Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Ingegneria Civile, Ambientale e dei Materiali

Ciclo XXVIII

Settore Concorsuale di afferenza: 08/A3

Settore Scientifico disciplinare: ICAR/05

EVALUATING CYCLIST PATTERNS USING GPS DATA FROM SMARTPHONES.

Presentata da: Silvia Bernardi

Coordinatore Dottorato **Relatore**

Prof. Alberto Lamberti **Prof. Federico Rupi**

Esame finale anno 2016

Acknowledgements

My journey through this three-years PhD study has been sometimes challenging, always interesting, and has taught me so much about the research subject, but also about myself. I have grown an invaluable set of skills, both work-related, and personal. I have had the opportunity of travelling a lot, working with many talented researchers and this improved significantly the value of my study and my personal growth.

The first person I feel like saying thanks to is Lissy: during this last year you have been my supervisor, my roommate, my life-coach but most of all my friend. I should thank you for the numerous lessons on transport modelling, but also for those on passion, hard work, and strength. Your example has been inspiring through these final months of work on my thesis.

I would like to say thank you to Karst Geurs, for his kindness, support, encouragement and patience. A cake is not nearly enough to thank you for the opportunity you gave me, not even my chocolate one.

Also thanks to all the CTS people at the University of Twente, you have been extremely welcoming and nice to me, and made those months one of the best memories from this years. A special acknowledgement is for Borja, who taught me tons of ArcGIS secrets, I will never thank you enough for your precious help.

I would like to thank my supervisor Federico Rupi, and all my colleagues from the University of Bologna, for your help and guidance during this experience.

Furthermore, Professor Aura Reggiani, from the Department of Economics of the University of Bologna and Professor Kevin J. Krizek, from the University of Colorado, are greatly acknowledged for their guidance and their precious inputs to my work. The conference time we shared, and the talks, have been greatly inspiring.

On a more personal note, I would like to thank all my friends, I will not name all, but each of you helped me and encouraged me during these three years. Also, I would like to thank Pier Paolo: among all the surprises I found throughout this journey, probably the least expected, certainly the most joyful.

Finally, this work is dedicated to my family, my parents Angela and Claudio and my brother Alessandro: thank you for your loving support during my education and your constant faith in me, without your encouragement nothing I did would have been possible.

Alla mia famiglia.

Part of the contributions of the candidate, and her co-authors, to the study presented in this PhD thesis have already been published as Journal papers, listed below. Data, methodologies and results published in these papers are therefore part of this document, and will be discussed in Chapter 1 and 3.

Journal Paper 1: The paper presented in Chapter 1 is first-authored by the candidate, co-authored by Professor Kevin J. Krizek of the University of Colorado Boulder, and co-authored by Professor Federico Rupi, and is published as:

Bernardi, S., Krizek, K.J., Rupi, F. (2016). Quantifying the role of disturbances and speeds on separated bicycle facilities. Journal of Transport and Land Use, 9(2), In Press.

Journal Paper 2: The paper excerpted from Chapter 3 is first-authored by Dr. Joerg Schweizer, co-authored by the candidate, and co-authored by Professor Federico Rupi, and is published as:

Schweizer, J., Bernardi, S., Rupi, F. (2016). Map-matching algorithm applied to bicycle GPS traces in Bologna. ITE Intelligent Transport Systems, In Press.

Summary

Abstract

In recent years the availability of GPS data has seen as a significant improvement in data accuracy, continuity and quality, due to the spread of smartphones and mobile applications for self-localization and navigation. GPS datasets provide analysts with geo-referenced information about users' mobility and habits.

The first part of the thesis consisted of an analysis of the context of bicycle facilities for the city of Bologna, Italy, made through experimental measures of cycleway and road usage rates, and cycling speed.

The following part of the study focused on the use of GPS traces datasets, which allow to record a wide amount of bike trips. The main advances in the last years' relevant literature have been described. Subsequently, two case studies have been analyzed: GPS traces datasets recorded through mobile devices, both for the city of Bologna and for the whole Netherlands.

First, the original dataset of GPS points have been properly filtered in order to exclude instrumental errors. The GPS points have then been matched to a high-detail network database, in order to obtain the actual routes chosen by cyclists. The considered road networks included both attributes of the roadway and attributes of the bicycle facilities, when existing.

