different options considered by a traveller to go from an origin to a destination. This is one of the main challenges of route choice modeling. The actual choice set is unknown from the modeler and can be extremely large due to the high number of links in a network. Still under investigation, is a method on how to generate a relevant choice set. Thus, both in the considered attributes and in the methodology, important research gaps remain and leave space for further research.

1.3 Project scope

In this master's thesis, the study of bicycle route choice is performed at a city scale with an emphasis on non-recreational trips. This project focuses on the city of Amiens, located in the north of France in the Somme region and centered at the crossroads of several strategic urban centers: Paris, located 140 km to the south, and Lille, located 180 km to the north (fig. 1.1). The territory was severely affected by the two World Wars and faced important losses and emigration. While Amiens was the 10th largest city in France in 1900, it is now the 27th largest city with 133,800 inhabitants living within 49.5 km^2 (INSEE, 2015). However, Amiens remains an area with a significant influence on a larger scale and is especially famous for its cultural heritage with its gothic cathedral and its traditional houses (Aduga, 2012). Originally well-known for its textile industry and then for its automobile industry, employment is today based on the service sector (healthcare, education, new technologies and logistics). Amiens belongs to a larger group, called Amiens Metropole, composed of 39 municipalities around Amiens that are mainly rural with only 19% of the area comprising artificial spaces. Amiens Metropole extends over 350 km^2 and has 182,600 inhabitants (Aduga, 2012).

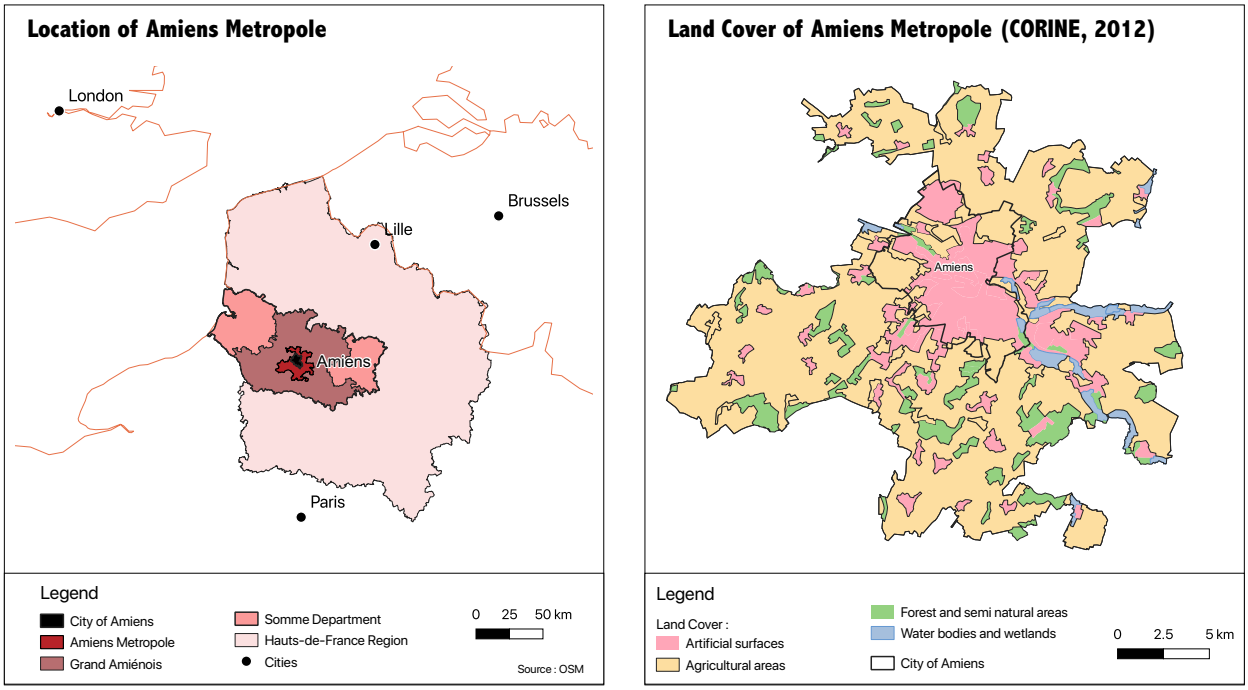

FIGURE 1.1: Background information about Amiens Metropole (Copernicus, 2012)

The choice of Amiens for investigating bicycle route choice was made for several reasons.

Firstly, data revealing route preferences were required to study the factors influencing bicycle route choice. Amiens Metropole had GPS data of bicycle trips, that came from the European Cycling Challenge, a cyclist competition between cities. Amiens took part in this challenge in 2016 and 2017. For one month, people were invited to use their smartphone to record their track. The challenge had two main objectives: encouraging people to cycle more and collecting GPS data for urban planning (CIVITAS, 2017). This enabled the city to record 4,452 bicycle trips. Amiens

Metropole kindly allowed me to work on the anonymized data.

The second reason that explains this choice is that France and especially Amiens are places where cycling conditions are a key issue. The most recent survey from 2010 revealed that only 2% of trips were made by bike in Amiens (Pays du Grand Amiénois, 2013). Moreover, a survey conducted online in France between September and November 2017 by a cyclist association (Fédération française des Usagers de la Bicyclette, 2018) ranked Amiens 20^{th}out of 29 cities with between 100,000 and 200,000 inhabitants. The majority of the respondents believe that current cycling conditions in Amiens are unsatisfactory and that they do not allow cycling in a comfortable and safe way. The average grades obtained for the different categories and the general perception of the cycling conditions can be seen in figure 1.2.

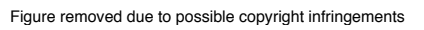

FIGURE 1.2: Survey results about cycling conditions in Amiens (Translated from Fédération française des Usagers de la Bicyclette (2018))

The safety issue is one of the most important problems mentioned in the the survey. The cycling network of Amiens is incomplete and not continuous. In 2012, there were 100 km of cycling facilities, mainly composed of bike lanes, mixed pedestrian/bicycle spaces or bus lanes. Since 2014, Amiens Metropole has worked on a cycling plan aiming to reach 200 km of cycling facilities by 2025. However, the quality of the infrastructures is highly criticized, especially by the cycling association Veloxygène, who regrets the lack of cycle paths separated from the traffic (Véloxygène, 2019). Therefore, it appears interesting to apply our study to this city, insofar as the potential improvements are big if the effects of the built environment on bicycle route choice are better understood.

Thus, investigating the factors affecting bicycle route choice is a main motivation. By identifying why some roads are avoided or preferred by bicyclists, transport planners could work on removing the bottlenecks and design better bicycle networks.

1.4 Objectives and workflow

The objective of this master thesis is to estimate a bicycle route choice model from revealed-preference data in the city of Amiens. The research questions are as follow: what are the factors influencing the bicycle route choice on urban areas? How does the built environment impact this choice?

In order to meet this objective, the recorded routes are first identified by using a geographical information software. Then, the different links are characterized particularly in terms of cycling infrastructures, land-use, topography and traffic volume by using different sources of open data. Lastly, a statistical analysis is performed to compare the characteristics of the selected paths to other possible paths. A discrete choice model is estimated to understand the magnitude of the effects of the various factors affecting bicycle route choice.

In this thesis, first of all, previous studies related to route choice modeling are presented in chapter 2. Then, chapter 3 provides a first analysis of the available data and aims to verify that there are appropriate for this study. Chapter 4 is dedicated to the methodology and all the different stages of the project are summarized: data filtering, map-matching, model preparation and model estimation. The results are presented and discussed in chapter 5. The conclusion in chapter 6 summarizes the main contributions and provides recommendations for further research.

Chapter 2

Literature Review

2.1 Introduction to the route choice problem

The route choice problem consists of identifying the chosen route for a given origin-destination pair.

2.1.1 Two modeling approaches

There are two main approaches to estimate a route choice model: a deterministic or a stochastic one. The first approach, called **deterministic utility maximisation**, is based on shortest path algorithms such as Dijkstra (1959). These algorithms select the sequence of links that minimizes a cost function. Even if the results can guide transportation investments, it cannot be used for prediction purposes as the model will always predict the same path between a given origin and destination. It assumes that all attributes leading to route choice can be identified and that drivers have a perfect knowledge of the link utilities. For example, Kang and Fricker (2018) applied optimization methods to calibrate a deterministic link cost function including distance and an indicator related to safety.

The second approach found in the literature is the use of **random utility models**. A random part in the utility function is added to capture the heterogeneity of the behaviors. These models are called discrete choice models. This chapter focuses on these types of models because they are behaviorally more realistic than deterministic ones.

2.1.2 Specific challenges of route choice modeling

In comparison to traditional discrete choice problems such as mode choice, the route choice has specific features that require appropriate modeling answers.

Firstly, the **choice set is very large** and difficult to identify. Because of the high number of links in a network, the number of possible routes going from an origin to a destination is enormous, or even infinite if loops are allowed. Moreover, all the feasible routes are not considered by travelers. Some alternatives may not be taken into account by drivers due to their preferences or experiences. Several alternatives may also be not perceived as distinct because they overlap (Prato, 2009). Thus, the actual choice set is unknown, and assumptions are required to generate the choice set.

Secondly, **correlation among alternatives** becomes a critical issue concerning route choice modeling. Some paths can overlap on several links. Therefore, their utilities share unobserved attributes, and the assumptions of independence of irrelevant alternatives cannot be made. This prevents the use of the simplest logit

models. A nested logit model is also inappropriate as it does not allow one link to be part of several nests (Frejinger and Bierlaire, 2007).

This chapter will first address the issue of choice set generation in part 2.2, then the model estimation will be considered in part 2.3.

2.2 Choice set generation

Choice set generation methods were highly discussed in the literature and remain an area under investigation. The actual choice set being unknown from the analyst, extracting the feasible alternatives is a main challenge in route choice modeling. Two types of common errors are not generating the observed route or including irrelevant alternatives. This part concerns the choice set generation methods and is decomposed into two parts. Firstly, **path-generation algorithms** are presented and then **new data-driven methods** that ensure that the actual path is included in the choice set are discussed.

2.2.1 Path generation algorithms

a. With explicit enumeration of the paths

Path-enumeration algorithms have been developed to extract the choice set from the network. The classification of the different techniques presented in the following is taken from Prato (2009). Choice set generation can be either deterministic or stochastic.

Deterministic methods

Deterministic path generation methods always generate the same choice set for a given origin-destination. Most of them are based on variations of the shortest path algorithm. The exact extraction of the k-shortest paths according to a generalized link cost function is rarely used because it generates very similar paths (Prato, 2009). Instead, some modifications are applied before searching for the next shortest path to increase the heterogeneity of the results. Three important techniques can be mentioned: link elimination, link penalty, and labeling. These are considered computationally attractive due to the efficiency of shortest path algorithms.

Firstly, the **link elimination** method initially proposed by Azevedo et al. (1993) removes the links (or part of the links) of previously selected paths before searching for the next shortest path The **link penalty** consists of penalizing the links included in previously selected paths (De La Barra et al., 1993). This technique is usually preferred to link elimination because it allows for further use of essential links and maintains continuity in the network. The procedure is repeated until no more new paths are obtained. However, this approach has an important drawback: the definition of the penalty rule is a crucial issue and highly influences the generated choice set. Moreover, as in the link elimination method, the number of paths that must be created is an arbitrary amount and the choice set does not depend on the individual preferences and knowledge of the network. The last common deterministic method based on the shortest path is called the **labeling approach**. It was first introduced by Ben-Akiva et al. (1984), and multiple modified versions of this technique can be found in the literature (Ramming, 2002; Broach et al., 2010). Instead of using a single objective function and multiple iterations, the labeling method defines several objective functions or labels, and one path will be generated for each criterion. This

method is based on the idea that travelers have different objectives when choosing their routes. For example, some drivers may want to minimize travel time, while others may try to avoid intersections. However, the results are highly dependent on how each label is defined, and the choice set is usually relatively small due to limited data and knowledge about preferences (Prato, 2009). Recently, Chen used a principal component analysis to create different cost functions of route alternatives and generates the route choice set according to the labeling approach (Chen et al., 2018).

To conclude, the deterministic path generation models are by far the most discussed and used group in the literature. However, they have significant limitations. The selection of the link elimination or penalization rules, the definition of the thresholds and the creation of the labels are very subjective and depend strongly on the knowledge of the researchers (Prato, 2009).

Stochastic methods

A smaller group of path generation techniques is based on **stochastic methods**. The advantage of these methods is the higher heterogeneity of obtained paths. They are based on the idea that travelers perceive path costs with errors and that different travelers have different perceptions. Variability is introduced in the network attributes (Ramming, 2002) by drawing the link impedances from probability distributions. The shortest path is then introduced in the choice set, and the process is repeated to generate other paths. Bovy and Fiorenzo-Catalano (2007) proposed the doubly stochastic generation function where both the parameters of the utility function and the attributes are stochastic. However, the selection of the probability distribution depends once again on subjective elements (Prato, 2009).

b. Without explicit enumeration of paths

All the models presented just above rely on the assumption that the choice set contains all the paths considered by travelers. However, due to the lack of an objective definition of relevant routes, the correctness of paths choice set cannot be ascertained (Prato, 2009). Due to this, conventional path-based models have important limitations and cannot be used for prediction. Therefore, two main other approaches were developed: a **probabilistic method** and a **link-based approach**.

Probabilistic approach

An alternative approach, known as the **probabilistic approach**, assumes that the true choice set is the universal choice set. A probability of being included in the choice set is attached to each route. However, a full probabilistic approach is very complex and not suitable for real-size application due to the high number of potential choices (Frejinger et al., 2009). A simplified approach is proposed by Cascetta and Papola (2001). A continuous variable representing the **availability/perception of an alternative** is introduced in the utility function of the route choice model. By doing so, the path-based model is restricted to more feasible routes. However, this model fails to develop a membership probability that depends on attitudinal and perception variables (Prato, 2009).

Frejinger et al. (2009) introduce a **sampling approach** from the universal choice set to overcome the impossibility to generate all the paths explicitly. The probability of selecting a link is calculated based on its distance from the shortest path. A subset

of the paths is generated by random walk, and a sampling correction is then applied to the paths to obtain unbiased parameter estimates.

A link-based approach

More recently, Fosgerau et al. (2013) proposed a new link-based approach. The underlying assumption is that travelers choose the next link at each node recursively. In other words, the route choice is modeled as a sequence of link choices. By doing so, it avoids enumerating paths before estimation, and it does not require any sampling of paths. The basic concept is the following: A traveler chooses the next link recursively that maximizes the sum of instantaneous utility and expected downstream utility at each node. The expected downstream utility refers to the expected utility for reaching the destination. Theoretically, Fosgerau et al. (2013) have shown that the recursive logit model (RL) is equivalent to a path-based model with unrestricted choice set.

The issue of this model is that it requires the calculation of huge (and often-ill defined) inverse matrices for each OD pair. The model can lead to **unstable estimations** if the network is large. Consequently, several modifications were proposed to reduce the computational complexity of the model. Mai et al. (2018) developed a **decomposition method** to evaluate multiple destinations more efficiently. Only one system needs to be solved for all destinations to evaluate path-choice probabilities instead of one system per destination. However, the **computational complexity** remains problematic to apply this model to large-scale problems. Kaneko et al. (2018) introduced the implicit availability/perception (IAP) factor, as proposed by Cascetta and Papola (2001), into the recursive logit model to restrict the universal choice set to more feasible alternatives. Mitigate results were obtained. The model including an awareness term does not work well for travelers that are familiar with the network and for short-trips.

2.2.2 New data-driven methods

None of the path generation algorithms succeed in including all the chosen routes and many routes that are in practice not considered by travelers are often part of the choice set. Moreover, these methods highly rely on the characteristics of the network. In the bicycle context, errors due to the bad quality of the network representation though, are very common. Bicyclists often use short-cuts or cycle against one-way streets, but this type of information is usually not included in the model. Recently, new data driven approaches were introduced to overcome these limitations and create more realistic choice sets (Ton et al., 2017; Ton et al., 2018; Bernardi et al., 2018).

Ton et al. (2017) proposed to include in the choice set all the chosen routes that were observed in the collected data. One requirement for this empirical method is that each OD pair contains multiple trips and more than two distinct routes. A clustering method was applied to aggregate similar OD pairs and increase the number of alternatives in the choice set for a given OD pair. However, the estimation model has shown a worse fit than the methods based on link elimination and labeling because of a lower variability between routes and their attributes (Ton et al., 2018). In parallel, Bernardi et al. (2018) developed an other data-driven approach for the choice set generation. In order to introduce a higher variability between the alternatives, similar routes are grouped together to constitute a unique alternative. This results in a lower number of alternatives but more differences between the alternatives. In

a first approach, Bernardi et al. (2018) used the trip length to create the different categories but a clustering analysis on multiple parameters is suggested by the authors for further research. The attributes of the different groups are calculated by taking the average of the attributes for all the trips belonging to the same category. For each OD pair, the choice set is composed of five alternatives: four coming from the aggregation of similar observed trips, and the shortest path.

To conclude, these two data-driven approaches used to generate the choice set allow the creation of a more realistic choice set. They provide interesting insights about travel behaviors and factors affecting route choice. However, several important limitations must be mentioned. Firstly, the choice set **depends on the data sample**. In other words, another set of data will generate another choice set. This issue may be reduced by using a higher sample size and a longer data collection period. It may introduce more variability in the collected routes collected leading to a better model estimation (Ton et al., 2018). Lastly, none of these models can be used for prediction purposes. Bernardi et al. (2018) relies on a high level of aggregation so that the model is only relevant for understanding behaviors, while the model of Ton et al. (2017) does not perform well out-of-sample.

Conclusion: None of the existing methods completely overcome the problem of choice set generation and this remains an area under investigation. Figure 2.1 summarizes the different methods found in the literature. Deterministic path generation methods with explicit enumeration of the routes are the most used group because of their efficiency. Among them, labeling approaches are especially popular. However, with the increase of available data, new data-driven methods could become new promising ways of finding the possible routes to go from an origin to a destination.

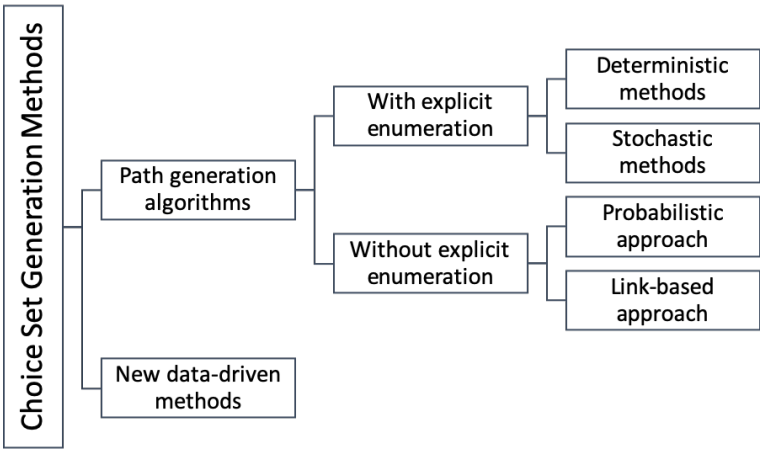

FIGURE 2.1: Summary of choice set generation methods

2.3 Route choice model estimation

In addition to the difficulty of enumerating the choice set, the **problem of overlapping paths between alternatives** is a main challenge in route choice modeling. For routes that have many links in common, the error terms cannot be considered independent and the independence of irrelevant alternatives (IIA) assumption is not valid. As for the blue/red bus paradox, applying a simple multinomial logit model (MNL) structure overestimates the probability of choosing similar alternatives, i.e. similar routes. A nested logit model is also inappropriate as it does not allow one link to be part of several nests (Frejinger and Bierlaire, 2007). To take correlations among the alternatives into account, two main approaches are found in the literature. The first possibility, which is the most often used because of its simplicity, consists of introducing a term in the deterministic part of the path utility that captures the similarities with the other paths of the choice set (part. 2.3.1). The second option, more complex, aims to capture explicitly the correlation through assumptions about the error terms (part. 2.3.2).

2.3.1 Deterministic correlation in a multinomial logit model

Modifications in the multinomial logit models were proposed to release the IIA assumption. The basic multinomial logit model is shown in equation 2.4.

$$P_k = \frac{\exp(V_k)}{\sum_{i \in C} PS_i \exp(V_i)} \tag{2.1}$$

where P_k is the probability of choosing route k, C is the choice set of paths, and V_k and V_i are the deterministic utilities of routes k and i, respectively. V_k can be written as $\beta.X$, where X is a vector of route attributes and β a vector of coefficients to be estimated.

a. C-logit model

Cascetta et al. (1996) were the first to propose a modification in the multinomial logit model. They introduced a **commonality factor** that measures the **degree of similarity** of each route with the other routes of the choice set. The correction is made within the **deterministic part** of the path utilities. The new expression for the probability of choosing route k within the choice set C is:

$$P_k = \frac{\exp(V_k + \beta_{CF} CF_k)}{\sum_{i \in C} \exp(V_i + \beta_{CF} CF_i)} \tag{2.2}$$

where CF_k is the commonality factor and β_{CF} is a parameter to be estimated.

Different formulations of the commonality factor (CF) were proposed in the literature. One possible specification is shown in equation 2.3.

$$CF_{kn} = \ln \sum_{i \in C} \left(\frac{L_{ki}}{\sqrt{L_k L_i}}\right)^{\gamma} \tag{2.3}$$

where L_{ki} is the length of links common to routes k and i; L_k and L_i are the overall length of routes i and j, respectively. The parameter γ may be estimated or constrained to 1 or 2. Thus, here, the commonality factor value depends on the common length between other routes within the choice set. The estimated parameter β_{CF} should be negative to reduce the utility of routes that share many links with the other paths of the choice set.

Other possible specifications of the commonality factor are presented in Cascetta et al. (1996). However, Cascetta et al. (1996) did not provide any guidance to select the most appropriate formulation of the commonality factor.

b. Path-size logit

Another modification of the multinomial logit model was suggested by Ben-Akiva and Bierlaire (1999). Similarly to the commonality factor, they introduced a **path size variable** in the utility of the path. The path size variable represents the **fraction of the path that constitutes a full alternative**. The contribution of a link is reduced according to the number of paths that share the link. The probability of choosing route k is

$$P_k = \frac{\exp(V_k + \ln PS_k)}{\sum_{i \in C} PS_i \exp(V_i + \ln PS_i)} \tag{2.4}$$

where PS_k is the path-size factor defined by

$$PS_k = \sum_{a \in \Gamma_k} \frac{l_a}{L_k} \frac{1}{\sum_{i \in C} \delta_{ai} \frac{L_C^*}{L_i}} \tag{2.5}$$

with Γ_k the set of links in route k, l_a length of link a, L_k length of route k and L_C^* the length of the shortest path in C. δ_{ai} is one if link a is part of path i and zero otherwise.

Ramming (2002) developed another formulation of the path size variable, called **generalized path-size**. The aim was to **reduce the influence of very long paths** on the utility of shorter, more reasonable paths.

$$PS_k = \sum_{a \in \Gamma_k} \frac{l_a}{L_k} \frac{1}{\sum_{i \in C_n} \delta_{ai} (\frac{L_k}{L_i})^\gamma} \tag{2.6}$$

where γ is a parameter greater or equal to zero. However, the model requires the estimation of more parameters and is computationally more costly.