The percentage of the trips done on a bicycle facility versus the roadway could thus be compared with the results from the first part of the study, and be used as a measure of the attractiveness and effectiveness of the bicycle facilities available. Furthermore, the chosen routes could be compared with the shortest routes for each origin-destination pair, and route choice models could be calibrated, based on different relevant attributes of the networks, of users, and of trips.

Abstract in italiano

Negli ultimi anni la disponibilità di dati GPS ha visto un deciso miglioramento in termini di accuratezza, continuità e qualità del dato, anche grazie al diffondersi di smartphone e applicazioni per l'auto-localizzazione e navigazione. I dati GPS forniscono agli analisti preziose informazioni geo-referenziate riguardo agli spostamenti e alle abitudini di mobilità degli utenti.

La prima arte del lavoro di tesi è consistita in un'analisi del contesto ciclabile per la città di Bologna, attraverso rilievi sperimentali dell'uso delle infrastrutture ciclabili, o della carreggiata stradale, nonché delle velocità dei ciclisti.

Successivamente lo studio si è concentrato sull'uso di dataset di tracce GPS, che permettono di registrare una grande quantità di viaggi in bici. Sono stati descritti i princiali avanzamenti della letteratura pubblicata sul tema negli ultimi anni. Successivamente, sono stati presi in considerazione due casi di studio: le tracce GPS regsitrate tramite applicazioni mobile, per la città di Bologna in Italia, e per tutto il territorio dell' Olanda.

Per prima cosa, i punti GPS che costituiscono i dataset originali sono stati opportunamente filtrat per escludere errori strumentali. I punti GPS sono quindi stati attribuiti ad un modello di rete opportunamente dettagliato, ottenendo così i veri e propri percorsi scelti dai ciclisti. I modelli di rete considerati includono sia le caratteristiche delle strade che delle piste ciclabili, laddove presenti.

Le percentuali di utilizzo delle piste ciclabili e della carreggiata stradale hanno così potuto essere confrontate con i risultati ottenuti dalla prima parte dello studio, e utilizzate come misura dell'attrattività e dell'efficienza delle piste ciclabili disponibili nei due casi di studio. Inoltre, i percorsi scelti sono stati confrontati con i percorsi di minima lunghezza calcolati per ciascuna coppia OD, e due modelli di scelta del percorso sono stati calibrati, sulla base di attributi della rete di trasporto, dei ciclisti e dei viaggi.

Introduction

Research background and aims

Over the years, non-motorized travel alternatives, as cycling, have been largely underrepresented in transport discussion both from city-administrators and policy makers, but also from research. Growing interest in sustainable transportation systems has increased the focus on policies and investments that are able to transform cycling in a real travel option, for everyday urban trips and for an increasing number of people belonging to different population segments. However, such investments are often made without being able to carry out a quantitative forecast of the impacts of the investment, thus risking to result in ineffective interventions. In this sense, what needs to be investigated in advance is the level of usage of existing bicycle infrastructures and travel patterns of existing cyclists. By knowing the "state of fact" of cycling mobility in a certain urban context, decision makers can effectively intervene for enhancing it.

The first approaches to the problem were mainly based on measuring the three fundamental traffic quantities – speed, volumes and density. The idea was to apply to bicycle infrastructures the same approach seen for the determination of the Level of Service for road segments. What these studies have determined, is that the same criteria adopted for car mode cannot completely be transferred to bicycle mode, as different are the factors that influence the quality of a trips as perceived by car travelers and cyclists.

As research has been demonstrating in the last decades, cyclists consider a variety of factors when evaluating the quality of their cycling trips in an urban context. Cyclists' travel decisions are often motivated by perceived travel time – and speed – as well as by perceived safety. Both travel time and safety are influenced by the geometric design of the facilities (turning radii, slope, lane width) and functional features (proximity to cars, speed of adjacent cars, proximity of car parking stalls, presence of dedicated facility). Moreover, other factors intervene, less easy to quantify and measure, as land-use, pleasantness of the environment, or perceived quality of a bicycle context.

Thus, researchers tried to develop synthetic indicators of bikeability, or stress level, or level of service, mainly based on stated preference surveys. Stated preference surveys has been largely used for understanding cyclists' preferences, as they constitute a relatively cheap data collection method, and they provide a controlled experimental environment.

Another consistent part of the literature on the subject has relied on revealed preference surveys, where analyst asked participants to describe their travel choices after the travel itself. In order to use this more reliable source of data, more information on the network is needed, as well as specific methodologies for data analysis.