The lack of theoretical guidance for the expression of the commonality factor and the better results found for the path-size logit model (Ramming, 2002) probably explains why the **path-size logit model** became more popular in the following route choice models estimation in the literature. These models that keep the logit structure are certainly easier to estimate but they are unable to capture all the correlation between the alternatives.

2.3.2 Explicit modeling of the correlation

Another approach explored in the literature is the use of models that capture explicitly the correlation in the error terms. However, due to their complexity, few of them have been applied for real size problems.

The **cross-nested logit model**, firstly applied by Vovsha and Bekhor (1998), is an extension of the nested logit model. In contrary to the nested logit model, links can be part of multiple nests. However, it is difficult to estimate because of the large number of nesting coefficients. The **multinomial probit model** (Yai et al., 1997) can also be used. However, this model lacks an analytical formulation of the probabilities and is therefore complicated to estimate. The **error component model** has also been proposed in the literature (Frejinger and Bierlaire, 2007).

Conclusion: Figure 2.2 summarizes the different options used in the literature to model route choice. Despite its drawbacks, path-size multinomial logit model is by far the most coomon model in the literature because of its computational efficiency and its simple structure.

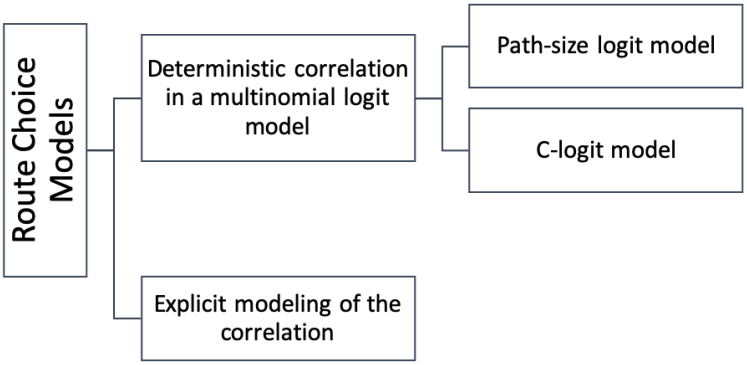

FIGURE 2.2: Summary of route choice models

2.4 Bicycle route choice

While the previous sections reviewed how route choice has been modeled in the literature, this section focuses more specifically on the use of these models in a cycling context.

2.4.1 Data collection

a. Stated-preference survey

The data used to study the factors influencing bicycle route choice have evolved over time.

First studies were based on stated-preference surveys. In many studies, people were put in hypothetical choice situations and were asked to choose a route between several options based on their main characteristics. As mentioned by Sener et al. (2009), most of these studies rely on a descriptive analysis of the collected data, and few multinomial logit or regression analysis methods were applied (Hunt and Abraham, 2007; Sener et al., 2009).

One of the advantages of stated-preference data is the relatively low effort required to collect and analyze the data. Notably, a generation of relevant alternatives is not necessary. Moreover, non-existing or rare facilities can be evaluated, and a large variety of attributes can be investigated. Hunt and Abraham (2007), Sener

et al. (2009), and Casello and Usyukov (2014) offered a detailed review of the considered attributes in the bicycle context. Sener et al. (2009) classified the attributes into six categories : individual characteristics, on-street parking, bicycle facility type, roadway physical characteristics (pavement surface, grade, number of stop signs, red lights, etc.), roadway functional characteristics (traffic volume, speed, etc.), and roadway operational characteristics (distance, travel time, etc.).

However, the difference between claimed and observed behavior is a major limitation of the stated-preference studies. It is difficult for the respondents to fully imagine the situation described. Stated-preference studies can give interesting insights about important factors affecting bicycle route choice but fail to represent behaviors accurately.

b. Revealed-preference data

Due to the important limitations of stated-preference surveys, the data collection process has evolved toward revealed preference studies. The objective is to represent more accurately the behavior by studying the actual chosen routes. Aultman-Hall et al. (1997) were one of the pioneers in using geographic information system (GIS) for bicycle route choice analysis. People were asked to draw their routes, and they were then compared to the shortest paths.

Recently, the emergence of GPS technology has offered new opportunities for data collection. The first model that used GPS data of bicycle trip was developed by Menghini et al. (2010) and applied to the city of Zurich. A link-elimination method was used to extract the alternatives. Results of the analysis suggest that cyclists are sensitive to trip length, grade and presence of cycling facilities.

2.4.2 Bicycle route choice models based on GPS data

After the model of Menghini et al. (2010), many other studies based on GPS data followed. They mainly differ by the location of the case study, the method used to generate the alternatives and the considered attributes.

Traditional path generation techniques presented in part 2.2, such as link elimination, link penalty, labeling, and stochastic path search, have been applied to the bicycle context. In Hood et al. (2011), the model is based on a doubly stochastic shortest path technique. Link attributes and generalized cost coefficients were randomized before extracting the next shortest path. The case study is San Francisco (USA), and socio-demographic and contextual attributes were included as well, such as gender, age, rain, sunset and sunrise times, etc. However, most of them were not significant, only the cycling frequency showed good results. Then, Broach et al. (2012) estimated a model with a calibrated labeling approach and distinguished themself by the large amount of road network attributes. The number of stop signs, of intersections are examples of new parameters that were included. Recently, Chen et al. (2018) also took advantage of a labeling method to extract possible other routes in Seattle network (USA). A principle component analysis was performed to define the objective functions for the labels. This aimed to increase the number of alternatives. Moreover, Chen et al. (2018) were the first to include a large variety of land-use attributes, such as city feature density, proportion of water and parks, or density of street trees. Unlike the other studies mentioned above, the model was not estimated with a multinomial logit formulation and a path-size factor but had a mixed logit structure with a path-size.

Alternative methods to generate the choice set have also been used for bicycle route choice models. Zimmermann et al. (2017) used the recursive logit model (RL) and the nested recursive logit model (NRL) that do not require enumerating alternatives. Finally, Ton et al. (2017) and Bernardi et al. (2018) employed a data-driven approach and considered choice sets with the observed routes.

Table 2.1 reviews the main bicycle route choice models found in the literature and presents the attributes considered. Several comments can be made about the findings of these studies. Trip length was an important parameter in most of the studies, with a negative sign showing that bicyclists are discouraged by long routes. Bicyclists prefer routes with bike facilities and avoid streets with a lot of traffic. Routes with important slopes are less attractive and bicyclists are ready to take a detour to reduce the number of bridges, intersections, traffic signals, and turning movements. Socio-demographic and contextual were not found significant (except the cycling frequency in Hood et al. (2011) and morning peak hours in Ton et al. (2017)). Finally, very few studies investigated the impact of land-use on bicycle route choice.

Conclusion: Thus, in the literature, several bicycle route choice models were estimated by using different techniques to generate the choice set. Several important research gaps remain. Few studies have considered a rich set of land-use attributes and research on how to extract relevant choice sets from the network must be pursued.

Attributes	Menghini et al., 2010	Hood et al., 2011	Broach et al., 2012	Casello and Usyukov, 2014	Zimmermann et al., 2017	Ton et al., 2017	Chen et al., 2018	Bernardi et al., 2018
Road network attributes								
Type of facility (proportion of bike lane, bike path, etc.)	✓	✓	✓	✓	✓	✓	✓	✓*
Gradient	✓	✓	✓	✓	✓		✓	
Traffic volume	✓	✓*	✓	✓	✓			✓*
Number of traffic lights	✓		✓					✓*
Turning movements		✓	✓		✓			
Speed limit		✓*		✓			✓	
Presence of bridge			✓		✓			
Number of intersections						✓		
Number of stop signs			✓					
Number of lanes		✓*						
Street lights							✓	
Trip characteristics								
Trip purpose		✓*	✓			✓		✓
Trip length	✓	✓	✓	✓	✓	✓	✓	
Cycling speed	✓						✓	
Travel time							✓	
Multimodal trip						✓		
Socio-demographic attributes								
Gender		✓*					✓*	✓*
Age		✓*					✓*	✓*
Cycling frequency		✓					✓*	
Contextual attributes								
Rain		✓*				✓*		
Sunset and sunrise times		✓*				✓*		
Crime rate		✓*						
Morning peak hours						✓		
Land-use attributes								
Land use mixture							✓	
City feature density							✓	
Proportion of water and parks							✓	
Proportion of beautiful links								✓
Average floor area ratio							✓	
Density of street trees							✓	

TABLE 2.1: Review of bicycle route choice model attributes (* : factor not significant)

Author	Case-study	Data Origin	Data Collection Period	Number of collected bicycle trips	Choice set generation	Model Estimation
Menghini et al., 2010	Zurich (Switzerland)	GPS device	2004	3 387	Breadth-first search link elimination	Multinomial logit model (MNL) with path-size (PS)
Hood et al., 2011	San Francisco, CA (USA)	App : CycleTracks	Nov- April 2010	2 777	Doubly stochastic shortest path search	MNL with PS
Broach et al., 2012	San Francisco, CA (USA)	GPS device	March -Nov 2007	449	Calibrated labeling method	MNL with PS
Casello and Usyukov, 2014	Waterloo, ON (Canada)	GPS device	February 2010-March 2011	724	Link penalty method	MNL with PS
Zimmermann et al., 2017	Eugene, OR (USA)	App : CycleLane	-	648	-	Recursive logit model + nested recursive logit model with and without PS
Ton et al., 2017	Amsterdam (Netherlands)	App : CyclePrint	September 14–20, 2015	3,045	Chosen alternatives	MNL with PS
Chen et al., 2018	Seattle, WA (USA)	App : CycleTracks	2009-2014	3 310	Labeling method	Mixed logit models with PS
Bernardi et al., 2018	Netherlands (country-wide)	App : MoveSmarter	April-May 2014	3 500	Chosen alternatives + Shortest path	Mixed logit model with PS

Chapter 3

Data Description

3.1 Data source

3.1.1 Data collection

To study the factors influencing bicycle route choice, a sufficient amount of data in a given location was required. As shown in the literature review, **GPS data** offer a huge opportunity for observing real behaviors. Collecting data would have considerably limited the sample size and time remaining for model estimation. However, few bicycle routes can be found in open access. Some traces are shared by bicyclists via cycling smartphone applications but are most of the time round trips for recreational purposes. As the objective was to study route choice for utilitarian trips, this data source was not appropriate.

The second idea was to identify events that could have led to a data collection process. The **European cycling challenge** was one of them. It is an urban cyclists´ team competition to encourage people to cycle more. The challenge was initially created by the city of Bologna in 2012 and took place every year in May from 2012 to 2017. Citizens could record their bicycle trips with a GPS-based smartphone App and help their city to collect kilometers. In 2016, 52 cities from 17 countries joined the game and four million kilometers were cycled in a month by 46,000 participants (CIVITAS, 2017). The challenge had two main objectives:

- Encouraging people to change their mobility behaviors towards a more sustainable mode
- Collecting GPS data for urban planning

In France, four cities have ever taken part in this challenge: Lille, Nantes, Amiens, and Louviers. These French municipalities were contacted. In parallel, cycling apps and cyclist associations were also approached. For privacy reasons, several organizations were reluctant to share even anonymized data but finally, the municipality of **Amiens** kindly accepted the use of their data for a research purpose.

3.1.2 Description of the collected data

Amiens Metropole took part three times in the European cycling challenge from 2015 to 2017. However, the municipality had only access to the results for 2016 and 2017. Table 4.1 shows the amount of data collected in Amiens. They are anonymous and organized into two files (GENERIC and DETAIL file), written in a csv format. The GENERIC file includes one line per trip with information about travel distance, travel time, average speed and sometimes socio-demographic characteristics of the participant. The DETAIL file contains the GPS data. It consists of one point per

line defined by its latitude, longitude, altitude and the time stamp. Thus, a route is represented by several points with the same ID.

	2016	**2017**
Number of trips recorded	2,106	2,346
Number of GPS points recorded	424,910	904,274

TABLE 3.1: Amount of data collected

Between 2016 and 2017, the mobile application used to collect data was changed (Cycling 365 in 2016, Naviki in 2017) and there are some variations in the variables and the way they are coded. Table 3.2 and 3.3 summarize the variables included in the GENERIC and DETAIL files of 2016 and 2017.

Name	**Definition**	**Type in 2016**	**Type in 2017**
TripID	*ID of the trip*	Character	Integer
TimeStamp	*Starting time in seconds from 01.01.1970*	Integer	Integer
StartDate	*Date of the trip*	2016-05-DD	05.DD.17
StartTime	*Starting time of the trip*	THH:MM:SS.000Z	HH:MM:SS
Duration	*Trip duration in seconds*	Integer	-
AvgSpeed	*Average speed in km/h*	Floating	Floating
MaxSpeed	*Maximum speed in km/h*	-	Integer
Distance/TotalLength	*Total length of the trip in km*	Floating	Floating
TrackType	*Type of track*	Urban bicycle or other	-
Sex	*Gender*	M or F	male or female
Year/YearOfBirth	*Year of birth*	Manually written	MM.DD.YY
Frequent User	*Frequent use*	Boolean	-
Profession	*Job*	Manually written	-
ZIP	*Zip Code of the participant*	Integer	-
TypeOfBike	*Type of the bike*	MyBike	owned bike
		MyEBike	owned e-bike
		BikeSharing	shared bike
		EBikeSharing	shared e-bike
		–	n/a
TypeOfTrip	*Trip purpose*	HomeToWork	home-to-work
		HomeToSchool	home-to-school
		Leisure	leisure
		Other	other
		–	n/a
Source/Uploaded	*How the trip was tracked*	cy-mobile (Cycling365)	No (Naviki)
		cy-wep-gpx (as gpx)	Yes
		cy-web-manually (manually)	

TABLE 3.2: Variables of the GENERIC table of 2016 and 2017

Name	Definition
TripID	ID of the trip
TimeStamp	Starting time in seconds calculated from 01.01.1970
Latitude	In degrees
Longitude	In degrees
Distance	Distance to the previous point
Altitude	In meters
Speed	In km/h
Type	start (if first point), mid or end (last point)

TABLE 3.3: Variables of the DETAIL table of 2016 and 2017

3.2 Representativeness of the sample

In this part, the socio-demographic data are analyzed and compared to the 2010 mobility survey *Enquête déplacements grand territoire réalisée dans le Grand Amiénois en 2009-2010* (EDGT, 2010) realized in Amiens Metropole (Pays du Grand Amiénois, 2013). This aims to examine the extent to which the sample collected during the challenge is representative of the cycling population. However, the results should be taken with some caution: first of all, these questions were optional and not all the participants in the challenge answered to them (table 3.4) and secondly, the most recent travel survey was in 2010 and the situation has probably changed since that time. It is also important to mention that it is not possible to identify which trips were made by the same person as the data collected are anonymous. Therefore, all trips are considered independently.

Variables	2016	2017
Gender	75%	44%
Age	82%	38%
Trip Purpose	53%	50%
Type of Bikes	53%	50%

TABLE 3.4: Percentage of answers to the optional questions

The attributes that are investigated in the following are gender, age, trip purpose, trip length and departure time.

3.2.1 Gender: a consistent overrepresentation of male among participants

Firstly, the gender of the participants was analyzed. More than 62% of the trips collected during the challenges of 2016 and 2017 were made by men. A similar trend was observed in the Amiens Metropole mobility survey of 2010. A comparison of gender distributions is presented in figure 3.1. Thus, gender can be considered representative. This overrepresentation of men among cyclists is common in other low-cycling countries, such as the UK, the USA, and Canada. On the contrary, in higher cycling countries, gender differences tend to be much less marked or even reversed. For example, in Netherlands or Denmark, more than 45% of cyclists are women (Aldred et al., 2016).

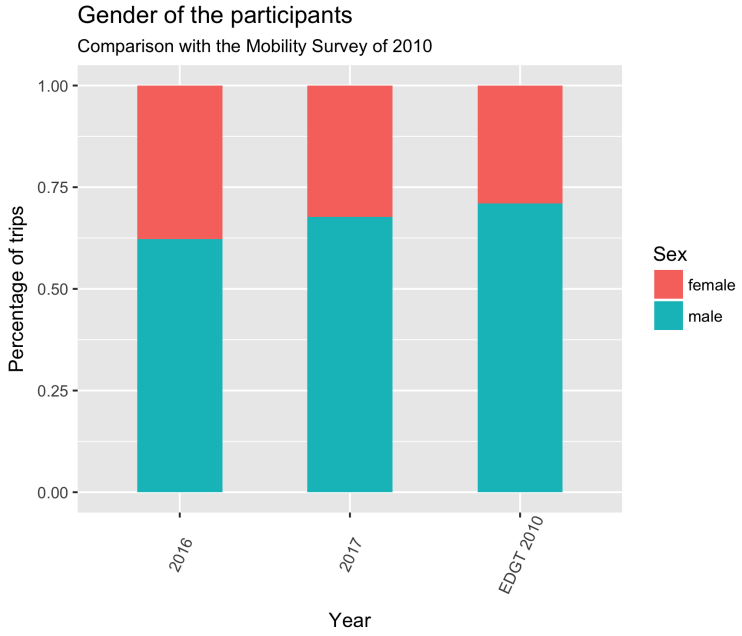

FIGURE 3.1: Gender Distribution

3.2.2 Age: a majority of 25-49 years old

Concerning the age distribution, important differences can be observed as shown in figure 3.2.

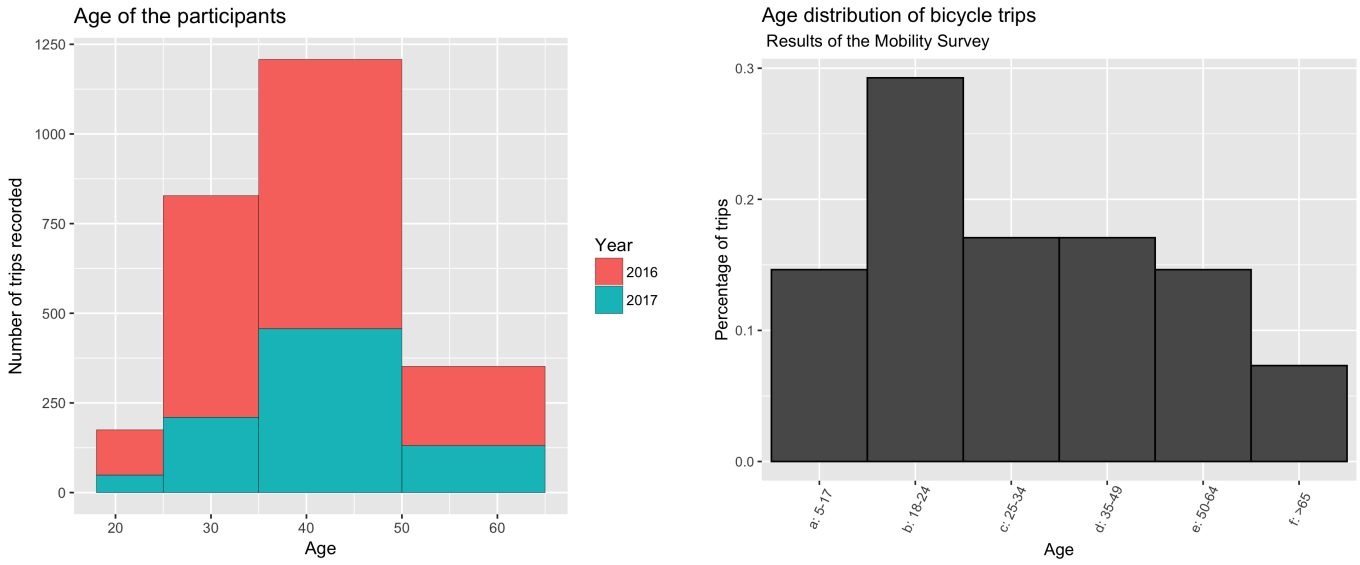

FIGURE 3.2: Age distribution

While in the mobility survey, the 18-24 years old are by far the most represented, they are almost absent among the participants of the cycling challenge. The 25-49 age group was the most involved (69%). This result is comparable to the one observed during the European cycling challenge in Lille in 2016, where 70% were between 25 and 49 years old (Drouaults, 2016). Thus, it was above all workers that took part in the challenge. This could be due to a positive emulation in companies that encouraged people to contribute to the challenge.

3.2.3 Trip purpose: a majority of home-work trips

Trip purposes of the GPS data collected during the cycling challenge are presented in figure 3.3.

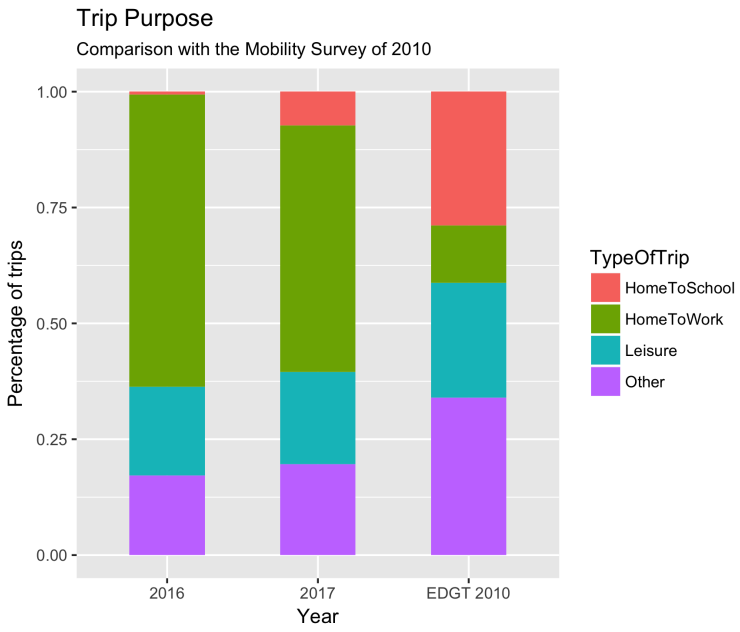

FIGURE 3.3: Trip purpose

Nearly 60% of the trips collected during the 2016 and 2017 challenges are home-work trips, whereas this category was the least represented in the travel survey (21%). Moreover, there is an under-representation of trips to school (4% for both challenges against 25% in the mobility survey). These results are in line with the age distribution observed. Finally, the proportion of leisure trips and other trip purposes is much smaller than in the mobility survey (38% against 55%).

3.2.4 Time of the trip: a typical time distribution

Information concerning the dates of the trips and the departure time confirms the high proportion of home-work trips. The temporal distribution of the trips recorded during the challenges is very similar to the one obtained by the mobility survey (fig. 3.4) and is typical of commuting trips. Three peaks can be observed. A first one very concentrated between 8am and 9am, a second one smaller around 1pm and a last one wider around 5pm. Concerning the trip dates, a big distinction between week-ends and week-days is observed in figure 3.5. The number of trips was much higher during the weekdays than the weekends. There are two public holidays in France in May: on the 1st and the 8th of May, during which similar trends to weekends can be observed.