Recently, research on cyclists' preferences has grown thanks to the spread of the Global Positioning Systems (GPS) recording devices. On one side, the level of accuracy of GPS devices has increased, thanks to the investments on the involved technologies. On the other side, people habits and attitude towards the use of smartphone changed very rapidly, to the point that nowadays it is absolutely common for people to possess a smartphone, keep self-localization devices activated, declare their position in order to make it available for a vaste set of applications and services, register, monitor and share with people their position, trips and activities. Data availability for investigating cyclists travel choices and habits has seen an incredible enhancement.

In parallel with these changes in technology and society, research followed: several of the methodological developments relevant to modelling route choices from GPS-based travel surveys are relatively recent.

This work aimed at understanding which are the most critical aspects of cycling in an urban context, and individuating what makes a path more attractive of another for cyclists. To do so, we started to explore some experimental data from Bologna, an Italian medium-sized city that has one of the highest bicycle split rate, considering the Italian context. The first measurements of usage rates and speed patterns on cycleway and road segments of the network of Bologna indicated that cyclists often cycle on the road, mixing with motorized traffic, even if a bicycle facility next to the roadway is provided. This in mainly due to the frequent interruptions that a cyclist encounter along a dedicated bicycle path, and to the presence of disturbances. One of the main factors influencing cyclists' choices are indeed his perception of speed and safety.

The research then moved towards the use of GPS smartphone-based datasets. Such data provides a broader point of view on cyclists' behaviour, as well as more information on the cycling patterns of a much wider sample of cyclists on a much wider portion of the network. The benefits of the availability of such rich travel datasets comes, of course, with the burden of a much more complicated process of data elaboration and analysis. The main processing methodologies for modelling route choice from GPS traces recorded from smartphone have been described, and can generally fall into three main categories: (1) filtering and map-matching procedures, (2) choice set generation and (3) route-choice modelling.

Each of these steps has been considered in this work, and both novel and existing methodologies have been applied to real case studies. The first case study considered is, again, the city of Bologna. A rich database of GPS traces collected through a smartphone application by cyclists, in the framework of the European Cycling Challenge of 2013, has been processed and analysed.

The second GPS-recorded trip dataset available has been collected in the Netherlands, in the framework of the Mobile Mobility Panel, using a dedicated smartphone application. Participants from all over the Netherlands has been registering their bicycle trips for weeks, indicating the trip purpose. The database also comprised personal socio-economical information of participants.

Thesis outline

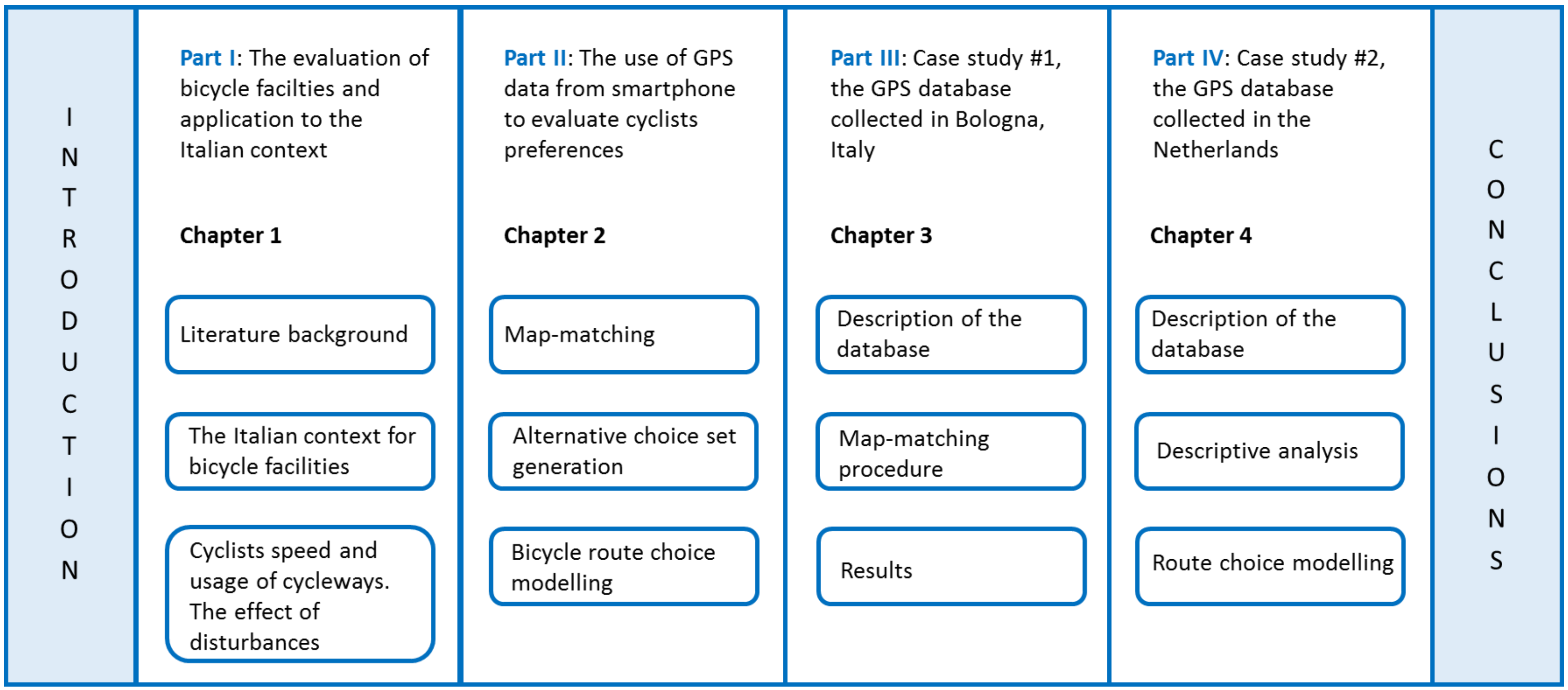
This PhD thesis consists of four main parts, which are illustrated in next page scheme. The first part of the thesis will present an analysis of the context of bicycle facilities for the city of Bologna, Italy. From first Chapter a Journal paper has been published, and the outline of the Chapter reflects the main structure of the paper. The aim of this first part of the study is to shed light on cyclists' usage of the bicycle facilities, in order to describe the specific context of bicycle facilities in Italy.

The second part of the work will focus on the use of GPS data coming from smartphone to investigate cyclists' preferences and habits. The literature background on the subject will be descripted in Chapter 2, focusing on the three main phases of the GPS data analysis process, i.e. the map-matching of the GPS points to the network database, the choice set generation methods, and route choice modelling.

Further, two applications to two case studies will be described: Chapter 3 will provide detail on the case study of the GPS database of cyclists trips collected in Bologna in the framework of the European Cycling Challenge. First, the GPS trips database and the network database will be described. Subsequently, a novel map-matching procedure will be proposed and applied to the case study; from this map-matching phase of the study a second Journal paper has been excerpted and published. The last part of Chapter 3 will provide a descriptive analysis of the traces obtained as an output of the map-matching algorithm previously proposed.

Chapter 4 focuses on the second case study, the bicycle GPS traces collected in the Netherlands for the Mobile Mobility Panel. This second database available for the study will be described, the analysis of matched trips will be illustrated and two models for route choice will be estimated and discussed.

Finally, general conclusions from the whole PhD study will be drawn in the final part of the document, summarizing the applied methodologies, the main results, and the lessons that could be derived, also considering possible future developments to this research.



The structure of the PhD thesis.

Chapter 1: Quantifying the role of disturbances and speeds on separated bicycle facilities

Journal Paper 1: The paper presented in Chapter 2 is first-authored by the candidate, co-authored by Professor Kevin J. Krizek of the University of Colorado Boulder, and co-authored by Professor Federico Rupi, and is published as:

Bernardi, S., Krizek, K.J., Rupi, F. (2016). Quantifying the role of disturbances and speeds on separated bicycle facilities. Journal of Transport and Land Use, 9(2), In Press.

A preliminary version of this paper was presented at the World Symposium of Transport and Land Use Research in Delft (NL), in June 2014. The paper was then updated, rewritten and accepted for publication from the Journal of Transport and Land Use, scheduled for Scheduled for JTLU vol.9 (2), 2016.

1.1. Overview

As cities worldwide aim to spur more bicycling in cities, a key issue revolves around the nature and design of specific, bicycling-oriented facilities. What design treatments might be necessary to best separate cyclists and pedestrians from traffic? What are associated land use constraints? Available guidelines (e.g., see CROW 2007, NACTO 2012, AASHTO 2012) are usually specific to various countries or contexts and provide differing levels of specificity about the degree to which cycling facilities mix with pedestrians. What remains unknown is detailed knowledge about how pedestrians mix with or impede cycling behavior in different contexts. This research therefore aims to quantify the impact of impedances along different types of cycling facilities, focusing on, but certainly not limited to, the role of pedestrians.