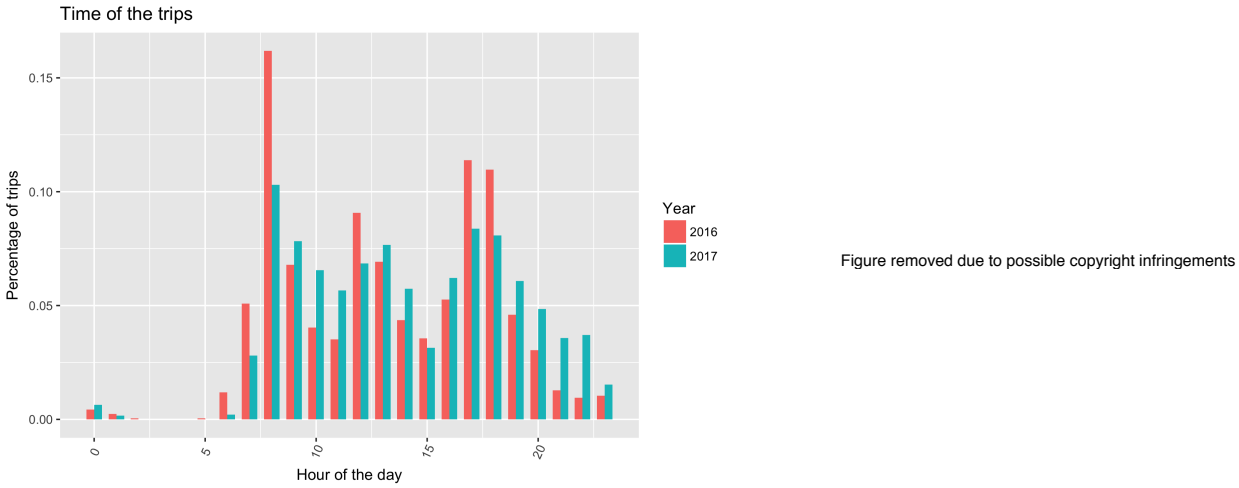

FIGURE 3.4: Temporal distribution

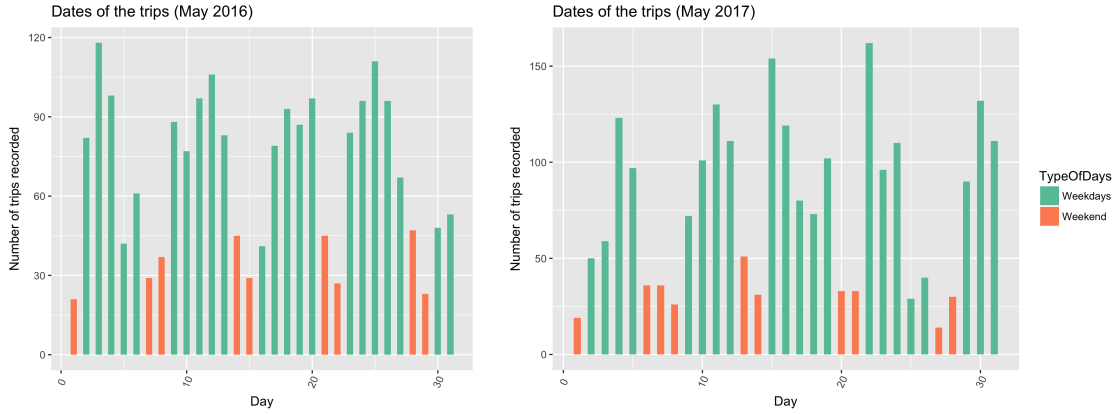

FIGURE 3.5: Trip dates

3.2.5 Trip length: a representative travel distance distribution

Finally, the observed travel distance distribution was analyzed in figure 3.6. The results are affected by outliers and there is a significant difference between the mean (4.1 km) and the median (2.9 km). The average distance is close to the one obtained in the mobility survey (3.9 km). Therefore, it seems that the participants did not try to increase their travel distance to collect more kilometers for their city. This is an important result for the rest of the study on route choice.

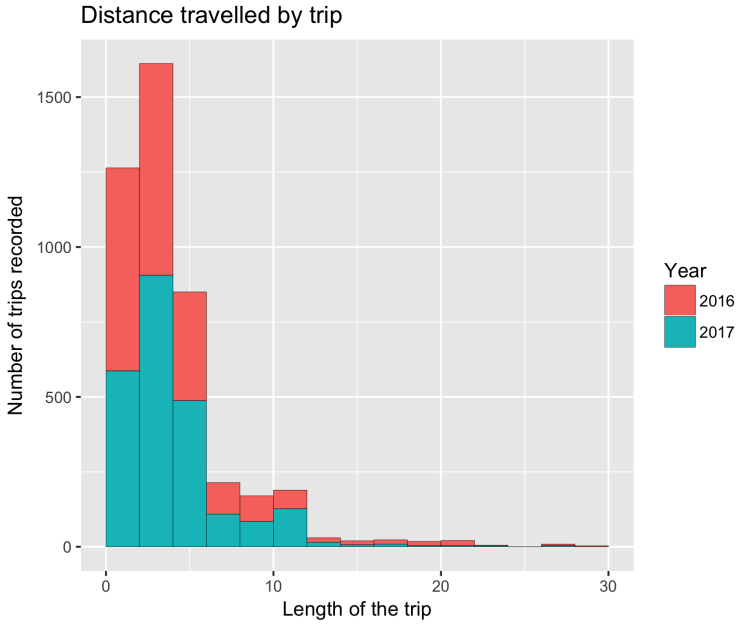

FIGURE 3.6: Travel distance distribution obtained during the challenge

Conclusion : Thus, the analysis of the dataset shows a majority of home work-trips realized by people between 25-49 years old. This overrepresentation of this group compared to the mobility survey of 2010 is not problematic because it is consistent with the study objectives consisting of analyzing non-recreational trips. Therefore, the dataset is suitable for our analysis. Moreover, due to a representative trip length distribution, the following hypothesis can reasonably be made: the participants of the challenge have similar behavior in terms of route choice than they will have without recording their data.

3.3 Spatial analysis of the data

3.3.1 Spatial distribution of the collected data

After comparing the socio-demographic answers, a spatial analysis was performed on the data. The spatial resolution depends on the year of the challenge. While in 2016 points were recorded every 5 seconds, in 2017 the new application allowed to decrease the time step to one second. Figure 3.7 shows the density of GPS points, while figure 3.8.a is a flow map showing the origin-destination of the recorded trips. In figure 3.8.b, the flow is aggregated by districts and only arrows representing more than 50 trips are displayed. It can be observed that the trips are highly concentrated inside the city of Amiens and above all between three districts: Amiens Center, Amiens South-West and Amiens South-East.

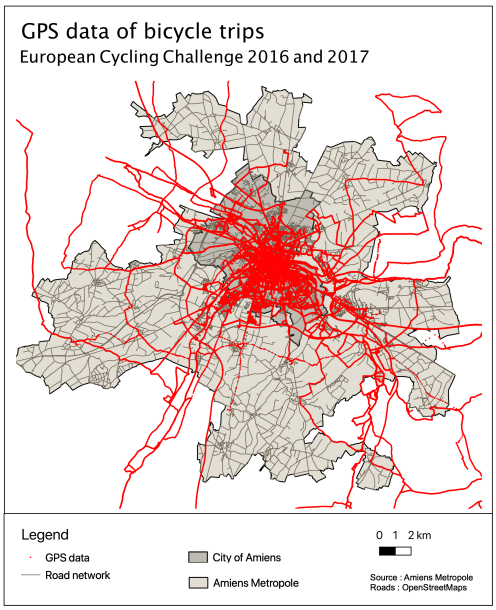

FIGURE 3.7: Density of GPS points

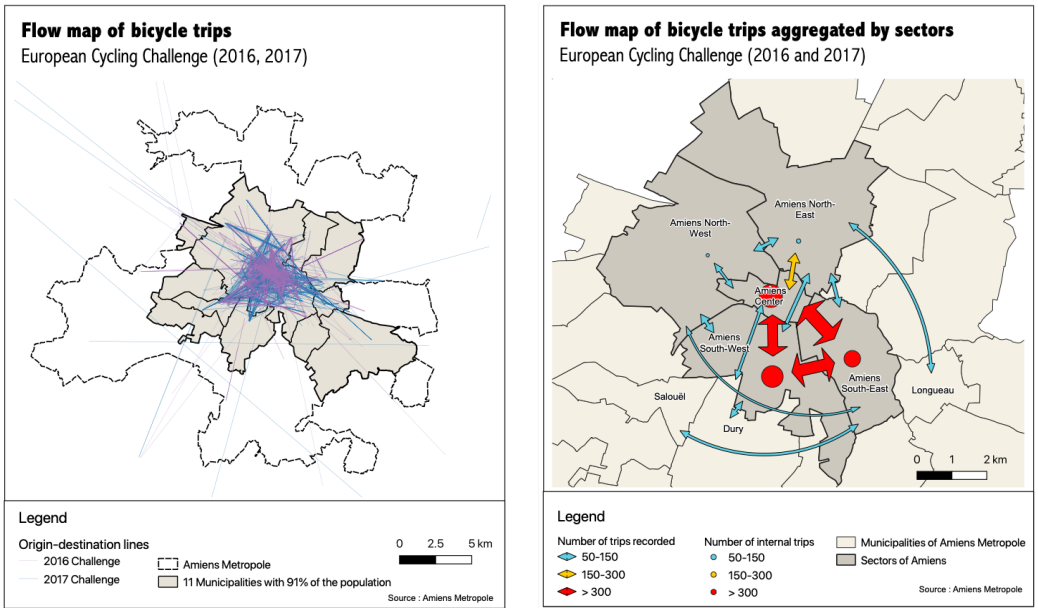

FIGURE 3.8: Flow maps of recorded trips (a: all trips, b: aggregated trips)

3.3.2 Comparison with the features of Amiens

These results are in line with the concentration of population of Amiens Metropole within the city, as it can be seen in figure 3.9. Amiens Metropole is a very contrasted territory characterized by a high concentration of the population (74%), jobs (84%) and equipment inside the city of Amiens. The urban community is structured by a central area of 10 municipalities representing 91% of the population of Amiens Metropole. The other 28 municipalities are less densely populated and only four of them have more than 1 000 inhabitants (INSEE, 2015).

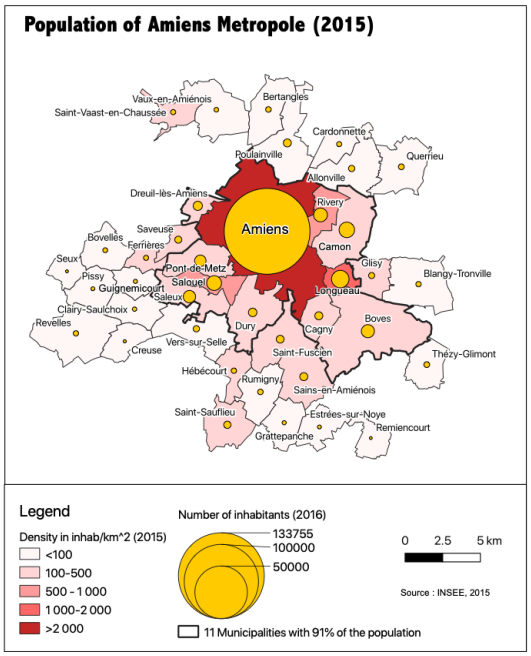

FIGURE 3.9: Population of Amiens Metropole (INSEE, 2015)

Conclusion: Thus, the data collected during the European cycling challenge are consistent with the spatial distribution of Amiens Metropole.

Chapter 4

Methodology

4.1 Introduction to the methodology

As seen in the literature review, estimating a route choice model especially requires two key decisions: the way the choice set is generated and the way the model corrects for overlaps between routes. For this study, a data-driven approach inspired by Ton et al. (2017) was combined with a labeling method (Ben-Akiva et al., 1984) to generate possible alternatives. For the model estimation, a path-size logit structure was selected as proposed by Ben-Akiva and Bierlaire (1999). The main originality of the methodology lies in the choice set generation method. A data-driven method based on the observed routes is a very promising approach because of the increase amount of available data. However, as mentioned by Ton et al. (2017), a low variability of collected routes can be obtained that can lead to worse fits than other generation techniques. To overcome this issue, this thesis combines the data-driven approach with the labeling one. In other words, in addition to the observed routes, for each origin-destination (OD) pair, the choice set contains several other routes, such as the shortest path and the path maximizing the proportion of cycling facilities.

To apply this methodology, the steps summarized in figure 4.1 were performed. Firstly, the trips were filtered to eliminate irrelevant trips and they were clustered to increase the number of possible routes for each origin-destination pair (part 4.2). Trips with very close origins and destinations were grouped together in the same cluster. Thus, one cluster contains possible routes to go from a given origin to a given destination. Then, the GPS points were matched to identify the succession of links that were taken by the bicyclists (part 4.3). After that, the different routes belonging to the same cluster could be studied in detail and the number of different routes per cluster was calculated. To increase the cluster size, some additional routes were added in part 4.4 by using the labeling approach. In part 4.5, a large variety of route attributes were defined based on the city characteristics and the literature and calculated for each trip. Finally, a discrete choice model was estimated with a path-size logit formulation (part 4.6).

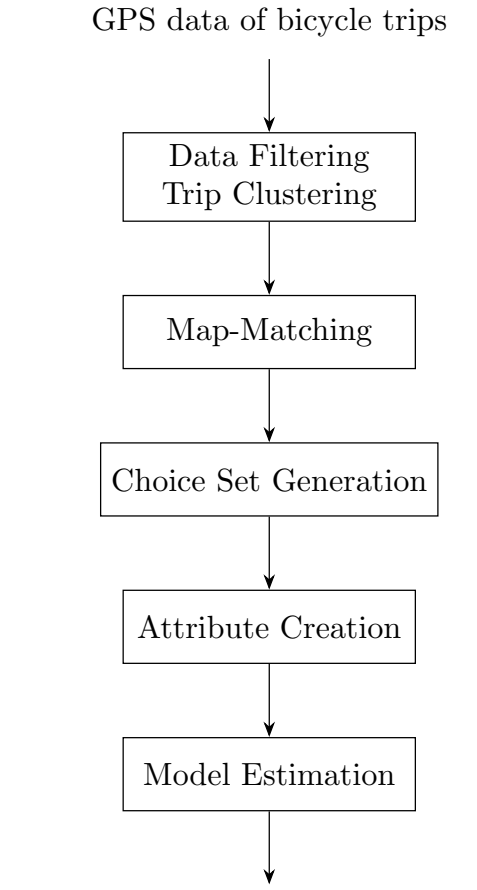

FIGURE 4.1: Summary of the methodology

4.2 Data filtering

This section describes the steps performed to prepare the data before identifying routes used by bicyclists and estimating a model. First, a plausibility check was performed on the 4,457 trips collected during the European cycling challenge in part 4.2.1. 424 irrelevant trips were eliminated because of implausible speed, very short distances or round trips. Then, a spatial filtering was realized in part 4.2.2. The aim of this step was to obtain clusters of similar trips that will be part of the same choice set for model estimation. At the end, 2,919 trips remain.

4.2.1 Elimination of irrelevant trips

a. Problem of multiple origins and destinations per ID

The study of the DETAIL file containing one line per GPS point shows some inaccuracies. Some trips have multiple departure and arrival points. These points can be identified by the variable "Type", which is equal to "start", "mid" or "end". As a result, several trips have the same trip ID. To solve this issue, it was necessary to identify whether trips with the same IDs correspond to different portions of the same trips or to different trips. Two situations were distinguished according to the way the trip was recorded.

First of all, for routes that were manually drawn by participants, trips with the same IDs were treated as different trips. It was assumed that the problem of multiple ID was due to the fact that participants reported several trips one after another when connecting to the website.

Secondly, trips collected directly via the app that have the same ID were treated as unique trips. An analysis of a sample of these trips revealed that trip portions were continuous over space and time. A manipulation of the phone while traveling may have caused these inconsistencies.

b. Elimination of very short trips and round trips

After dealing with the problem of multiple ID, some other data filtering process were performed to eliminate trips that cannot be used for route choice estimation. It is unlikely that a route choice is possible for very short trips. Consequently, trips shorter than 300 m were suppressed. Then, round trips were identified and eliminated. A round trip was defined as a trip with a distance between its origin and its destination less than 200 meters. Finally, only trips with speeds lower than 45 km/h were kept in the dataset. This is intended to eliminate irrelevant trips or trips traveled by car.

The flowchart in figure 4.2 summarizes the different steps performed during the plausibility check.

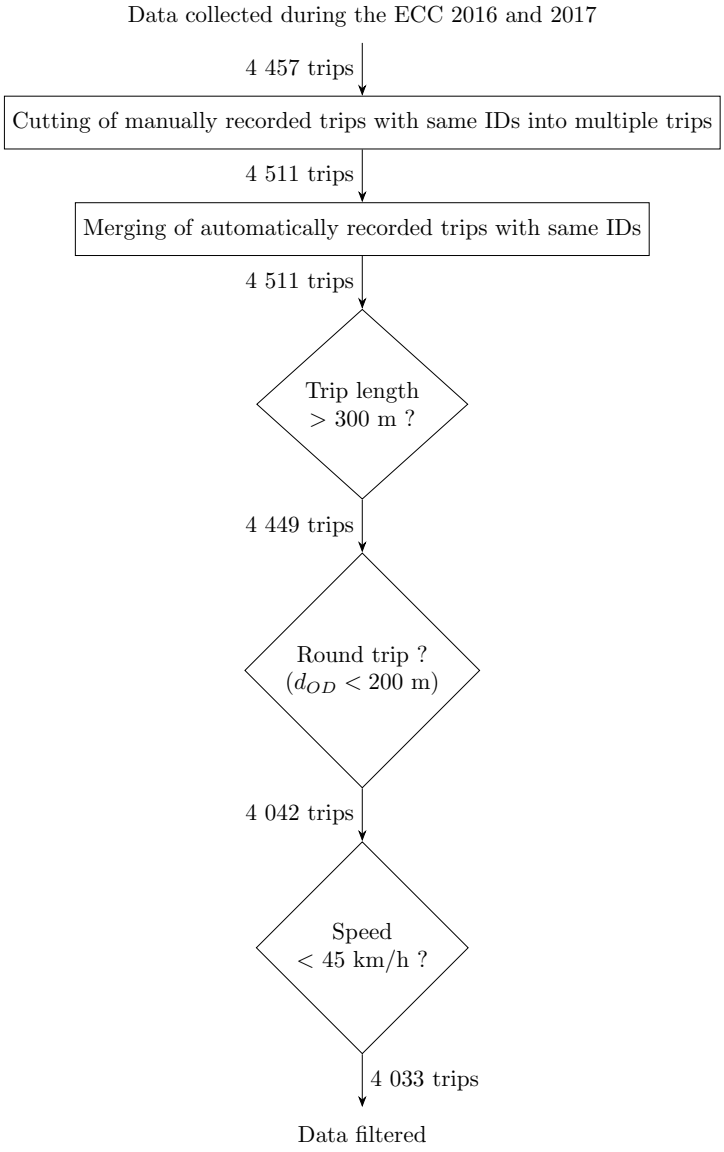

FIGURE 4.2: Flowchart of trip filtering steps during the plausibility check

4.2.2 Trip Clustering

a. K-Mean

The data-driven approach for generating the choice set consists of considering the observed routes between each OD pair. However, because of the high resolution of GPS data, it is unlikely that two trips have exactly the same origin and same destination. A clustering method was applied to gather trips that have similar OD pairs. Ton et al. (2017) used K-Mean (Hartigan and Wong, 1979). Each cluster is associated with a centroid and each point is assigned to the cluster with the closest centroid. The intra-cluster distances are minimized whereas the inter-cluster distances are maximized. The number of clusters needs to be specified in advance. A K-Mean clustering analysis was performed on the dataset with K=500. An example of a cluster can be visualized in figure 4.3.

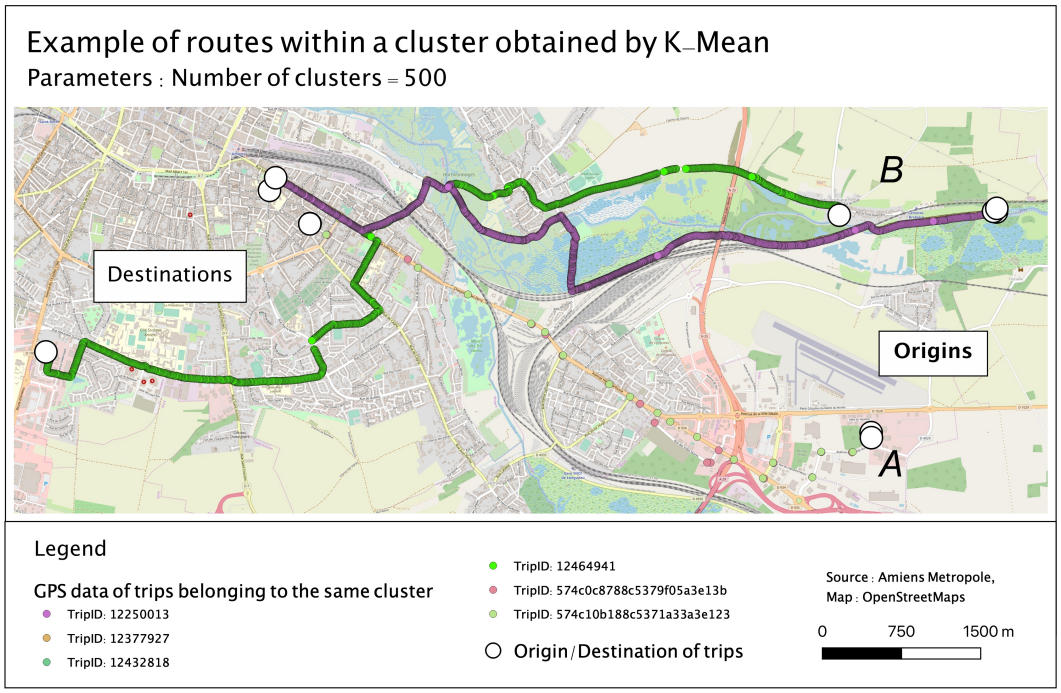

FIGURE 4.3: Example of cluster obtained by K-Mean

It can be seen that the origins and destinations of trips in this cluster cannot be considered similar for a route choice estimation. The trips towards Amiens Glisy aerodrome (A) and Larmotte Brebière (B) and are gathered in the same cluster whereas they are located 1.3 km away. One of the issues of the K-Mean method is that the four coordinates of the origin and destination are treated as separate variables. Therefore, a very close origin can compensate for a more distant destination. Distances between points are parameters that are difficult to control. Moreover, K-Mean method has problems when clusters have different sizes and densities, and when the data set has outliers. The data used have a high heterogeneity of location densities and have many outliers. The density of points is especially much higher in Amiens South than in the rest of the region. As a consequence, K-Mean method does not perform well. Therefore, another approach was preferred in the following.

b. Implementation of an alternative method inspired by He et al. (2018)

An alternative method for clustering origin-destination trips was implemented. This method was inspired by He et al. (2018) and its **Simple line clustering method** recently proposed. This intuitive approach searches for neighboring lines for every OD trip within a certain radius. Two parameters are required: a **searching radius** and a **minimum number of lines** in a cluster. With these parameters composition of clusters can be carefully controlled. The searching radius applies to both origin and destination. For example, a very close destination cannot compensate for a further destination as it can be the case for K-Mean clusters.