Several research efforts address the effectiveness of dedicated bicycle facilities on outcomes such as cyclist safety, user satisfaction or other. Using stated preference surveys (Abraham et al. 2002, Sener et al. 2009, Stinson and Bhat 2005, Krizek and Roland 2005, Tilahun et al. 2007), revealed preference surveys (Broach et al. 2012, Menghini et Al. 2010, Hood et Al. 2011), and accident data (Aultman-Hall and Hall 1998, Lusk et al. 2011, 2013, 2014), the research base generally suggests that dedicated facilities help spur levels of cycling (Schweizer and Rupi, 2014), provide safety benefits (perceived and real), and advance a more pleasurable cycling experience.

Cyclists consider a variety of factors when evaluating the utility of a route and more specifically, a particular segment of a route. Decisions are often motivated by perceived safety and a desire to maintain their ideal cycling speed (i.e., they tend not want to be slowed by various obstructions). Safety and speed are influenced by the geometric design of the facilities (turning radii, slope, lane width) and functional features (proximity to cars, speed of adjacent cars, proximity of car parking stalls, presence of disturbances on the facility). But key dimensions that have received less attention in the design and research about bicycle facilities revolve around disturbances: specific obstacles in the facility that affect user satisfaction. Disturbances may be stationary (e.g., intersections, utility poles, bollards) or non-stationary (e.g., other cyclists or pedestrians).

For example, intersections that punctuate bicycle facilities have shown to be particularly vexing: a body of literature examining intersections along off-street paths is burgeoning (Phillips et al. 2011, Schepers et al. 2011, Sorton and Walsh 1994, Strauss and Miranda-Moreno 2013). In the path/route choice literature, using both SP and RP-based studies Broach et al. (2012) and Sener et al. (2009) found that route utility reduced by high-traffic crossings, stop signs, and traffic signals. Although the delay of these intersections can vary and be modified to favour cyclists, depending on traffic control devices used (signals or stop signs), more intersections than fewer are often perceived as a nuisance and impedance.

A different type of disturbances are those that are non-stationary in nature—other cyclists or users of other modes. Sometimes, the utility of a separated bicycle facility may be undetermined because of the intended or unintended need to mix with other modes. Pedestrians are a prime example of such. Little research, however, has focused on the extent to which pedestrians mixing with cyclists decrease the speed and therefore lessen the utility.

The most relevant literature we are aware of in this respect refers to "pedestrian hindrances." For example, knowing volumes and speeds of pedestrians and cyclists, Botma (1995) initially proposed a model to evaluate the number of passing and meeting events on off-street facilities. This work was later adopted and applied to different contexts (Allen et al 1998, Virkler et al. 1998, Green et al. 2003, Highway Capacity Manual 2010) to demonstrate how hindrances affect functional characteristics of bicycle facilities. The main contribution of these studies was the determination of a level of service for shared off-street facilities based on the number of passing and meeting events. This work, however, left issues unanswered, such as the specific impact of hindrances on users speed. Also, Botma's methodology supposes a normal distribution of bicycle speeds and fails to consider alternatives such as mixing with traffic.

We aim to fill this gap in the literature by systematically analyzing three transportation segments in Bologna (Italy). We considered two different environments: off-street bicycle facilities and mixed traffic conditions in the roadway. By definition, off-street facilities include both separated bicycle facilities (exclusive to cyclists – cycleways – or shared with pedestrians) and cycle tracks (, i.e. bicycle-exclusive paths) (European Economic and Social Committee, 2012).

We focus here exclusively on non-stationary disturbances to cyclists: pedestrians and bicyclists for the off-street environments and motorized vehicles for roadway environment. We compared cycling travel speeds in each environment to quantify speed reductions, largely types and amounts of non-stationary disturbances.

Our results contribute to the literature by quantifying cyclist speed reductions from pedestrians (on off-street facilities) and motorized traffic (on the roadway). This work has direct implications for developing decay curves that relate travel speed on cycling facilities with pedestrian volumes; such decay curves prove useful for the future planning of dedicated facilities in urban areas. We describe the context of our research and the specifics of the three segments. The following section details our data collection process and results. The conclusions discuss implications and future research needs.