This method is based on **neighboring lines** concept. Neighboring lines of a given O_iD_i pair is defined as all the OD lines that have both origin and destination within the searching radius d of O_i and D_i (fig. 4.4). In other words, neighboring lines $NLs(L_i)$ of a centerline L_i are:

$$NLs(L_i) = \{L_j \in L | dist(O_i, O_j) \leq d \cap dist(D_i, D_j) \leq d\}$$

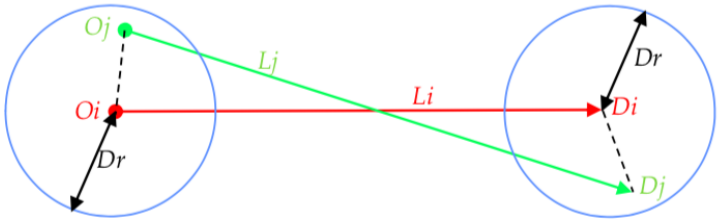

FIGURE 4.4: Definition of neighboring line of a center line (He et al., 2018)

Steps can be described as follows. The OD pair with the highest number of neighboring lines and its associated neighboring lines will form a cluster. The related lines are then extracted from the dataset and the next iteration is performed with the remaining OD pairs. The different steps of the clustering process are summarized in figure 4.5. The code in R is available upon request.

In this study, the searching radius is a function of the **trip length**. If the trip is very long, we may accept together in a cluster OD pairs that have origins or destinations further away. On the contrary, if the trips are very short, considering a smaller radius seems more appropriate. Expression 4.1 is considered for the searching radius.

$$\text{Searching Radius} = \min(\text{Ratio} * \text{tripLength}, D_{max}) \tag{4.1}$$

where D_{max} is the maximum searching radius. It is defined to avoid very distant origins or destinations together in a cluster.

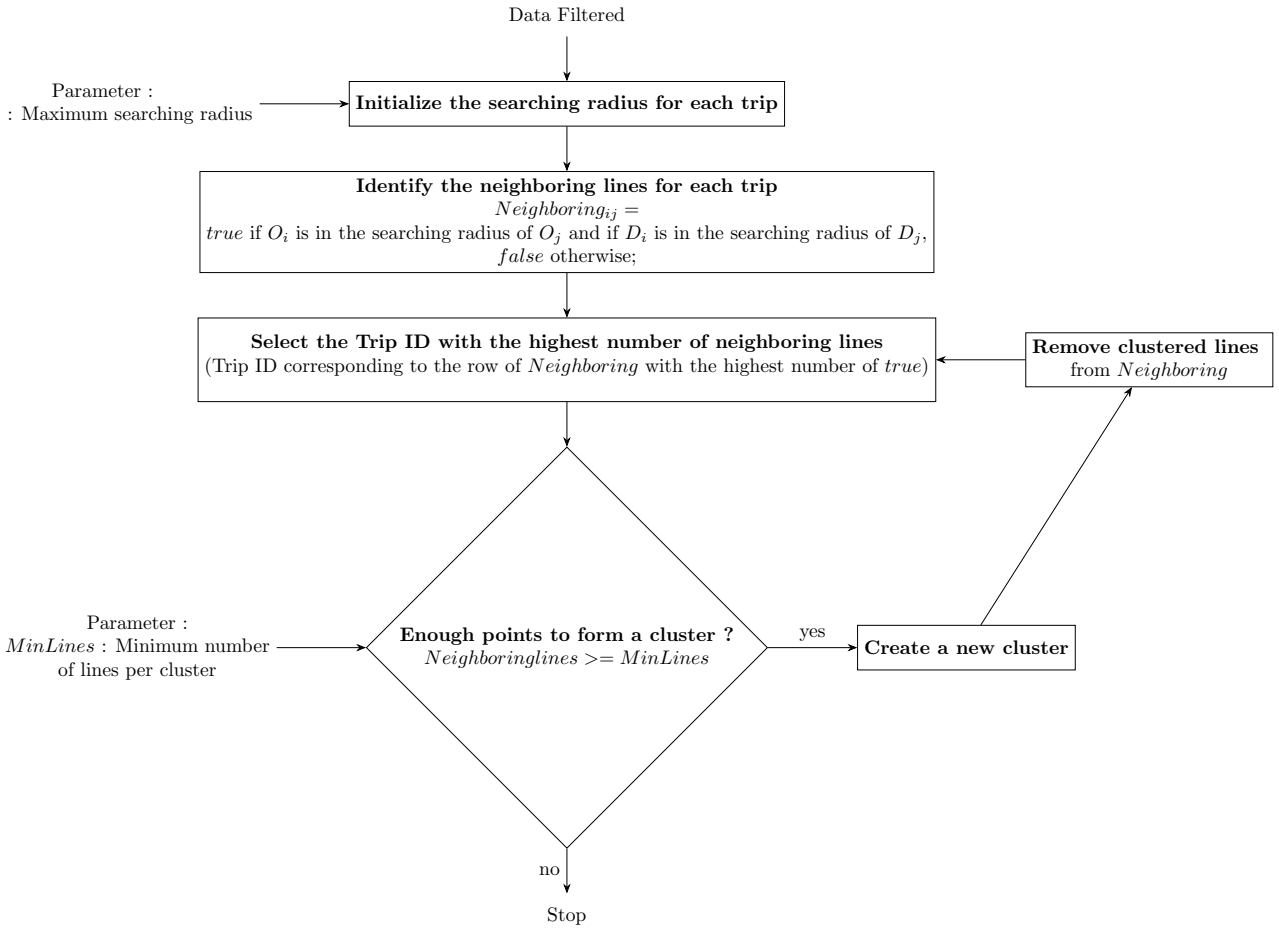

FIGURE 4.5: Flowchart of the clustering process

The methodology used in this study differs from the one proposed by He et al. (2018) by the **expression of the searching radius**. He et al. (2018) proposed two methods a basic one and a more advanced one. In the first basic method, the searching radius is set as a constant. However, by doing so, He et al. (2018) showed that it is impossible to cluster trips with length shorter than 0,83d because there is then no guarantee that clustered trips will be in the same direction. Therefore, with this method some trips must be excluded from the dataset. The second more advanced method consists of iteratively decreasing the radius when all the remaining trips are too short to be matched. Thus, when all the trips that could be clustered were clustered, the algorithm is run again for the rest of the trips with a smaller radius. In this study, the approach is different. Within one iteration, the searching radius is not the same for all the OD pairs but depends on the trip length. It has the advantage of using a clear and intuitive expression for the searching radius and requiring to run the program only once.

c. Parameters selection

The **searching radius** needs to be carefully selected. The objective is to minimize the number of non-clustered trips, while having a reasonable distance between the origins and between the destination within one cluster. The higher is the searching radius, the further are the OD lines belonging to the same cluster. Several tests were performed to evaluate the influence of the maximum searching radius D_{max} and the ratio of trip length that appear in the expression of the searching radius. The minimum number of trips in a cluster is set as two. Table 4.1 presents the results obtained in terms of cluster characteristics and mean distance within a cluster.

Parameters for clustering		**Cluster characteristics**			**Mean distance (in m) within a**		
D_{max} (in m)	Ratio of trip length	% of lines in clusters	Number of clusters	Average trips per cluster	**cluster** Origins	**between:** Destinations	O and D
300	10 %	79 %	445	12.5	53	36	89
300	20 %	84 %	442	12.8	75	59	134
400	10 %	80 %	454	12.7	64	48	113
400	**15 %**	85 %	449	12.4	86	74	160
400	20 %	86 %	429	12.7	100	84	185
500	10 %	81 %	450	12.8	67	50	117
500	20 %	88 %	394	13.4	119	102	222
K-Mean Clustering							
k=500		99%	446	8.5	141	141	282

TABLE 4.1: Parameters selection for trips clustering and comparaison with the K-Mean method

The objective was to keep an average distance between the OD lines within a cluster of less than 160 m, while maximizing the number of OD clustered. The clusters obtained with a ratio of 15% and a maximum radius of $D_{max} = 400$ m are the ones that meet the best these requirements. With these parameters, 3,414 trips are clustered in a total of 449 groups. There is an average of 12.5 trips per cluster.

A comparison with the clusters obtained by the K-Mean method shows that the method applied with the selected set of parameters outperforms the K-Mean method. The characteristics of the nearly 450 clusters are very different. The percentage of lines clustered is much higher (+14%) for the K-Mean but the distances are not accurately controlled. The average distance between two OD is equal to 282 m against 160 m for the Simple line clustering method. The most striking difference concerns the maximum distance between two OD belonging to the same cluster. It is equal to 14 km for K-Mean against 760 m for the selected method, i.e. 19 times more. Thus, the method developed allows a more homogeneous and controlled composition of clusters compared to the K-Mean method.

Figure 4.6 shows an example of a cluster obtained by the Simple line clustering method containing four trips.

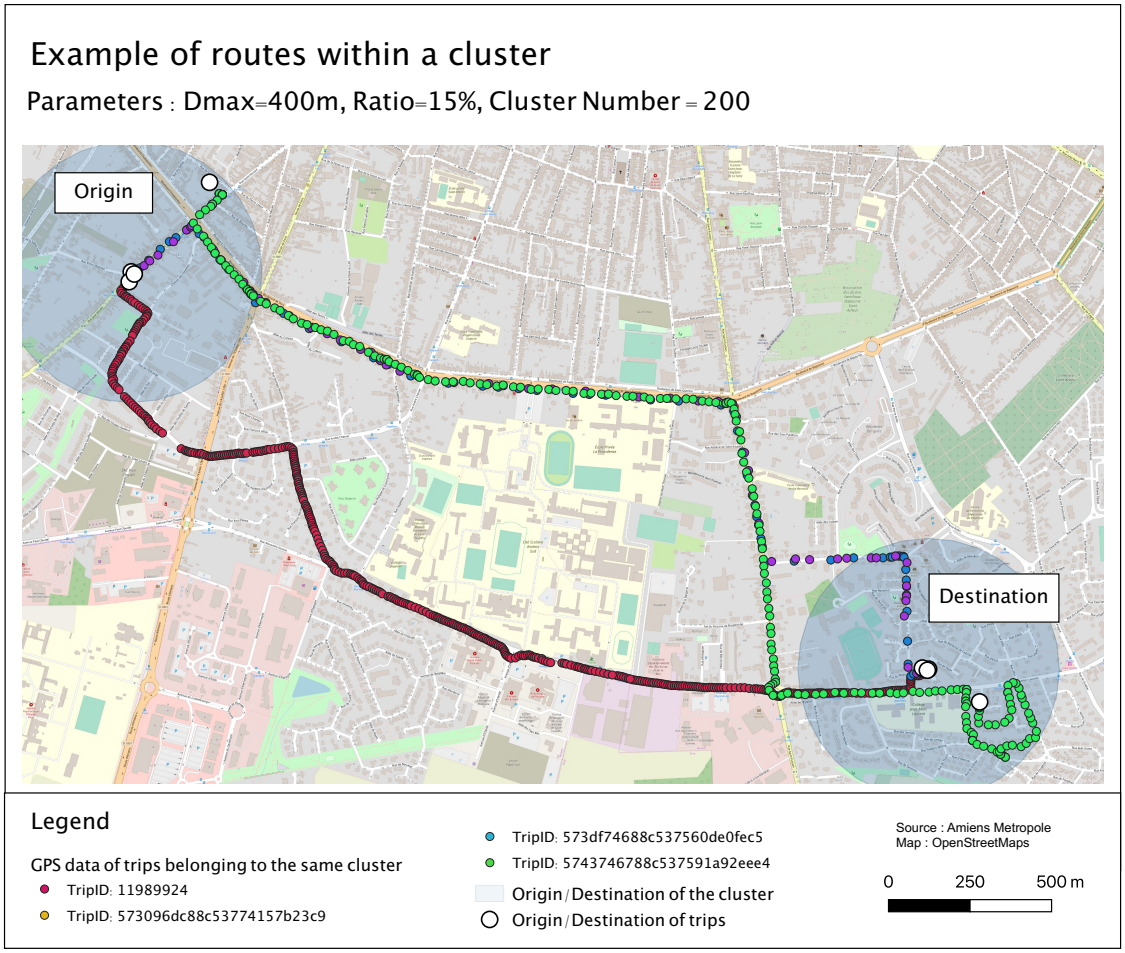

FIGURE 4.6: Example of trips belonging to the same cluster

4.2.3 Boundary definition

In the following, the boundaries of the study were adjusted. 77% the trips clustered are located inside the city of Amiens. Only these trips were kept. This decision avoids the estimation of a route choice model with very different alternatives that are not easily comparable. The land-use of the city of Amiens differs significantly from the rest of Amiens Metropole, which is mainly made of rural areas. By filtering the clusters created, 780 trips that have their origin or their destination outside the city were excluded. Several steps were performed in the following to decrease the number of trips that must be eliminated.

a. Methodology to reduce the number of eliminated trips

The compromise consists of keeping in the dataset not only the trips inside Amiens but also part of the trips crossing Amiens that have an origin or a destination located outside the city. The trips that have at least 70% of their length inside the city were cut and the parts inside Amiens were included in the data used for clustering. It is assumed that the route choice for 70 % of the trip does not significantly differ from the choice for the entire route. This process created some short routes, for which a

route choice was unlikely. Therefore, the new trips shorter than 300 m were eliminated. The same clustering algorithm as presented earlier was applied to the new dataset. As a result, 403 clusters were found for a total of 2,919 clustered lines. Thus, this compromise allowed keeping 285 additional trips. The two flowcharts illustrate the differences between the initial methodology and the one with a compromise found to reduce the number of eliminated trips.

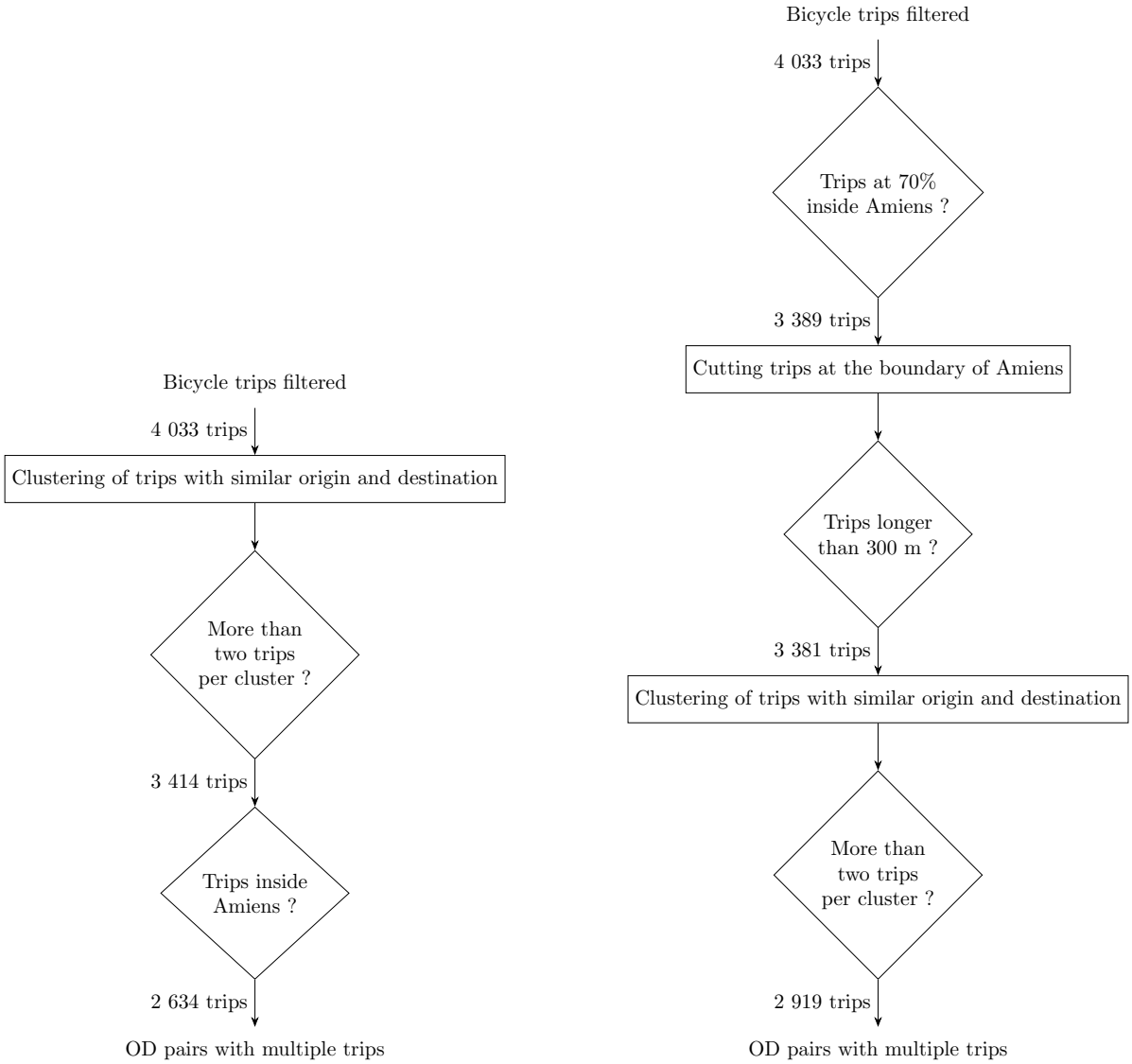

FIGURE 4.7: Comparaison of spatial filtering methods : OD inside Amiens (left), 70% of trip length inside Amiens (right)

Table 4.2 presents the characteristics of the obtained clusters. They can be visualized in figure 4.8. Trips belonging to the same cluster are represented in the same color.

Clusters characteristics with D_{max}=400 m and r=15%	
Number of clusters	403
Number of clustered trips	2,919
% of clustered trips	86 %
Average number of trips per cluster	11.6
Maximum number of trips per cluster	99
Number of clusters with only two trips	134
Mean distance (in m) within a cluster between:	
Origins	90
Destinations	74
Origins and Destinations	163
Maximum distance (in m) within a cluster between:	
Origins	769
Destinations	771
Origins and Destinations	1,288

TABLE 4.2: Characteristics of the clusters obtained for trips with at least 70% of their lengths inside Amiens

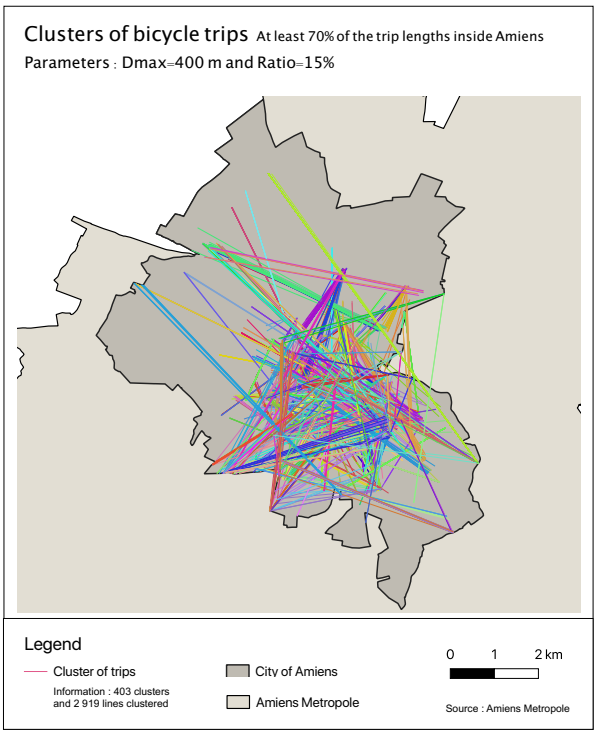

FIGURE 4.8: Clusters obtained for trips with at least 70% of their lengths inside Amiens

4.2.4 Socio-demographic analysis

Some of the bicycle trips recorded have additional socio-demographic information such as age and gender of the participant or trip purpose and type of bike used (table 4.3). The last step of the data processing part consists of evaluating the number of trips that cannot be considered if the socio-demographic attributes are included as attributes in the model estimation. Table 4.4 shows the number of remaining trips.

Variables	**2016**
Gender	58%
Age	58%
Trip Purpose	49%
Type of Bikes	49%

TABLE 4.3: Percentage of answers to the optional questions among the trips filtered

Remaining trips after elimination of trips without :	
Type of bike, age, gender and trip purpose	624
Age, gender and trip purpose	632
Gender and trip purpose	685
Trip purpose	1,429

TABLE 4.4: Number of remaining trips after elimination of trips without socio-demographic information among the 2,919 trips

Because the loss is consequent, it was decided as a first model estimation not to consider socio-demographic attributes. Moreover, age and gender were not found significant in the models reviewed in the literature (Hood et al., 2011; Chen et al., 2018; Bernardi et al., 2018). Trip purpose was successfully included in Broach et al. (2012), Ton et al. (2017) and Bernardi et al. (2018), and could be considered in a smaller model in a second time.

Conclusion: Thus, after the trips filtering steps, 2,919 trips remain. However, for creating a data-driven choice set for a route choice model, more than one route is required per OD pair. This condition will be checked in a second step, after matching the GPS routes to the network.

4.3 Map-matching

4.3.1 Map-matching in the literature

The sequence of GPS points needs to be processed in order to accurately identify the roads that were taken by its users. This process is called map-matching (Quddus et al., 2007).

a. Map-matching problems

Map-matching is a challenging task for several reasons. Identifying the route is not easy, firstly, because of the limited GPS accuracy and the sparseness and noise in the data. Secondly, the digital road network is only a representation of the real world. It can have missing segments and connectivity problems (Van Dijk and De Jong, 2017). Moreover, networks can have complex road geometry with complicated intersections, multi-layer roads or parallel links close to each other (Dalumpines and Scott, 2011). All these cases make the identification of traveled routes especially difficult.

Moreover, map-matching of bicycle routes shows additional issues compared to car routes. Bicyclists are much more flexible than car drivers and tend to behave both like cars and pedestrians. For example, they can ride in parks, cross pedestrian areas and ride in opposite directions in some streets. These aspects are often not taken into account in network representations built for cars but these behaviors must be allowed in order to get an appropriate map-matching.

b. Map-matching algorithms in the literature

Map-matching algorithms are classified by Quddus et al. (2007) into three main categories:

- **Geometric methods**: use geometric information of the spatial road network data (point-to-point, point-to-curve, curve-to-curve, etc.) by considering only the shape of the links.
- **Topological methods**: consider the connectivity of the network links and the sequence of measurements in addition to links geometry.
- **Other advanced techniques**: probabilistic algorithms based on error regions around GPS signals, Kalman Filter, Hidden Markov models, Multiple Hypothesis Techniques, etc.

Map-matching algorithms are used in two different contexts: for real-time and for post-processing applications. **Real time algorithms** associate the position of drivers to the road network during the recording process, whereas **post-processing** or off-line map-matchings are used after the end of the trips. This study of bicycle route choice belongs to the second case because the analysis of routes is realized after collecting data. The first type of map-matching is by far the largest group investigated in the literature because of its applications in navigations systems. This explains why post-processing map matching problems often use the same algorithms as for real-time applications. However, as mentioned by Dalumpines and Scott (2011), the two problems have important differences. Firstly, unlike real-time map-matching, a post-processing map-matching does not require identifying exact locations of vehicles within links over time but only the list of traveled links. Secondly, post-processing map-matching enables the use of other methods. All points

can be considered and slower performance in favor of accuracy is tolerated.

Dalumpines and Scott (2011) proposed to use a different approach than those developed for real-time applications. They invented a **GIS based post-processing map-matching**. The algorithm creates a buffer region around GPS trajectories, which delimits the network in which a shortest path is searched between the first and the last GPS points. With this method, 88 percent of 101 trips were accurately matched by Dalumpines and Scott (2011) in Halifax, Nova Scotia, Canada with a buffer size of 50 meters.

More recently, Van Dijk and De Jong (2017) reported two additional methods based on a GIS environment. The first one is called *Connected subset assignment procedure* and connects consecutive GPS points by mini shortest path assignments. The second one, called *Impedance reduction method* consists of counting the number of nearest GPS-measurements for each road segment and adjusting the link impedance accordingly before running a shortest path algorithm.

4.3.2 Map-matching method applied: shortest path search in subnetworks

The method developed by Dalumpines and Scott (2011) based both on a **subnetwork delimited by a buffer area around GPS points** and on **shortest path analysis** was selected for this study. It has the advantage of being easily implemented while obtaining satisfactory results. It only requires combining available tools on GIS softwares, like shortest route analysis and buffer creation. The different steps performed for matching the recorded trips to the digital road network are presented in this part. Figure 4.9 summarizes the workflow. Because of the high number of trips, it was necessary to automate processes. Python was used for this purpose because of its efficient existing libraries dealing with graphs and geographical data frame (OSMnx, NetworkX, geopandas). The code is available upon request.

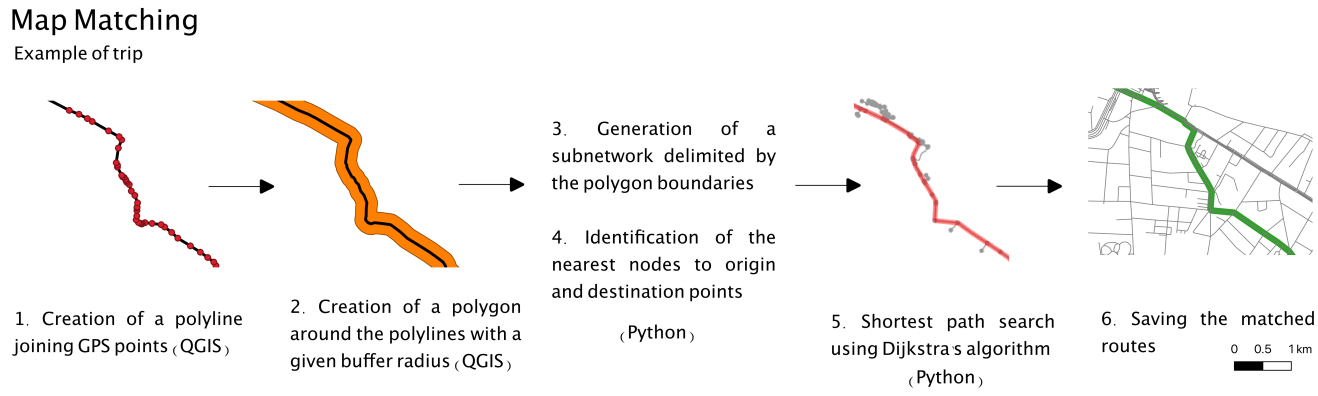

FIGURE 4.9: Methodology for map-matching the routes

The first step consisted of **converting GPS points to lines**. It was performed in QGIS. Only the trips that were automatically recorded are used. Then, **buffers were created around the lines** with QGIS. The buffer size is a parameter that had to be given. The section 4.3.3 details how it was selected.

The rest of the steps were performed directly on Python. Street networks were extracted from OpenStreetMap with OSMnx, which is a Python package developed by Boeing (2017). It allows automating the download of streets from OpenStreetMap and their construction into graphs. The network can be download by providing a polygon of the desired street network's boundaries. This functionality was used to **generate the street-network inside the buffer polygon** created around the GPS points. Thus, one subnetwork was loaded for each trip. The network type had to be specified (drive, walk, bike, all, etc.). For this study, all non-private OpenStreetMap streets and paths were loaded because the bike network did not include all the links traveled by bicyclists in the collected data. Moreover, the graphs were transformed into undirected graphs for map-matching. In an undirected graph, edges point in both directions. This step was performed because it was observed that many streets were used in the opposite direction by bicyclists.

Before applying a shortest path algorithms, an origin node and a destination node had to be identified for each trip. A **search for the nearest node along the nearest edge** was preferred to a basic nearest node function because the nearest nodes may not be connected to the link were the bicyclists stops or starts as illustrated in figure 4.10. Finally, the Python package NetworkX (Hagberg et al., 2008) was used to compute the **shortest path between origin nodes and destination nodes**. It is based on Dijkstra's algorithm (Dijkstra, 1959). The link lengths were used as weights. As a last step, the shortest route for each subnetwork was saved into a shapefile.

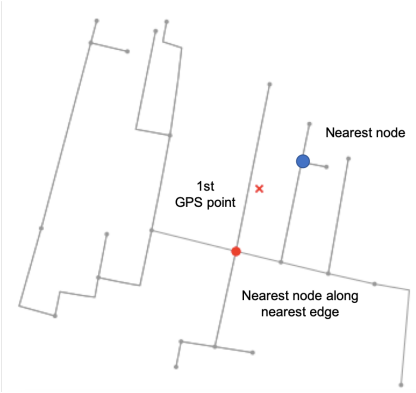

FIGURE 4.10: Difference between the nearest node and the nearest node along edge

4.3.3 Selection of the buffer radius

To apply this method, one parameter is required: the buffer radius. Based on the results of a sensitivity analysis, Dalumpines and Scott (2011) recommends to use a buffer equals to five to six times the horizontal accuracy of GPS data. However, the adequate buffer distance also depends on the road network and on the distance between two consecutive GPS points. If the buffer radius is too small, some links used will not be part of the subnetwork generated. On the contrary, if it is too large, the subnetwork will contain many alternative routes to go from A to B and the shortest path algorithm will probably not give as a result the real traveled route. Several expressions for the buffer radius were tested on a **sample of 50 trips** in order to select the best parameter. The sample was randomly selected.

a. Definition of expressions for the buffer radius

Before matching the routes, additional filtering steps were performed on the data. 166 trips with too distant points where excluded from the dataset. In other words, trips with points more than 500 m apart were not considered in this study. Moreover, when the maximum distance was between 200 and 500 m, a distinction was made according to the number of distant points. If only one or two pairs of points were distant from each other ($> 200m$), a satisfactory map-matching was still expected and the trips were kept, otherwise, they were excluded.

Three different expressions were tested for the buffer radius. Firstly, **two expressions depending on the distance between two GPS measurements** were used. The idea is that closer points can enable a smaller buffer size. The minimum buffer radius was chosen according to the horizontal accuracy and the road network. An analysis of the road network showed that a minimum buffer size of 25 m is required to include the traveled links. This value was measured in a pedestrian street (Rue des Trois Cailloux) where the bicyclists can ride over all the street width. The last buffer radius that was tried corresponds to a **constant size of 50 meters**. This value is the one used by Dalumpines and Scott (2011). The different expressions (in meters) for the buffer radius can be visualized in figure 4.11 and have the following equations:

- Test 1 : Buffer Radius = $\min(100, \max(25, \frac{\max \Delta d}{2}))$
- Test 2 : Buffer Radius = $\min(50, \max(25, \frac{\max \Delta d}{2}))$
- Test 3 : Buffer Radius =50

with Δd, the distance between two consecutive GPS points.

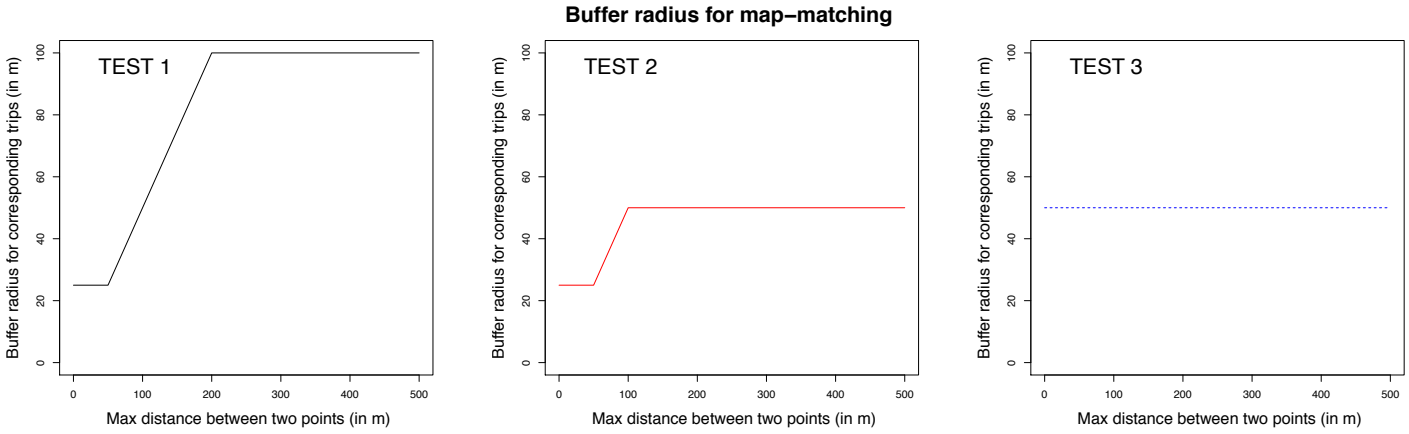

FIGURE 4.11: Tests of buffer radius

b. Error classification

Classifying the errors and counting the number of errors for different radius expressions helped selecting the best buffer radius among the three tests. The errors are mainly due to an inadequate buffer size, wrong network coding or complicated network layouts. The points on the maps in the following pictures represent the GPS measurements, while the lines are the route obtained by the map-matching algorithm.

i. Problems due to a too large buffer radius

The first most common type of error is wrong matches caused by a too large buffer radius. Another alternative route, shorter than the real one, is included in the subnetwork and selected by the shortest path algorithm. This type of error especially happens when there are streets close to each other or intersections with possible short-cuts (fig. 4.12). They can be avoided to some extend by decreasing the buffer size.

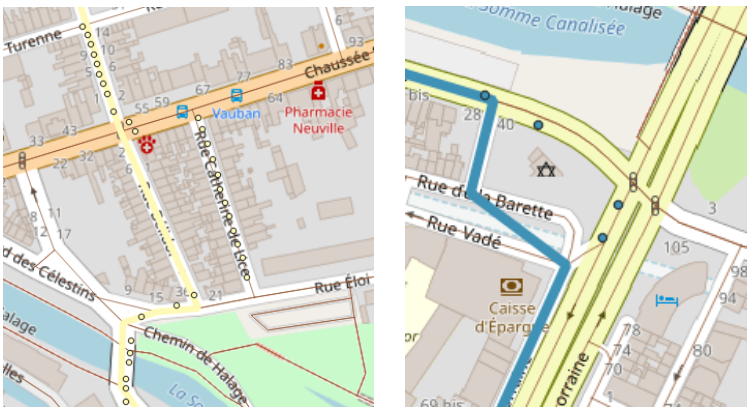

FIGURE 4.12: Examples of errors due to a too large buffer radius

ii. Problems due to a too small buffer radius

The second type of error is on the contrary due to a too small buffer radius that cannot compensate for the horizontal inaccuracy of GPS measurement. The route used by bicyclists is not included in the subnetwork and the algorithm does not succeed in joining trip origin and trip destination and stops in the middle of the travel (fig. 4.13).

FIGURE 4.13: Example of errors due to a too small buffer radius

iii. Problems due to parallel streets

Thus, a trade-off is required in the choice of the buffer size. It must be neither too big nor too small. Parallel streets that are located at a distance of less than one radius cannot be properly identified. Therefore, the method applied is not capable of identifying the traveled lane. This is not problematic for estimating the effects of the built environment on bicycle route choice when the lanes on a street have the same properties but in some cases, the differences can be significant. For example, one part of the lanes can be coded in OpenStreetMap as *Secondary street* and the others as *Residential streets*. This is often the case in Amiens. An example of such situations is presented in figure 4.14.

Figure removed due to possible copyright infringements

FIGURE 4.14: Examples of errors due to a too small buffer radius

The lanes are separated into two parts: one main part is in the middle of the street, where there are multiple lanes in both directions, and another part next to the sidewalk on each side, surrounded by on-street parkings. The two parts are separated by trees and parked cars. The bicyclists tend to use the side parts, where the traffic is less important but the associated traces are often matched to the wrong lanes. Therefore, manual corrections of routes along this type of road geometry will be required in the following. This type of error is called problem of parallel streets.

iv. Problems due to a wrong network coding

Finally, some errors in map-matching are due to a wrong network coding. Some links on which bicyclists can drive are not included in the OpenStreetMap network. This is especially common on squares. As explained before, the bicyclists are much more flexible than cars and behave sometimes like pedestrians. For example, some of them cross streets in the middle of links, which is not allowed in the digital representation of the network, or they take short-cuts where there is no path. Examples of problems due to a wrong network coding can be visualized in fig. 4.15.

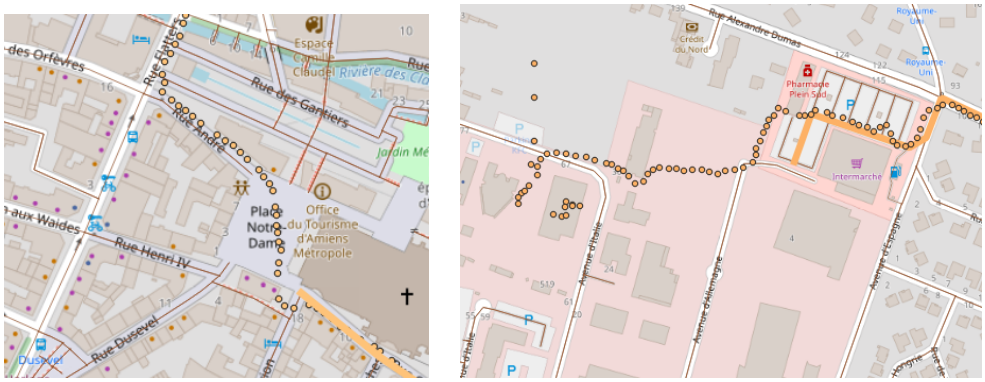

FIGURE 4.15: Example of errors due to missing links

Missing links are very problematic for map-matching. Either, the buffer is large and another alternative route is selected, or the matched route fails in joining trip origin and trip destination. In the second case, the problem occurs at step 3 of the process (part 4.3.2) when the subnetworks delimited by the boundaries of the polygons are downloaded. The graph generated via OSMnx only contains connected links. If there is a missing link in the middle of a route, the subnetwork can become discontinuous and only a continuous part of it will be generated. In this case, the subnetwork does not have the same size as the buffer and either origin or destination is not included. Therefore, the shortest path algorithm will fail in linking true origin and destination and only a subpart of the route will be obtained.

c. Analysis of the results for the sample

The routes obtained for the sample of 50 trips and the three different buffer expressions (part 4.3.3) were analyzed and the errors were classified as described in part 2.3.3.b.. Table 4.5 shows the results for the three tests. Figure 4.16 represents the percentage of errors due to a too large buffer radius among trips belonging to the same distance group. It confirms the relevancy of adapting the buffer radius to the distance between GPS points.

Test number	**1**	**2**	**3**
Maximum buffer radius (m)	100	50	50
Minimum buffer radius (m)	25	25	50
Number of correct matches	25	33	20
Errors :			
- due to a buffer radius too big	14	4	17
- due to a buffer radius too small	4	6	5
- due to network coding	4	4	4
- due to parallel streets	3	3	3

TABLE 4.5: Tests on a sample of 50 trips with different equations for the buffer radius

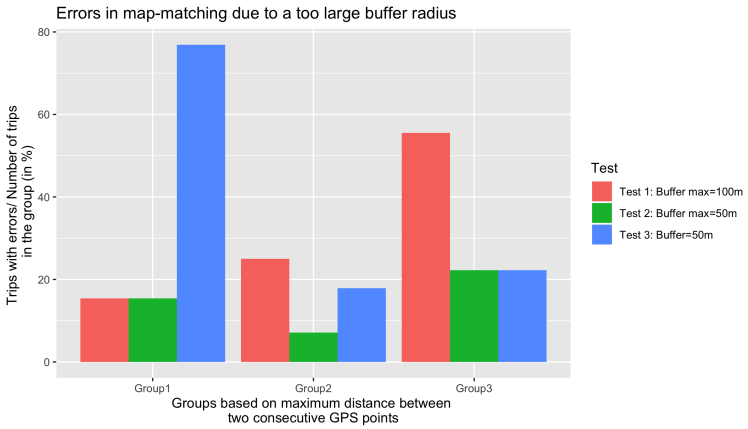

FIGURE 4.16: Distance between two consecutive GPS points (Group1:$\Delta d \leq 50m$, Group2: $50 < \Delta d \leq 200m$, Group3: $\Delta d > 200m$)

Thus, test 2 has the best results with 66% of perfect matches. 14% of trips are not correctly matched due to wrong network coding or problems of parallel streets and must be partly manually corrected. Therefore, at the end, 80% of correct matches can be expected. Moreover, test 2 also offers the lowest error rate in each distance group. Reducing to 25 m the buffer size for trips that have a distance between two GPS points smaller than 50 m succeeds in decreasing the error ratio by 60%. Moreover, a maximum buffer of 50 m for trips with points more than 200 m away is more appropriate than a radius of 100 m.

Therefore, for each trip, the buffer radius presented in equation 4.2 was selected for map-matching.

$$\text{Buffer Radius } (\max \Delta d) = \min(50, \max(25, \frac{\max \Delta d}{2})) \tag{4.2}$$

where Δd is the distance between two consecutive GPS point in the recorded trip.

4.3.4 Post-processing

a. Analysis of the results

The steps presented in 4.3.2 were run with the selected buffer (equation 4.2). Around 10 hours were required to match 2,711 trips. Comparing manually each route with

the measured points is impossible due to the high number of trips. Therefore, automatic checks were imagined to eliminate problematic trips. True closest node from origin and matched origins, and true closest node from destination and matched destination were compared. Errors due to wrong network coding or too small buffer radius can result in a partial path that does not join origin and destination. 72% have exactly the same origins as expected and 67% the same destinations. In total, 49% of matched trips have correct origins and destinations. This low percentage is due to the difficulty of identifying the routes for the first and the last meters. Bicyclists can park their bikes in private places that are not reachable with the OpenStreetMap network. However, the difference can be sometimes insignificant compared to the entire trip. For example, if the true destination differs from the matched one by only a few meters, it is reasonable to consider that the trip is correctly matched and that the characteristics of the route will stay unchanged for the model estimation.

Trips were considered properly matched if :

- $\Delta D \leq 10\%.TripLength$
- or $10\%.TripLength < \Delta D < 15\%.TripLength$ and $\Delta D \leq 350$ m
- or $15\%.TripLength < \Delta D \leq 20\%.TripLength$ and $\Delta D \leq 250$ m

with $\Delta D = \Delta D_o + \Delta D_d$ and $\Delta D_{o/d}$ is the distance between the closest node from origin/destination and matched origin/destination. By doing so, 78% of matched trips were considered as properly matched, which represents 2,109 trips. Figure 4.17 shows the distribution of the difference of distance for problematic trips. The trips that have the same true and matched origin/destination are not represented in the graph.

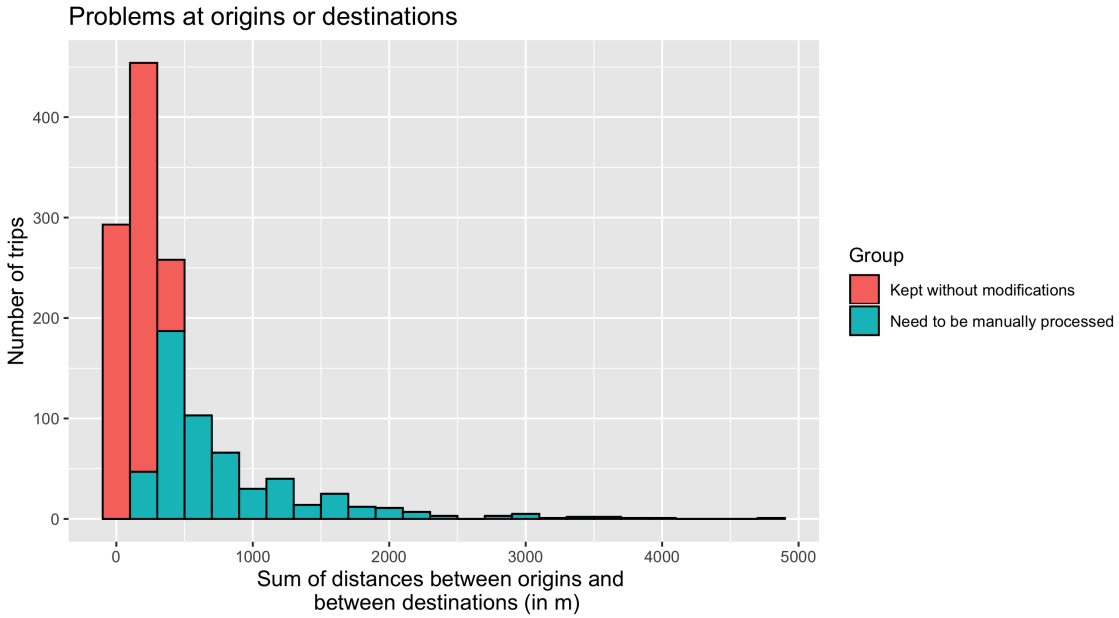

FIGURE 4.17: Problems at origins/ destinations

b. Correction of errors

Because correcting manually 602 trips is unrealistic, an automatic correction of the errors was required. Eliminating completely the errors was not possible without applying another map-matching method, but reducing the errors was still possible. The idea consisted of allowing matched routes to be discontinuous. Until now, if the GPS points could not be matched on a link, the matched route was stopped at the problematic spot without reaching the destination even if the rest of the trip could have been identified.

The same methodology as before was applied but only to the part of the GPS points that was not correctly matched. Thus, another shortest path was generated and then merged with the previous one. Examples of improvements can be seen in figure 4.18. Figure 4.19 presents in detail the methodology applied to reduce the missing part of the matched routes. The steps were repeated twice to allow two discontinuities in a matched route.

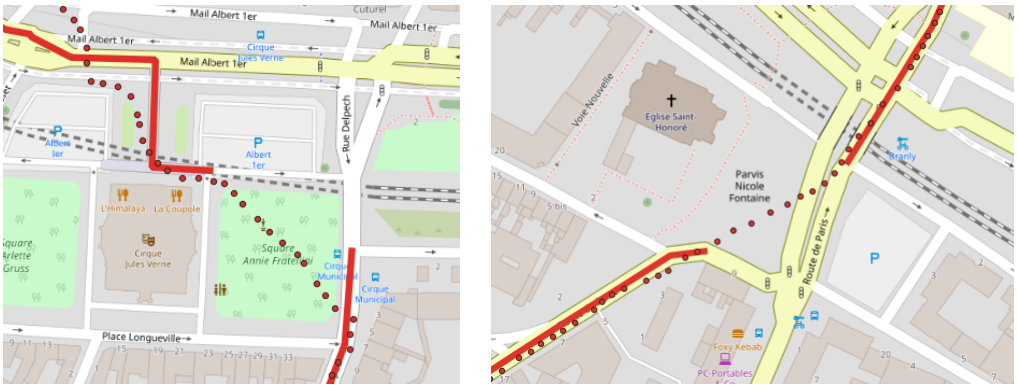

FIGURE 4.18: Example of reduced errors

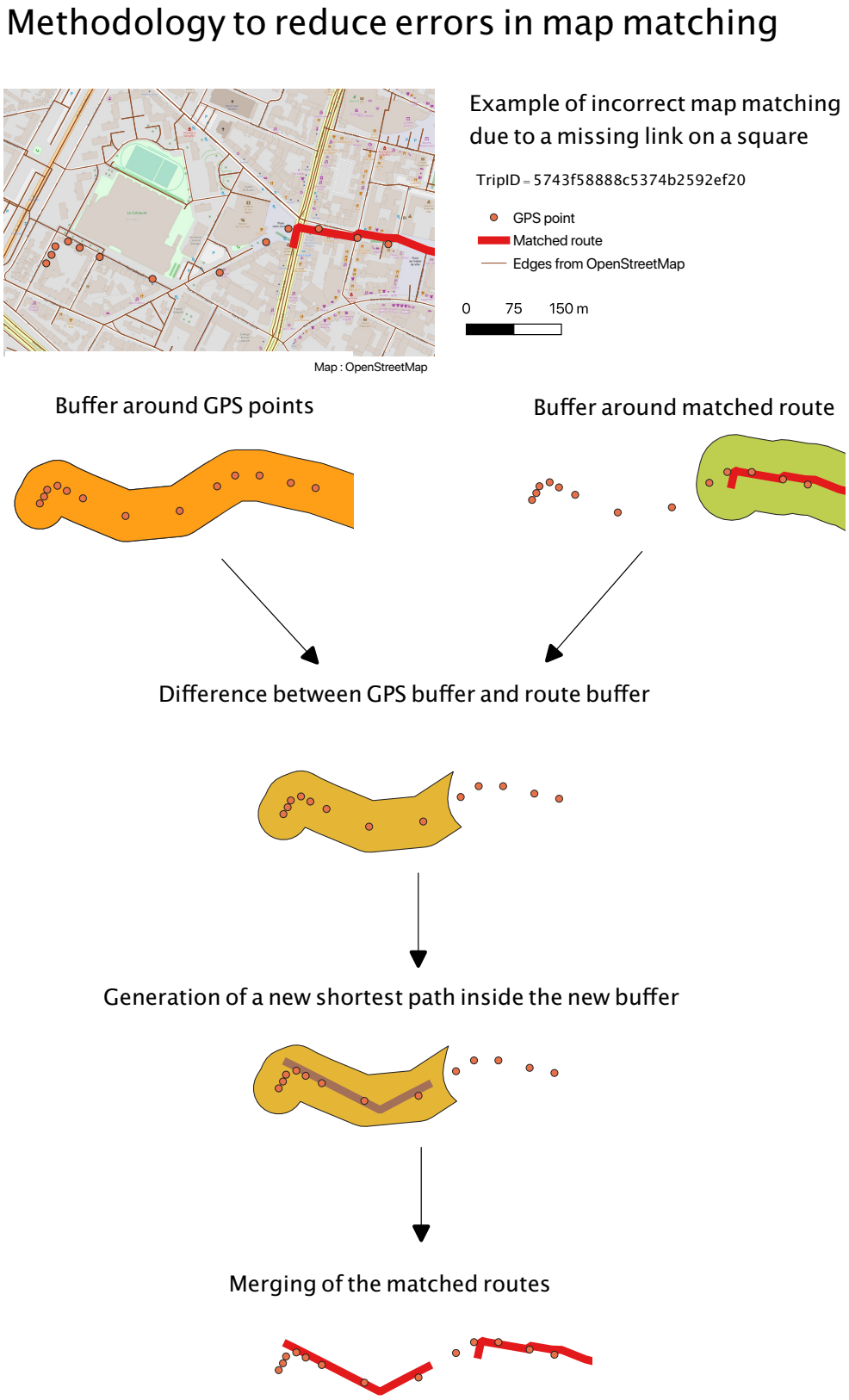

FIGURE 4.19: Steps performed to correct the problematic trips

c. Analysis of the corrected trips

After automatically correcting the problematic trips, matched origins and matched destinations were compared one more time to the theoretical ones to evaluate how well the correction process performed. The graphs in figure 4.20 compare the distances between true and matched origin/destination before and after the correction step. It can be observed, that the errors were significantly reduced. 253 additional trips meet the requirements defined in 2.3.4.a..

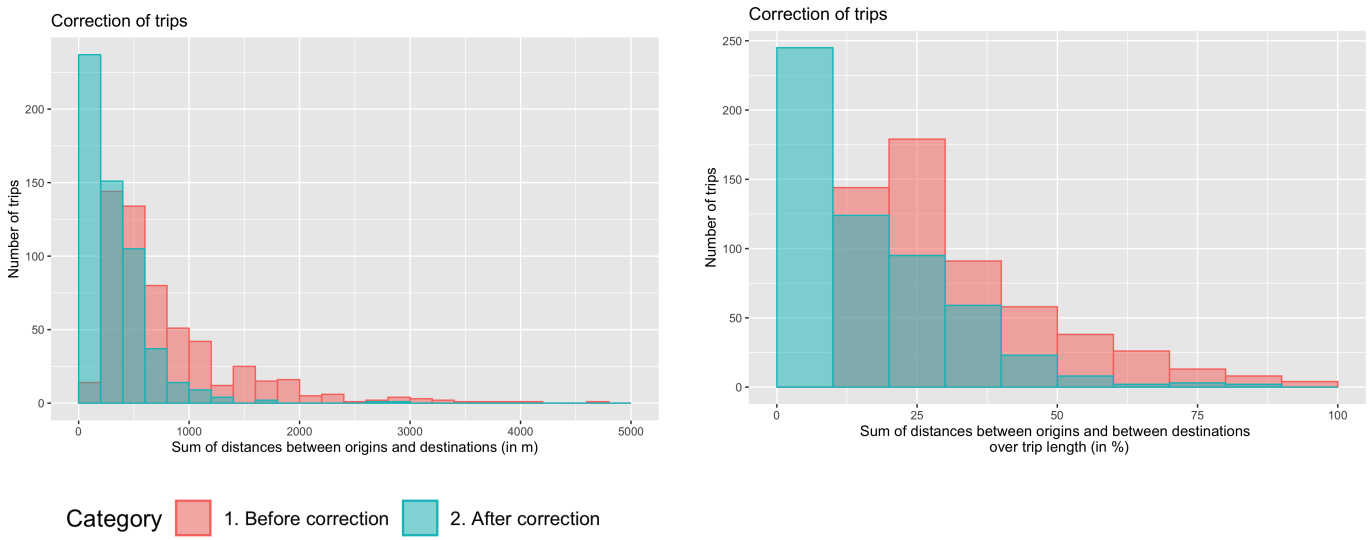

FIGURE 4.20: Comparaison between true and matched origin/destination

Conclusion: Thus, at the end, 2,403 trips could be successfully matched, which represents 82 % of the filtered trips. The method used has the advantage of being relatively easily implemented and straightforward. However, it is very demanding in terms of network quality. Moreover, because bicyclists have flexible behaviors and do not always respect the rules, obtaining a perfect match was not possible. Other more advanced map-matching methods could be used for further research. Figure 4.21 summarizes the different steps performed for identifying the routes.

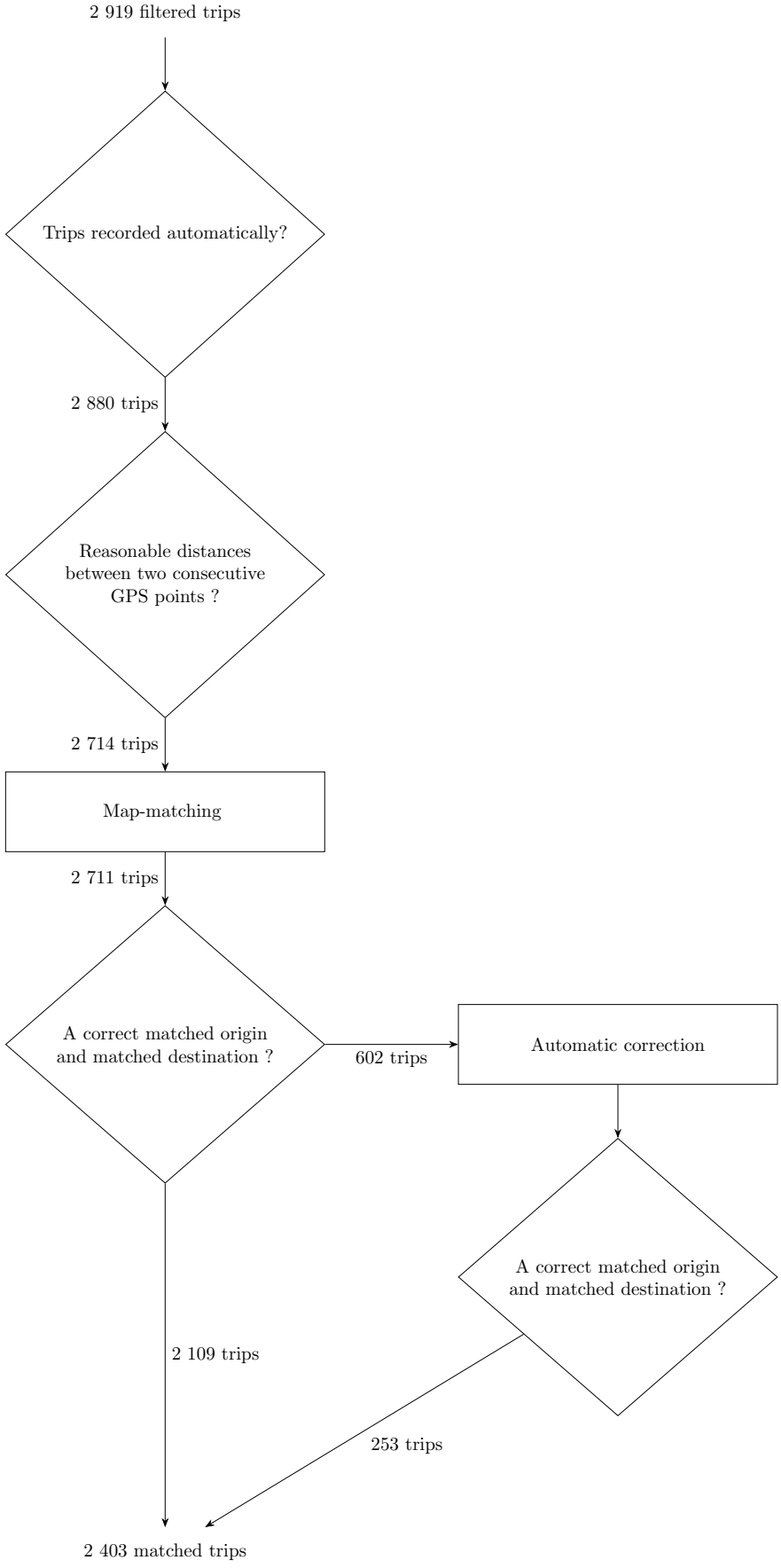

FIGURE 4.21: Summary of the steps performed for identifying the routes

4.4 Choice set generation

After map-matching the routes, the choice set composition was examined in detail to ensure consistent results and additional plausible routes were included in the choice set.

4.4.1 Choice set checking

The first step was to analyze the composition of the choice set. To estimate a route choice model, each cluster of similar trips must contain at least two different routes. These checks were performed thanks to the results of the map-matching process. Two routes are considered different if they have at least a certain number of different links. In other words, the sum of numbers of different links between routes A and B and between routes B and A must be bigger than a threshold. The threshold value (17) was defined based on an empirical analysis of several routes. It must be sufficiently high to allow small differences at the origin or destination among the trips in the same cluster.

A statistical analysis of the cluster composition after the map-matching can be found in table 4.6.

	Median	Mean	Range
Number of trips per cluster	3	6.42	1-76
Number of different routes per cluster	2	2.38	1-17

TABLE 4.6: Characteristics of the clusters after map-matching

In addition to that, figure 4.22.a shows the number of different routes in each cluster, whereas figure 4.22.b illustrates the number of matched trips that can be used for the model according to the minimum number of different routes allowed per cluster.

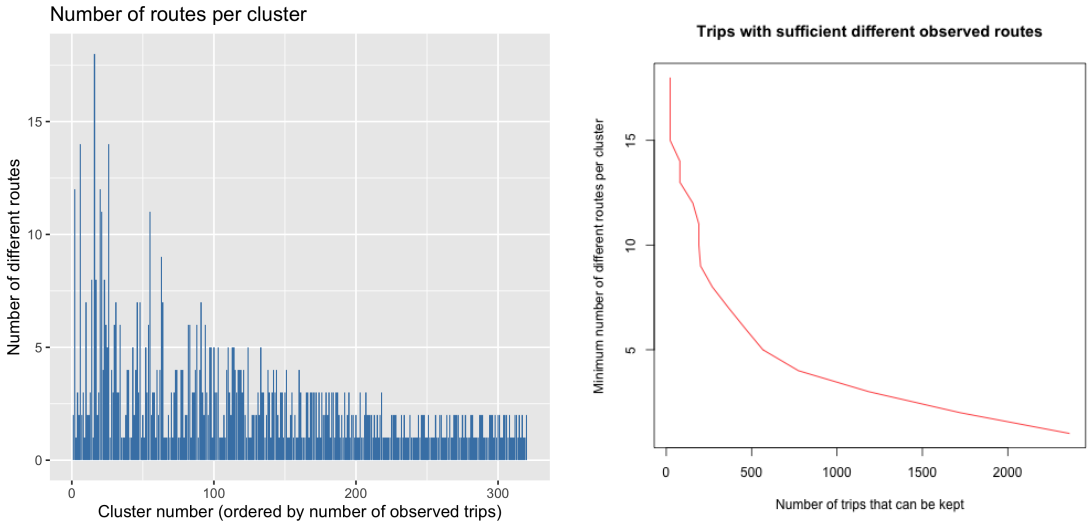

FIGURE 4.22: a: Number of routes per cluster after map-matching,
b: Trips kept according to the minimum number of routes in a cluster

Two comments can be made. Firstly, a high heterogeneity in the number of different routes per cluster can be noticed. Secondly, some clusters do not contain enough

different routes for model estimation but excluding them will cause an important data loss. Therefore, additional plausible routes must be included in the choice set.

4.4.2 Enrichment of the choice set

To increase the number of different routes in each cluster, the choice set was enriched by using the labeling method introduced by Ben-Akiva et al. (1984). It consists of defining several objective functions and generating one path for each criterion. A new impedance is defined for each link before searching for the shortest path. Table 4.7 summarizes the additional routes that were created.

Label	**Link impedance**
Shortest path	$Length$
Shortest path in the car network	$Length$
Path minimizing grade	$Length * f(Grade)$ *with* $f(x) = x$ *if* $x \leq 0$ *and* $f(x) = -\frac{1}{x}$ *otherwise*
Path minimizing the number of intersections	$Length + 200$
Path minimizing the traffic volume	$Length * Road\ Category$ *where Road Category = 4 for a primary link, 3 for a secondary road, 2 for a tertiary road,1 for path, residential or living streets*
Path maximizing bike paths	$Length * Bicycle\ Facility$ *where Bicycle Facility = 0.2 when the link has a bicycle facility, 1 otherwise*
Path maximizing green, water areas and landmarks	$Length * Nice\ Landscape$ *where Nice Landscape = 0.2 when the link has either green areas, water or landmarks in a 100m buffer, 1 otherwise*

TABLE 4.7: Labels, by default the network used is the one with all non-private links

This process was realized on Python with the OSMnx (Boeing, 2017) and NetworkX (Hagberg et al., 2008) packages. Except for the shortest paths in the car network and the paths minimizing grades, the network used contains all the non-private links of OpenStreetMap. Unlike map-matching, the unidirectionality of roads is taken into account to ensure the respect of traffic regulations. The elevation was obtained by using Google Maps Elevation API (GoogleMaps, 2019). The elevation could be obtained only for the driving nodes, which explains why the car network was used for generating the path minimizing grades and not the entire network. Route origin and route destination correspond respectively to the centroid of the cluster origin and the centroid of the cluster destination. At the end, 7 new routes were created for each cluster. Some may be identical or already included in the data-driven choice set.

Figure 4.23 shows the number of different routes for the enriched choice set and table 4.8 presents a statistical analysis of the cluster composition.

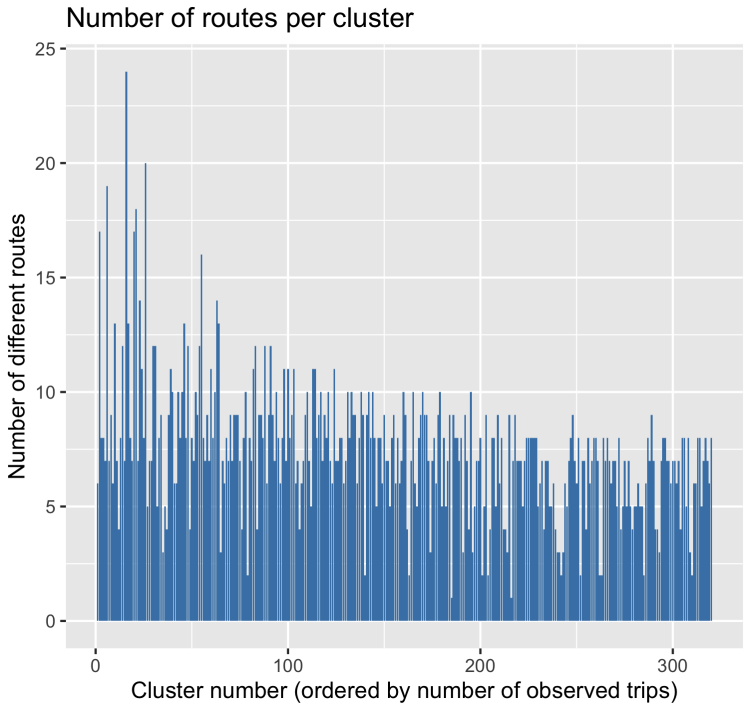

FIGURE 4.23: Number of routes per cluster after enrichment of the choice set

	Median	Mean	Range
Number of trips per cluster	10	13.4	8- 83
Number of different routes per cluster	7	7.22	1-24

TABLE 4.8: Characteristics of the clusters after enrichment of the choice set

Thus, the labeling method succeeded in introducing additional routes in the choice set. In average, there are 7 different routes between each OD pair. Two cluster contained only one route and were removed. At the end, 2,362 trips remains to estimate a model.

4.5 Attribute creation

This section presents the selected attributes for the model estimation. The choice of the attributes to characterize each route were based both on the existing literature on bicycle route choice and on the particularities of Amiens. A field trip was done on the 19^{th} and 20^{th} of October in Amiens to help selecting relevant attributes.

4.5.1 Case study analysis

Attributes for the model estimation should take the characteristics of Amiens into account. The following part investigates the possible factors influencing bicycle route choice in Amiens. Four main areas are explored: traffic volume, intersections, bike facilities, and the land-use.

a. Traffic volume

Firstly, an emphasis was put on traffic analysis. In Amiens, car is by far the dominant mode. In 2010, 56% of all trips were made by car (Pays du Grand Amiénois, 2013). This high mode share generates an important traffic in the city. As suggested by many authors, traffic volume plays an important role in bicycle route choice. In Amiens, the road network has a radial pattern and four concentric roads structure the territory. The bigger circular road called Rocade forms a bypass around the city. However, the bypass does not fulfill entirely its function consisting in absorbing the traffic outside the city centre (fig. 4.24) because the western part of the bypass (A16) is charged. This causes a high traffic inside because the drivers tend to avoid the charged section of the bypass. Two times less vehicles are counted on this part. Consequently, some streets allowing crossing the city record a high average annual daily traffic of more than 20 000 cars (Amiens Métropole, 2013).

Figure removed due to possible copyright infringements

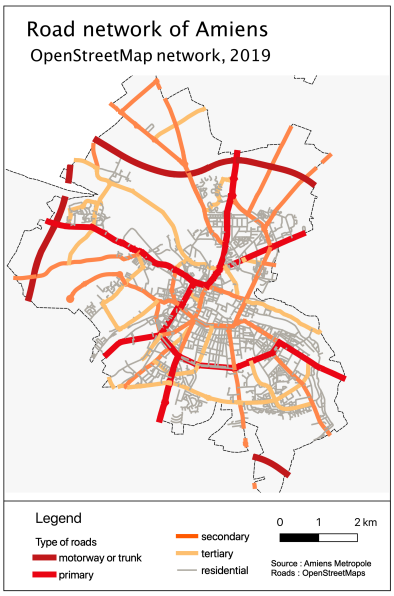

FIGURE 4.24: a: Traffic map translated from Amiens Métropole (2013), b: OpenStreetMap road network (2019)

Thus, traffic volume is an important parameter for Amiens. In this study, the traffic volume information was included by considering the OpenStreetMap hierarchy.

As shown in figure 4.24, the different categories are comparable. Considering the road type has the advantage of being easily reproductible in other locations where traffic volume data are not available.

b. Intersections

Moreover, many large intersections were observed, where turning left can be particularly complicated for cyclists (figure 4.25). Bike markings at pedestrian crossings were installed at some places but these imply waiting several traffic signals to turn left. This safety issue at intersections were also particularly deplored in the online cycling survey (Fédération française des Usagers de la Bicyclette, 2018) as shown in figures 4.26.

FIGURE 4.25: Example of large intersection in Amiens

Figure removed due to possible copyright infringements

FIGURE 4.26: Category concerning safety (Fédération française des Usagers de la Bicyclette, 2018)

c. Bicycle facilities

During the field visit, particular attention was also paid to cycling infrastructures. In Amiens, cycling network is highly discontinuous and in many places confusing. It is composed above all of bike lanes that are not separated from the traffic. Very few bike paths were observed and there are most of the time located along the river and have above all a recreational purpose. There are two common types of bike lanes in Amiens, either shared with cars or with buses. Another very common type of infrastructures are contraflow bike lanes. These are bike lanes that can be taken in the opposite direction by bicyclists (figure 4.27). Contraflow bike lanes enable many shortcuts. However, they create additional conflict situation with cars at intersections. Reaching these links may be very difficult and dangerous for cyclists, especially when they require crossing car lanes. Some colorful parts were installed to overcome this issue but are not yet widespread everywhere.

Thus, three attributes regarding cycling infrastructures were included in the study: the proportion of bike path, bike lane and contraflow bike lanes. For this, data from

FIGURE 4.27: a: bike lane shared with car, b: with bus, c: contraflow bike lane

FIGURE 4.28: Conflicts at intersections due to contraflow bike lanes

GeoVelo (2019) were used and are shown in figure 4.29. For easiness, the current cycling network of 2019 was considered and the changes made in the network are neglected.

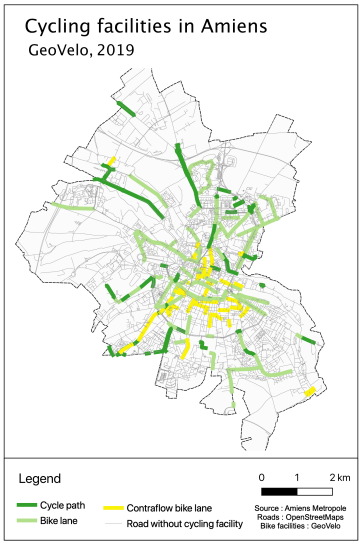

FIGURE 4.29: Cycling facilities (GeoVelo, 2019)

d. Land-use

Finally, the visit of Amiens helped to quantitatively translate the notion of pleasant landscape. Three main pleasant type of land-use were identified: water areas, green areas and the pedestrian city center (figures 4.30 and 4.31).

Firstly, the field trip revealed that the areas closed to **water** were especially attractive. Amiens is crossed by the river Somme and the network of canals has always been an important asset for the city. It led to the construction of textile mills and the installation of draperies and dyes. The textile industry was the main activity of the city until the 20th century. Today, the canals of Amiens and especially the district Saint Leu with its colorful houses and its numerous restaurants are the historical heart of Amiens. Thus, including the presence of water seems relevant for our study.

Amiens is also famous for its **green areas** and especially for the hortillonages (floating gardens), located the east of the city and and in the neighboring municipalities. These are 300 hectares of marshland surrounded by a grid network of man-made canals. These market gardens have been cultivated since the middle age. Today only 7 market gardeners remain, the rest of the area is own for leisure purposes. These market gardens have been cultivated since the middle age. A long path called Chemin de Hallage is running along the hortillonage and is especially appreciated by cyclists and pedestrian. Several other green areas are spread throughout the cities.

Finally, a last important element concerning the land-use can be mentioned. There is a **large pedestrian zone** in the city center of Amiens, where cyclists are also allowed to ride.

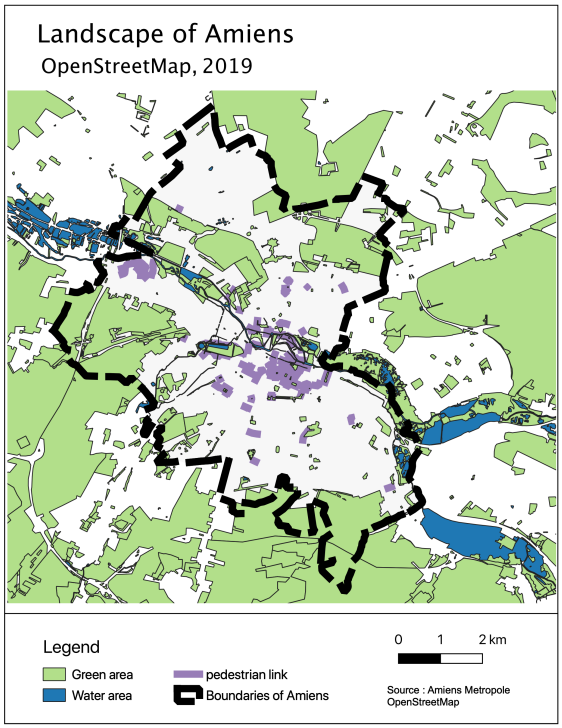

FIGURE 4.30: Land-use of Amiens

Figure removed due to possible copyright infringements

FIGURE 4.31: Example of green and water areas in Amiens

4.5.2 Considered attributes

Given the particularities of Amiens, the existing literature and the available data, the attributes in table 4.9 were investigated. In addition to these attributes, more detailed factors were tested such as the proportion of a given type of road with and without bike path and the proportion of minor road taken in the opposite direction.

Variables	Description
ROAD NETWORK ATTRIBUTES	
Length	Length of the trip (in km)
Cycling facilities	*(GeoVelo, 2019)*
Proportion of bike paths	Proportion of route on off-street bike path, separated from the traffic
Proportion of bike lanes	Proportion of route on an on-street lane, dedicated for bicycles and marked with paint or on shared bus lane
Proportion of contraflow bike lanes	Proportion of route on on-road painted lane added to one-way street to allow cycling in the opposite direction of all other traffic
Proportion of link types	*(OpenStreetMap contributors, 2019)*
Proportion of primary roads	Major roads intended to provide large-scale transport links within or between areas
Proportion of secondary roads	Roads supplementing main roads, usually wide enough and suitable for two-way
Proportion of minor roads	Residential roads or unclassified roads
Proportion of pedestrian street	Proportion of route on pedestrian streets
Traffic signals density	*(OpenStreetMap contributors, 2019)*
Traffic signals	Number of traffic signals per km
Intersections	Number of intersections per km
Crossings	Number of pedestrian crossings per km
Roundabouts	Number of roundabouts per km
Left turn	Number of turning movements per km at an angle between 60° and 179°
Comfort	
Gradient	Maximum or mean gradient on the route from Google Elevation Model
LAND-USE ATTRIBUTES	*(OpenStreetMap contributors, 2019)*
Design	
Trees	Number of trees per km in 25m buffers
Green Areas	Proportion of green areas in 50m buffers, including parks, forests and allotments
Water	Proportion of water in 50m buffers, including rivers, wetlands and lakes
Bridge	Number of bridges per km
Landmarks	Number of landmarks per km in 50m buffers, including churches, memorial, monuments and museums.
Density	
Amenities	Number of amenities per km in 25m buffers (bakery, restaurant, shop, theater, school ...)
Buildings	Proportion of buildings in 100m buffers

TABLE 4.9: Considered attributes

Different sizes of buffers were used to calculate the land-use attributes. A small buffer of 25m was used for characteristics that are expected to influence the perception of a route if they are located really close to the road (amenities, trees) and a bigger buffer of 100m for other areas like green or water areas that can be seen from further away. Table 4.10 summarizes the different steps performed to calculate the attributes. Most of them required the use of buffers, intersections and counting functions in QGIS. R was used for some additional operations.

Variables	**Methodology**
Road network attributes	
Link types	Intersection between routes and network edges (QGIS)
Node attributes	Intersection between routes and network nodes (QGIS)
Turning Left	- Extract vertices (to identify the order of nodes) (QGIS) - Calculate angle between two consecutive nodes (R) - Calculate angle between two consecutive links (R)
Grade	- Extract vertices (to identify the order of nodes) (QGIS) - Intersection with network nodes (QGIS) - Calculate gradients (R)
Contraflow links	- Extract vertices (QGIS) - Intersection with network nodes (QGIS) - Identify direction taken (R)
Land-use attributes	
Amenities, Landmarks	- Create Buffer around routes - Count points in polygon (QGIS)
Green, Water, Building	- Create Buffer around routes - Intersection between green, water or buildings and buffers - Calculate areas (QGIS)

TABLE 4.10: Methodology for attribute creation

4.6 Model estimation

Multinomial logit models were estimated using mlogit package (Croissant, 2019) in R. As mentioned in the literature review in part 2.3, a path-size logit factor (PS) can be included in the utility of a path to take the correlation among alternatives into account. The probability of choosing route k is defined by equation 4.3.

$$P_k = \frac{\exp(V_k + \ln PS_k)}{\sum_{i \in C} PS_i \exp(V_i + \ln PS_i)} \tag{4.3}$$

where C is the choice set of paths, V_k and V_i are the deterministic utilities of routes k and i, respectively. V_k can be written as $\beta.X$, where X is a vector of route attributes and β a vector of coefficients to be estimated. Their values should maximize the likelihood that the process described by the model produced the data that were actually observed.

The path-size factor was calculated for each different route in R with equation 4.4 proposed by Ben-Akiva and Bierlaire (1999).

$$PS_k = \sum_{a \in \Gamma_k} \frac{l_a}{L_k} \frac{1}{\sum_{i \in C} \delta_{ai} \frac{L_C^*}{L_i}} \tag{4.4}$$

with Γ_k the set of links in route k, l_a length of link a, L_k length of route k and L_C^* the length of the shortest path in C. δ_{ai} is one if link a is part of path i and zero otherwise. It reduces the utility of overlapping alternatives.

Before estimating a model, a correlation analysis was performed on the data. In discrete choice modeling, high correlations between attributes must be avoided. Thus, if two attributes were highly correlated, only one of them was kept in the model estimation. All attributes were first included in the model estimation and insignificant or correlated attributes were removed step-by-step based on the log likelihood, p-values and AIC. The last indicator refers to the Akaike information criterion and includes a penalty that is an increasing function of the number of estimated parameters. It aims to avoid overfitting. The model results estimated with 2,362 trips are presented in chapter 5.

Chapter 5

Analysis and discussion

This chapter presents the results of this study. First of all, part 5.1 analyzes how the actual routes differ from the shortest paths. Then, a discrete choice model is estimated in part 5.2 to better understand the influence of each attribute. Finally, the results are discussed and compared with other bicycle studies found in the literature in section 5.2.3.

5.1 Comparison between the actual and the shortest paths

5.1.1 Trip Distance

Trip distance statistics are summarized in table 5.1 and present the comparison between actual routes and shortest routes generated in a car network. T-test is applied to see if the actual routes lengths are significantly different from the shortest ones. T-test relies on the assumption that both samples are random, independent, and normally distributed with unknown but equal variances. The results below show a p-value < 2.2e-16 supporting the alternative hypothesis that true difference in means is not equal to 0 and that the route lengths are significantly different. The route actually taken was on average 0.89 km longer than the modeled shortest route [95% confidence interval (CI): 0.75 to 1.0].

Route	Median	Mean	Range	Standard Deviation (SD)	Mean Difference	95% CI	P-value
Actual route	3.6	4.3	0.37-16	2.8	0.89	0.75-1.0	< 2.2e-16
Shortest route	3.1	3.4	0.087-10	2.0			

TABLE 5.1: Distances of shortest and actual routes (in km)

The ratio of actual trip distance to shortest trip distance is shown in figure 5.1. This ratio can be interpreted as a detour factor. In other words, a ratio higher than 1 occurs when bicyclists use a path longer than the shortest one. A ratio smaller than 1 results from the use of sidewalks, streets in the opposite direction or informal shortcuts. The average ratio is 1.34 which means the actual trip was on average 134% of the distance of the computer-generated shortest path (table 5.2).

Moreover, it is also interesting to mention that the modeled shortest route in a car network is on average 0.34 km longer than the observed shortest route (95% CI: 0.22 to 0.45). However, this study used the model shortest route and not the observed shortest path as a reference route, because it provides a fixed comparaison point that does not depend on the collected dataset and for which the traffic regulations are respected.

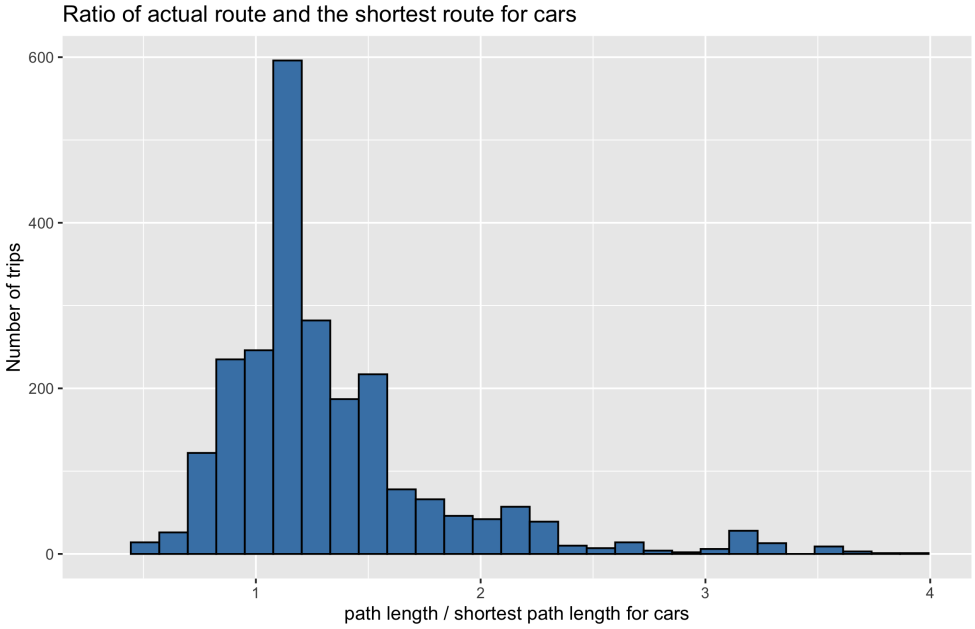

FIGURE 5.1: Shortest path ratio

	Median	Mean	Range	SD	95% CI
Detour factor	1.2	1.3	0.32-9.5	0.55	1.32-1.36

TABLE 5.2: Analysis of the detour factor (ratio of actual route and the shortest route

Thus, an important willingness for detours can be observed in the data. This trend was also observed in other studies. However, the deviation from the shortest path observed in Amiens (1.34) is significantly higher than in the literature (1.09 in Winters et al. (2010), 1.11 in Broach et al. (2012)). The reasons for detours can be multiple, including the lack of acceptable infrastructures in the area or the preference for a given land-use type. The following section aims at investigating the possible explanatory factors for such detours.

5.1.2 Road attributes

Table 5.3 shows the comparison between the means of the attributes for the actual and the shortest route.

The results clearly highlight the preferences for bike facilities. Whereas the shortest route in a car network predicted that 18% of the trip is along bike facilities, the actual route is at 25% along bike paths, bike lanes or contraflow bike paths (95% confidence interval for the mean difference: 6.5% to 8.7%). Another important difference was in the appeal for low traffic streets. Thus, actual bike trips had significantly less proportion along primary (-4.1% to - 2.5% with a 95% confidence level) and secondary roads (-13% to - 10% with a 95% confidence level) than predicted by shortest-route models and significantly more along pedestrian streets (+8.5% to +9.9% with a 95% confidence level). Moreover, the results show that cyclists tend to choose routes with fewer intersections (signalized, non-signalized, roundabouts) and turning left movements than in the shortest routes.

Attributes	Actual Route		Shortest route	T-test
	SD	*mean*	*mean*	*p-value*
propBikePath	0.086	3.6%	1%	***
propBikeLane	0.16	17%	15%	***
propContraflowBikeLane	0.058	4.9%	0.0%	***
propPrimaryRoad	0.11	4.9%	8.2%	***
propSecondaryRoad	0.19	15%	26%	***
propMinorRoad	0.26	53%	53%	
propPedestrianStreet	0.17	9.2%	0.0%	***
nbSignalsPerKm	0.86	0.95	5.8	***
nbIntersectionsPerKm	3.9	12	41	***
nbCrossingsPerKm	1.2	1.0	3.2	***
nbRoundaboutsPerKm	0.37	0.20	0.67	***
nbLeftTurnsPerKm	0.59	2.4	1.9	***
maxGrade	0.025	5.0%	5.9 %	***
meanGrade	0.012	-0.014%	0.037 %	.

TABLE 5.3: Comparison between the actual and the shortest route significance level p: <0.0001 '***', <0.001 '**', <0.01 '*', <0.05 '.' , <1 ' '

5.1.3 Land-use attributes

Concerning the land-use characteristics, seven attributes were evaluated but five from them show significant differences between the actual and the shortest routes (table 5.4). A higher proportion of green, water, bridges and denser areas are preferred. However, the number of trees along the road shows surprising results with a smaller proportion for actual routes than for shortest routes but the mean values are very small in both cases.

Attributes	Actual Route		Shortest route	T-test
	SD	*mean*	*mean*	*p-value*
nbTreesPerKm	2.7	1.6	2.3	***
propGreen	0.060	4.9%	4.0%	***
propWater	0.042	2.0%	1.5 %	***
nbBridgesPerKm	0.00058	0.00044	0.00021	***
nbSightsPerKm	0.36	0.32	0.34	
nbAmenitiesPerKm	20	18	18	
propBuildings	0.10	30%	29%	***

TABLE 5.4: Comparison between the actual and the shortest route significance level p: <0.0001 '***', <0.001 '**', <0.01 '*', <0.05 '.' , <1 ' '

Thus, a first analysis comparing the observed routes with the shortest route already give some insights about how cyclists choose their routes. However, a more detailed analysis is required to understand the influence of each parameter. For this, a discrete choice model is estimated in the next section.

5.2 Discrete choice model

The results of the discrete choice models are presented in this section. They are discussed and compared with other studies.

5.2.1 Correlation analysis

The correlation matrix is presented in figure 5.2. Positive correlations are displayed in blue and negative correlations in red. The intensity of the color and the size of the circles are proportional to the correlation coefficients.

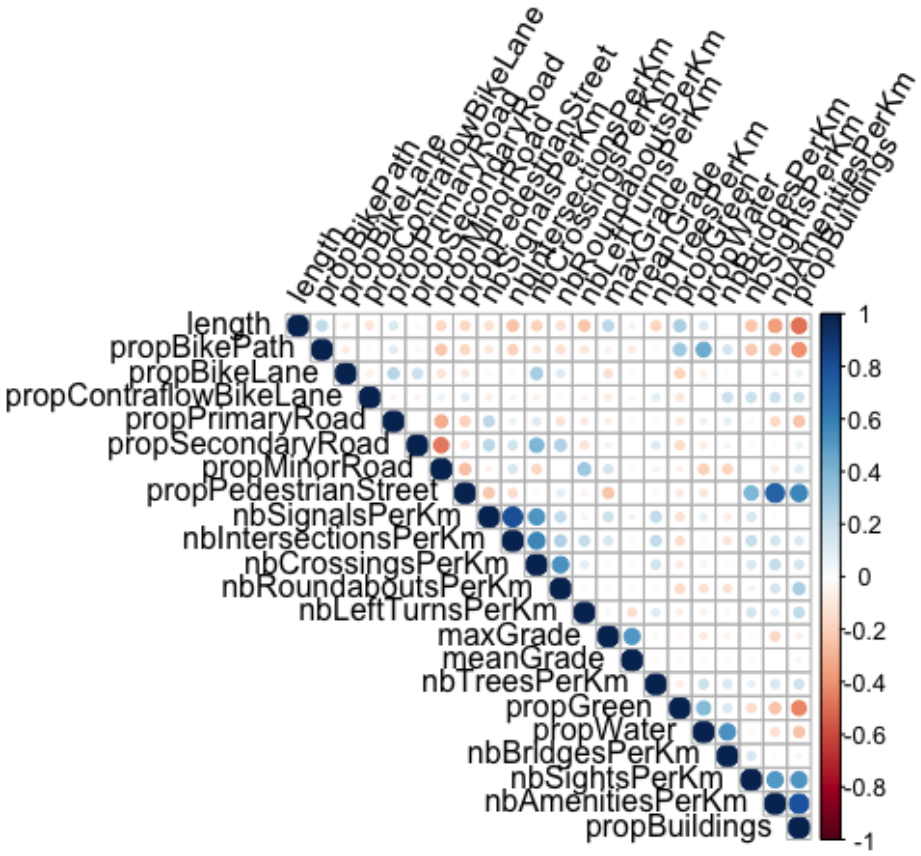

FIGURE 5.2: Correlation between the attributes

Several groups of attributes presenting a correlation higher than 0.5 are identified. The attributes in italics have a correlation coefficient between 0.5 and 0.6.

- Group 1: Proportion of pedestrian streets + Number of amenities per km + Proportion of buildings + *Number of landmarks per km*

Correlation among the variables of group 1 is easily understandable. It can be explained by the concentration of amenities, buildings, and sights in the city center, where many of the streets are reserved for pedestrians or bicyclists (figure 5.3).

FIGURE 5.3: City center with pedestrian streets and a high number of amenities and sightseeings

- Group 2: Number of traffic signals + Number of intersections + *Number of crossings*

The number of traffic signals is highly correlated with the number of intersections (0.80). Another trial was done by replacing the number of intersection by the number of intersections without traffic signals, but the correlation value remains high (0.71). Thus, routes with many non-signalized intersections tend to have many traffic signals as well. Finally, the proportion of water along the route and the number of bridges crossed on the one hand and maximum and mean grade on the other hand are as expected highly correlated.

- *Group 3: Proportion of water + Number of bridges per km*
- *Group 4: Mean grade + Max grade*

Finally, the proportion of water along the route and the number of bridges crossed on the one hand and maximum and mean grade on the other hand are, as expected, highly correlated. To overcome this issue, only one attribute of each group was kept. Only the group concerning the intersections had a different treatment because the number of intersections with and without traffic signals provides different information. It was decided to keep the number of intersections without traffic signals in the model but to add an interaction term with the number of traffic signals. Different combinations were tested to choose the model with the highest log-likelihood and the highest significance of the coefficients.

5.2.2 Model results

The best path-size logit model in terms of significance of the coefficients, sign coherence and log-likelihood is shown in table 5.5. The coefficient associated with the logarithm of the path size variable has as expected a negative sign to decrease the utility of similar alternatives. It corrects for route overlap.

Attributes	Estimate	Std. Error	z-value	p-value	
length (km)	-0.181	0.0374	-4.844	1.28e-6	***
propBikeLane	0.632	0.201	3.148	1.65e-3	**
propContraflowBikeLane	8.08	0.651	12.4	< 2.2e-16	***
propPrimaryRoad	-2.02	0.316	-6.39	1.65e-10	***
nbIntersectionsNoSignalsPerKm	-0.0695	0.00989	-7.03	2.12e-12	***
nbIntersectionsNoSignals*nbSignals	-0.0237	0.00288	-8.15	4.44e-16	***
nbLeftTurnsPerKm	-0.566	0.0633	-8.95	< 2.2e-16	***
meanGradeInPercent	-0.131	0.0423	-3.10	1.95e-3	**
propWater	6.09	1.06	-5.74	9.24e-09	***
propGreen>36%	2.44	0.982	2.49	1.29e-2	*
nbAmenitiesPerKm	0.0147	0.00302	4.87	1.09e-06	***
log(path-size)	-1.03	0.0771	-13.5	< 2.2e-16	***

TABLE 5.5: Model estimation
significance level p: <0.0001 '***', <0.001 '**', <0.01 '*', <0.05 '.' , <1 ' '

Log-Likelihood: -3916, AIC: 7855

All parameters estimated have the expected sign. Cyclists prefer shorter routes with fewer intersections and fewer turning left movements. They are also highly put off by important slopes and prefer to ride on bike lanes. Contraflow bike lanes are especially attractive for cyclists in Amiens. Finally, in terms of land-use, the number of amenities along the route and the proportion of route along the river are important parameters that influence positively the route choice.

5.2.3 Analysis of the coefficients

a. Trip length

Trip **length** is as expected negative, on average cyclists prefer to ride on shorter routes. However, the disutility of riding one additional kilometer is rather low compared to other studies (Menghini et al., 2010, Hood et al., 2011, Broach et al., 2012). Ton et al. (2018) explained that data-driven choice sets provide lower parameters than link elimination methods or labeling approaches. The reason is the lower variability in the choice set because no irrelevant routes are included. All the routes were selected at least once. The absence of unrealistic route with very high distances can cause a lower model fit and lower parameter values. For example, Ton et al. obtained a coefficient for the distance 1.5 times higher in absolute value for link-elimination choice set and 8.2 times more for the labeling method. In our case, a data-driven approach was combined with a labeling method but despite the enrichment of the choice set, the parameters estimated remain low. The coefficient for trip length in this study (-0.18) is of the same order of magnitude of the one obtained by Ton et al. (2018) (-0.23) in their data-driven study.

b. Road attributes

There are different types of **bike facilities** in Amiens : bike lanes shared with cars or with buses, contraflow bike lanes and bike paths (figure 5.4).

Figure removed due to possible copyright infringements

FIGURE 5.4: a: bike lane, b: contraflow bike lane, c: bike path

Bike facilities are all associated with positive signs, showing that cyclists prefer routes with more bicycle facilities. The variable related to the proportion of route on a bike path was not found significant. This can probably be due to the relatively small amount of bicycle paths in Amiens that are reserved for bicycles and separated from the traffic. On the contrary, **contraflow bike lanes** and **normal bike lanes** are significant parameters to explain the route choice. Contraflow bike lanes are especially attractive and are even associated with a higher coefficient than traditional bike lanes. Contraflow bike lanes are 13 times more attractive than normal bike lanes. This result might seem surprising at first sight but is easily understandable by the important advantages of such facilities. They allow cyclists to use streets in the opposite direction and provide important shortcuts in the city center where most of the roads are one-lane streets. The presence of contraflow bike lanes is one of the main strategies of the cycling plan of the municipality to encourage cycling. This result is also consistent with the choice set composition. All the new routes added to the data-driven choice set were generated with a network respecting the street directionality, and only the chosen route could use streets in the opposite direction. It explains the significantly high parameter estimate for contraflow bike lanes.

Concerning the road type, three attributes were tested: the proportion of primary, secondary or minor road (figure 5.5). Only the parameter associated with the proportion of **primary roads** was significant and has a negative sign. It proves that this road type with high traffic volume and many lanes are avoided by cyclists. The current model shows that to compensate for one additional percent on a primary road, 3.2 percent needs to be traveled on a bike lane or 0.25 on a contraflow bike lane. The proportions of minor roads and secondary roads were not significant. Concerning the minor roads, the non-significance of the parameter can be due to the high proportion of residential streets in Amiens. The differences in means for the observed paths and the shortest paths were also not significant when applying a t-test in the previous section. On average, 53 % of the observed routes are on minor roads and the standard deviation is relatively small (0.26). Thus, it is not a decisive criterion for route choice in Amiens. It would have been interested to have access to traffic volume data to enable a better and more precise distinction between the different streets of Amiens. Moreover, other more detailed parameters were tested

such as the proportion of a given link type with and without bike facilities in order to investigate the impact of bike facilities according to the link type but the results were not significant. A bigger data set is necessary to perform such a detailed analysis.

Figure removed due to possible copyright infringements

FIGURE 5.5: a: primary road, b: secondary road, b: minor road

Finally, the coefficient associated with the number of **intersections without traffic signals** was highly negative indicating that cyclists prefer routes with long streets and fewer conflict points. **Turning left** at an intersection is perceived 8.1 times more painful than crossing a junction without traffic signal. The number of traffic signals was included as an interaction term with the number of intersection without traffic signals. The negative coefficient shows that bicyclist try to avoid routes with many intersection with and without traffic signals.

The **slope** also affects the way cyclists perceive the route. Several parameters were tested for the gradient: the maximum gradient, the mean gradient, and the mean up-slope. The mean gradient showed higher significance and a negative sign as expected. Cyclists are highly discouraged by important grades.

c. Land-use attributes

After eliminating the non-significant and highly correlated attributes, four parameters concerning the **land-use** remain: the **proportion of water**, the proportion of green along the route and the number of amenities per kilometers. The first variable indicates that routes along water are very attractive. This result is not surprising given the characteristics of Amiens. Amiens is crossed by the river Somme and the river is an important part of its identity. Many streets run along the river and offer a pleasant perspective on the Somme.

The second attributes related to the **proportion of green** along the route showed, first of all, a surprising result with a negative sign. It meant that routes along green areas are less attractive. This result was not in line with previous studies (Chen et al., 2018, Ghanayim and Bekhor, 2018). Therefore, the definition of this attribute was slightly changed. The variable was set to 0 when the proportion of green was less than 36%. This percentage was the lowest value that results in an estimated parameter with a consistant sign and a good significance. With this variable, it is considered that green areas have a significant influence only when the proportion along the road is high. This variable definition is consistent with the characteristics of Amiens. There is a big heterogeneity in the park sizes. In the city center, most of

green areas consist of small parks distributed everywhere. Due to their small size, going along them does not probably worth the detour. However, bigger parks like les hortillonages (at the east of the city) and Parc-Saint Pierre at the west may be very attractive. This new variable definition was a success and a positive sign was obtained for the proportion of green areas along the route. It was obtained that going 1% along green areas is 2.5 times less attractive than going 1% along water.

Finally, the **number of amenities** per kilometer shows a positive sign, which indicates that dense areas with many shops and restaurants are very attractive. This parameter among the correlation group 1 with the proportion of buildings, the proportion of pedestrian streets and the number of landmarks is the one that showed the higher significance and the highest likelihood for the resulting model. 7.4 amenities per km can compensate for one intersection.

5.2.4 Analysis of marginal rates of substitution

a. Marginal rates of substitution

In this part, the model coefficients are first compared with each other. The equivalent percent changes of different attributes that are obtained when the trip length is increased by one kilometer are calculated in 5.6. For example, it was found that the disutility of one additional kilometer can be compensated if 29% of the route is ride on a bike lane. However, the effects of an additional kilometer must be studied carefully because if a cyclist ride an extra kilometer, the proportion of route on the different type of links can change too, as well as the number of a given facility per kilometer. This issue appears because the attributes are divided by length for the model estimation. Nevertheless, to facilitate the interpretation of the results, it is considered in the following that the other attributes remain unchanged.

Table 5.6 also shows additional marginal rate of substitution (MRS) when the number of intersections or the proportion of primary road is increased by one unit. For example, the model predicts that the disutility of 1% on a primary road can be compensated by 3.2% on a bike lane and one additional intersection per kilometer can be compensated by 11% on a bike lane.

MRS	Present study
$\frac{\beta_{length}}{\beta_{propBikeLane}}$ [%/km]	-28.7
$\frac{\beta_{length}}{\beta_{propContraflowBikeLane}}$ [%/km]	-2.24
$\frac{\beta_{length}}{\beta_{propPrimaryRoad}}$ [%/km]	8.97
$\frac{\beta_{length}}{\beta_{nbIntersectionsNoSignalsPerKm}}$ [1/km /km]	2.61
$\frac{\beta_{length}}{\beta_{nbLeftTurnsPerKm}}$ [1/km /km]	0.320
$\frac{\beta_{length}}{\beta_{meanGrade}}$ [%/km]	1.38
$\frac{\beta_{length}}{\beta_{nbAmenitiesPerKm}}$ [1/km /km]	-12.3

MRS	Present study
$\frac{\beta_{propPrimaryRoad}}{\beta_{propBikeLane}}$ [-]	-3.20
$\frac{\beta_{propPrimaryRoad}}{\beta_{propContraflowBikeLane}}$ [-]	-0.250
$\frac{\beta_{nbIntersectionsNoSignalsPerKm}}{\beta_{nbAmenitiesPerKm}}$ [-]	-7.43
$\frac{\beta_{nbIntersectionsNoSignalsPerKm}}{\beta_{nbLeftTurnsPerKm}}$ [-]	0.123
$\frac{\beta_{nbIntersectionsNoSignalsPerKm}}{\beta_{propBikeLane}}$ [% / (1/km)]	-11.0
$\frac{\beta_{nbIntersectionsNoSignalsPerKm}}{\beta_{nbAmenitiesPerKm}}$ [-]	-4.71

TABLE 5.6: Marginal rates of substitution

b. Comparison with other studies

Then, the results are compared with other bicycle studies in figure 5.7. Trade-off values between the trip length and the proportion of route on a bike facility vary a lot in the different studies found in the literature. The MRS obtained by Hood et al. (2011) is 52 times higher than in Casello and Usyukov (2014). In the present study, the trade-off value between distance and bike lane (-22) is closed to the average value found in the literature (-30). These important differences in model estimations show that the parameters highly depend on the location of the case study and on the choice set.

	Present study	Menghini et al., 2010	Hood et al., 2011	Broach et al., 2012	Casello and Usyukov, 2014	Ton et al., 2017	Chen et al., 2018
$\frac{\beta_{length}}{\beta_{propBikeLane}}$ [%/km]	-22		-78		-4.2	-1.5	-36
$\frac{\beta_{length}}{\beta_{propBikePath}}$ [%/km]		-2.3	-89	-3.1			
$\frac{\beta_{length}}{\beta_{propContraflowBikeLane}}$ [%/km]	-1.7						

TABLE 5.7: Comparisons of marginal rates of substitution

Other parameters commonly included in the literature are traffic volume and gradient. Traffic volume was not directly included in our study because of a lack of data and was replaced by the road hierarchy of OpenStreetMap that gives an indication about the traffic but it prevents the comparison with other studies. In the literature, grades are considered with various forms: mean gradient, maximum gradient, average up-slope or as a categorical variable. The trend observed is consistent with our observations: cyclists prefer to avoid important slopes but a more detailed comparison of coefficients among studies is delicate due to the different type of gradient variables included in models and the different topographies among case studies.

For the rest of the parameters, only Chen et al. (2018) included such a large variety of land-use variables. Our results are coherent with their study in terms of sign coefficients. More other studies including land-use parameters are required to compare the results, but as for the slope, the land-use is extremely dependent from the city characteristics and comparisons between models are difficult.

5.2.5 Limitation: overrepresentation of trips

Some routes are included many times in the dataset. On average a route is repeated 2.7 times and the maximum number of repetitions is 51. In total, 14 routes are repeated more than 20 times. An overrepresentation of one route selected by the same user is problematic because the estimated model will be highly influenced by the attribute characteristics of the chosen route. However, because of data privacy, it was not possible to identify if the repeated trips were done by the same person. Nevertheless, analyzing the socio demographic data and the time of the trips could help to make a hypothesis on the number of travelers that made these repeated trips. This analysis revealed that 9 out of 14 routes with more than 20 trips seem to come from the same users.

Several attempts have been made to solve this problem but without achieving satisfactory results. The idea was to reduce the weight of the overrepresented trips. Trips repeated more than 15 times were investigated in detail and a hypothesis was made on the number of different persons that traveled. This hypothesis was based on two elements: the socio-demographic data, when they were provided, and the time of the trips. For example, if all the repeated trips were done by persons with the same date of birth, it is very likely that they were done by a single person. However, It is impossible to analyze all the trips in detail and defining weights and thresholds are arbitrary decisions. This method added a certain level of uncertainty and was not improving the model. Thus, the overrepresentation of certain persons is one of the limitations of this study but is difficult to control without a person ID.

Chapter 6

Conclusion

6.1 Main contributions

This thesis presents the findings of a bicycle route choice model estimated for the city of Amiens. The objective was to investigate the effects of the built environment on bicyclists' preferences. It succeeds in evaluating the impacts of a variety of land-use characteristics in addition to the commonly included road features. Thus, green, water areas and amenities along the route are shown to be attractive for bicyclists. In terms of methodology, this study adopted an innovative approach with a data-driven method to generate the choice set, as proposed recently by Ton et al. (2017). It is a very promising method because it gathers real, observed routes between a given origin-destination pair and does not require an artificial enumeration of possible paths. In this study, the choice set was enriched with labeled routes (Ben-Akiva et al., 1984). By doing so, limitations of both techniques can be overcome: the low variability of the data-driven method due to the relatively limited number of cyclists that took part in the challenge, and the low number of labeled routes due to the difficulty of defining relevant objective functions. Another particularity of this study is the data source for the attribute calculation. All the considered attributes come from an open-data source, mainly OpenStreetMap. Therefore, the methodology can easily be reproduced in other cities.

6.2 Limitations and further research

However, the model is based on several assumptions and the different steps of the methodology have important limitations that will require further research.

Firstly, concerning the data filtering process, the study focuses on trips within Amiens but to decrease the data loss, trips that were at least 70% within Amiens were also kept. They were cut at the city boundary and included in the model. By doing so, it was assumed that the route choice for 70 % of the trip does not significantly differ from the choice for the entire route. A second assumption was made during the clustering of trips with similar origin-destination pairs. It was considered that small differences at the origin and the destination only have a minor impact on the resulting route attributes.

Secondly, the map-matching process also had some limitations. The method, which consists of searching for the shortest path in a subnetwork, provides good results but is very demanding in terms of network quality. However, bicyclists have a flexible behavior and the digital network representation does not take this into account. This issue led to an important data loss: 18% of the filtered data were not properly matched. Further research on how the network could be adapted to bicyclists, especially by a data-driven method, is an interesting area of investigation.

For the choice set generation, another important limitation is the overrepresentation of certain persons in the data. As the trips were not linked by a person ID, correcting for this issue is difficult and future research on another dataset that includes this information is necessary.

Concerning the attribute creation, several aspects can be mentioned. First of all, changes in Amiens network from 2016 to 2019 were not taken into account in this thesis due to the time constraints of this project. Furthermore, it would be interesting to include other variables in the model, such as socio-demographic characteristics, road pavement, road width and the number of lanes. Unfortunately, the socio-demographic variables were not available for all the trips and considering only those trips would have caused an important data loss. Moreover, in Amiens, many land-use attributes such as the number of landmarks, the proportion of buildings, and the proportion of pedestrian areas were highly correlated and it was not possible to estimate their effects. A model estimation in another city may provide additional results.

Finally, the discrete choice model was based on a multinomial logit formulation with a path-size factor. This additional term is not able to capture all the correlation among alternatives, more complex models, such as the cross-nested logit model or a multinomial probit model, could be estimated in further research. The recently developed recursive logit model that does not require generating the choice set would also be interesting to consider but is very computationally expensive.

6.3 Recommendations

Despite these limitations, the estimated model provides significant insights into the preferences of bicyclists for choosing their route. The model can be used by the municipality to improve the existing infrastructures. Based on the results, cyclists prefer bicycle facilities and avoid streets with high traffic. In addition, contraflow bike lanes that allow cyclists to ride in the opposite direction are especially attractive because they offer shortcuts. Therefore, the efforts of the municipality towards bike facilities must be continued. Creating these facilities along green and water areas or in the historical center, where many amenities are concentrated, is relevant. Another important aspect affecting bicycle route choice are intersections, they have a very negative impact. In Amiens, there are many large intersections, making it difficult or cyclists to cross and turn left. Moreover, intersections linking contraflow bike lanes require special attention. For example, colorful painting on the road could be generalized to enable bicyclists to reach and leave safely these contraflow links. Thus, this study has highlighted several actions that could be taken by the municipality to improve cycling conditions in Amiens. Adapting the road network to bicycle users is a huge opportunity to create a more sustainable transportation system.

Bibliography

Aduga (2012). *Portrait de territoire, Amiens Métropole*. Tech. rep. `http://www.aduga.org/index.php?lvl=cmspage&pageid=4&id_article=151`.

Aldred, Rachel et al. (2016). "Does more cycling mean more diversity in cycling?" In: *Transport reviews* 36.1, pp. 28–44.

Amiens Métropole (2013). *Plan de Déplacement urbains Amiens Métropole 2013-2023*. Tech. rep.

Aultman-Hall, Lisa et al. (1997). "Analysis of bicycle commuter routes using geographic information systems: implications for bicycle planning". In: *Transportation research record* 1578.1, pp. 102–110.

Azevedo, JoseAugusto et al. (1993). "An algorithm for the ranking of shortest paths". In: *European Journal of Operational Research* 69.1, pp. 97–106.

Ben-Akiva, Moshe and Michel Bierlaire (1999). "Discrete choice methods and their applications to short term travel decisions". In: *Handbook of transportation science*. Springer, pp. 5–33.

Ben-Akiva, Moshe et al. (1984). "Modeling inter-urban route choice behaviour". In: *Proceedings of the 9th international symposium on transportation and traffic theory*. VNU Press Utrecht, pp. 299–330.

Bernardi, Silvia et al. (2018). "Modelling route choice of Dutch cyclists using smartphone data". In: *Journal of transport and land use* 11.1, pp. 883–900.

Boeing, Geoff (2017). "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks". In: *Computers, Environment and Urban Systems* 65, pp. 126–139.

Bovy, Piet HL and Stella Fiorenzo-Catalano (2007). "Stochastic route choice set generation: behavioral and probabilistic foundations". In: *Transportmetrica* 3.3, pp. 173–189.

Broach, Joseph et al. (2010). "Calibrated labeling method for generating bicyclist route choice sets incorporating unbiased attribute variation". In: *Transportation Research Record* 2197.1, pp. 89–97.

Broach, Joseph et al. (2012). "Where do cyclists ride? A route choice model developed with revealed preference GPS data". In: *Transportation Research Part A: Policy and Practice* 46.10, pp. 1730–1740.

Cascetta, Ennio and Andrea Papola (2001). "Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand". In: *Transportation Research Part C: Emerging Technologies* 9.4, pp. 249–263.

Cascetta, Ennio et al. (1996). "A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks". In: *Transportation and Traffic Theory. Proceedings of The 13th International Symposium On Transportation And Traffic Theory, Lyon, France, 24-26 July 1996*.

Casello, Jeffrey M and Vladimir Usyukov (2014). "Modeling cyclists' route choice based on GPS data". In: *Transportation Research Record* 2430.1, pp. 155–161.

Chen, Peng et al. (2018). "A GPS data-based analysis of built environment influences on bicyclist route preferences". In: *International journal of sustainable transportation* 12.3, pp. 218–231.

CIVITAS (2017). *European Cycling Challenge*. https://civitas.eu/event/european-cycling-challenge. Accessed: 2019-05-30.

Copernicus (2012). *CORINE Land Cover*. https://www.statistiques.developpement-durable.gouv.fr/corine-land-cover-0. Accessed: 2019-06-03.

Croissant, Yves (2019). *mlogit: Multinomial Logit Models*. https://cran.r-project.org/web/packages/mlogit/index.html. Accessed: 2019-10-05.

Dalumpines, Ron and Darren M Scott (2011). "GIS-based map-matching: Development and demonstration of a postprocessing map-matching algorithm for transportation research". In: *Advancing geoinformation science for a changing world*. Springer, pp. 101–120.

De La Barra, Tomas et al. (1993). "Multidimensional path search and assignment". In: *PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom*.

Dijkstra, Edsger W (1959). "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1, pp. 269–271.

Drouaults, Kévin (2016). *Approche participative pour la mise en œuvre d'une politique publique. Le cas plan vélo de la MEL*. Tech. rep. https://opendata.lillemetropole.fr/explore/dataset/trajets-challenge-europeen-velo/information/. Institut d'Aménagement et Urbanisme de Lille.

Fédération française des Usagers de la Bicyclette (2018). *Baromètre des villes cyclables - résultats 2017*. https://www.parlons-velo.fr/copie-de-barometre-villes-cyclables. Accessed: 2019-07-20.

Fosgerau, Mogens et al. (2013). "A link based network route choice model with unrestricted choice set". In: *Transportation Research Part B: Methodological* 56, pp. 70–80.

Frejinger, Emma and Michel Bierlaire (2007). "Capturing correlation with subnetworks in route choice models". In: *Transportation Research Part B: Methodological* 41.3, pp. 363–378.

Frejinger, Emma et al. (2009). "Sampling of alternatives for route choice modeling". In: *Transportation Research Part B: Methodological* 43.10, pp. 984–994.

GeoVelo (2019). *Cycling infrastructures in France*. https://www.geovelo.fr/france/route. Accessed: 2019-08-12.

Ghanayim, Muhammad and Shlomo Bekhor (2018). "Modelling bicycle route choice using data from a GPS-assisted household survey". In: *European Journal of Transport and Infrastructure Research* 18.2.

GoogleMaps (2019). *Elevation API*. https://developers.google.com/maps/documentation/elevation/start. Accessed: 2019-09-10.

Hagberg, Aric et al. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hartigan, John A and Manchek A Wong (1979). "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108.

He, Biao et al. (2018). "A simple line clustering method for spatial analysis with origin-destination data and its application to bike-sharing movement data". In: *ISPRS International Journal of Geo-Information* 7.6, p. 203.

Hood, Jeffrey et al. (2011). "A GPS-based bicycle route choice model for San Francisco, California". In: *Transportation letters* 3.1, pp. 63–75.

Hunt, John Douglas and John E Abraham (2007). "Influences on bicycle use". In: *Transportation* 34.4, pp. 453–470.

INSEE (2015). *Observatoire des territoires*. https://www.observatoire-des-territoires.gouv.fr/outils/cartographie-interactive/. Accessed: 2019-06-20.

Kaneko, Noriko et al. (2018). "Route Choice Analysis in the Tokyo Metropolitan Area Using a Link-based Recursive Logit Model Featuring Link Awareness". In: *Transportation research procedia* 34, pp. 251–258.

Kang, Lei and Jon D Fricker (2018). "Bicycle-Route Choice Model Incorporating Distance and Perceived Risk". In: *Journal of Urban Planning and Development* 144.4.

Mai, Tien et al. (2018). "A decomposition method for estimating recursive logit based route choice models". In: *EURO Journal on Transportation and Logistics* 7.3, pp. 253–275.

Menghini, Gianluca et al. (2010). "Route choice of cyclists in Zurich". In: *Transportation research part A: policy and practice* 44.9, pp. 754–765.

OpenStreetMap contributors (2019). *OpenStreetMap data retrieved from https://download.geofabrik.de*. https://www.openstreetmap.org. Accessed: 2019-06-03.

Pays du Grand Amiénois (2013). *Enquête déplacements grand territoire réalisée dans le Grand Amiénois en 2009-2010*. Tech. rep. http://www.aduga.org/doc_num_data.php?explnum_id=7244.

Prato, Carlo Giacomo (2009). "Route choice modeling: past, present and future research directions". In: *Journal of choice modelling* 2.1, pp. 65–100.

Quddus, Mohammed A et al. (2007). "Current map-matching algorithms for transport applications: State-of-the art and future research directions". In: *Transportation research part c: Emerging technologies* 15.5, pp. 312–328.

Ramming, MS (2002). "Network Knowledge and Route Choice [PhD thesis]". In: *Cambridge, USA: Massachusetts Institute of Technology*.

Sener, Ipek N et al. (2009). "An analysis of bicycle route choice preferences in Texas, US". In: *Transportation* 36.5, pp. 511–539.

Ton, Danique et al. (2017). "How Do People Cycle in Amsterdam, Netherlands?: Estimating Cyclists' Route Choice Determinants with GPS Data from an Urban Area". In: *Transportation research record* 2662.1, pp. 75–82.

Ton, Danique et al. (2018). "Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam". In: *Travel behaviour and society* 13, pp. 105–117.

Van Dijk, Justin and Tom De Jong (2017). "Post-processing GPS-tracks in reconstructing travelled routes in a GIS-environment: network subset selection and attribute adjustment". In: *Annals of GIS* 23.3, pp. 203–217.

Véloxygène (2019). *Véloxygène: les aménagements*. https://veloxygene-amiens.com/category/les-amenagements/. Accessed: 2019-06-12.

Vovsha, Peter and Shlomo Bekhor (1998). "Link-nested logit model of route choice: overcoming route overlapping problem". In: *Transportation research record* 1645.1, pp. 133–142.

Winters, Meghan et al. (2010). "How far out of the way will we travel? Built environment influences on route selection for bicycle and car travel". In: *Transportation Research Record* 2190.1, pp. 1–10.

Yai, Tetsuo et al. (1997). "Multinomial probit with structured covariance for route choice behavior". In: *Transportation Research Part B: Methodological* 31.3, pp. 195–207.

Zimmermann, Maëlle et al. (2017). "Bike route choice modeling using GPS data without choice sets of paths". In: *Transportation research part C: emerging technologies* 75, pp. 183–196.