1.2. The specific context of separated cycling facilities

Our research focuses on disturbances, which are a situated in a particular cultural context. It is therefore helpful to each as they apply to our study—being situated in Bologna (Italy)—and some peculiarities relating to bicycle facilities. Structural and functional aspects of cycling facilities in Italy are regulated by the Codice della Strada, along with the Decreto Ministeriale number 557 from 1999, "Regolamento recante norme per la definizione delle caratteristiche tecniche delle piste ciclabili." These regulations offer guidelines to plan and design cycling facilities, stating as their aim, "the achievement of a proper level of safety and environmental sustainability" (Decreto Ministeriale number 557 of 1999, Article 1).

Different designs for off-street facilities described by the Italian guidelines are shown in Figure 1. These guidelines, similar to the situation in many countries, prescribe various types of bicycle facilities for different types of traffic situations. There is wide variation in the types of facilities and the manner in which they separate cyclists from cars.

Note at least three factors run in the face of these guidelines. The first is that it is not uncommon for a cycling facility to be constructed with design treatments misapplied in a given context; alternatively, specific standards might be waived or ignored. Secondly, contextual factors (most often lack of space) affect the overall quality of specific design elements. For example, in some instances, an “off-street bicycle facility” is comprised of a raised treatment (e.g., a curb with accompanying bricks) on one side of the street to separate the cyclist from cars. The other side of the cycling lane, however, might merely be a painted stripe to separate cyclists from pedestrians. This leads to a third issue: even when a facility is designed and rules for its use are specified, regulations are mere suggestions. When space for pedestrians is provided in adjacent space or banned from the cycling facility altogether, it is not uncommon for pedestrians to encroach on the cycling facility. In cases of high pedestrian flows, they merely “spill over” into the cyclist facility. This is commonplace not only in Bologna, but also in many cities throughout Europe where it is difficult to find space for non-motorized travel in extremely space-constrained situations.

Figure 1: Some examples of the off-street facility typologies in Italy, as categorized in the Italian regulation: a) separated path non-adjoining the roadway; b) separated path, separation obtained by a non-continuous barrier; c) cycle track, obtained from the sidewalk, separating pedestrian and bicycle areas with different pavements; d) cycle track, obtained from the sidewalk, separating pedestrian and bicycle areas by painted markings.

1.3. Research approach

1.3.1. Selection of segments studied and their characteristics

Our central objective was to quantify how different types of disturbances affect cyclist travel speed; we chose three segments of Bologna's (Italy) cycling network to do so (Figure

2)[1]. Our primary concern was on the interplay between pedestrians and cyclists in off-street bicycle facilities—but we aimed to do so, in part, relative to disturbances caused by motorized vehicles (cars, buses/trucks, motorcyclists) in mixed traffic. We therefore chose study segments with: (a) an off-street bicycle facility adjacent to the roadway, (b) varying levels of pedestrian use around them, and (c) varying traffic volumes on the roadway. To the extent possible, we chose segments away from key intersections, allowing us to better focus on primary characteristics of segments vis-a-vis cyclists travel behaviour. Similarly, we aimed to keep speeds of vehicular travel constant as best as possible. The segments were defined to be 20 meters long and the location of each is approximately two kilometres from the centre of Bologna. Characteristics of each segment are described below with locations shown in Figure 2. Summary characteristics are presented in Table 1.

Figure 2: Map of Bologna and locations of the three segments.

[1] We employ the following nomenclature throughout this manuscript to improve clarity. *Segments* are the three areas studied; each segment is approximately 20 meters long. These segments contain *an off-street bicycle facility* (that is separated from traffic via a physical means — raised curb, parked cars or median) and an adjacent travel lane where the cyclists can mix with motorized traffic, i.e. the roadway. Therefore, each segment has two different types of environment.

Segments	Type of off-street bicycle facility	Width of the bicycle area (m)	Width of the pedestrian area (m)	Pedestrian volumes	Posted speed limit of adjacent roadway (km/h)	Peak bus volumes on adjacent roadway /hour
(1) Ercolani	Exclusive separated facility	1.80	-	Low	50	40
(2) Fioravanti	Cycle track on pavement	2.10	1.90	Modest	50	5
(3) Matteotti	Cycle track on pavement	1.90	1.50	High	50	55

Table 1: Features of the three segments examined.

(1) Ercolani (Figure 3) is along the first ring road of the city. It includes two carriageways with 50 km/h traffic that is divided by a median strip. Both sides of the northbound and southbound carriageways carry buses: 40 per hour, southbound, during peak times, and 20 per hour, northbound. The off-street bicycle facility in this section is part of a separated facility with bidirectional lanes, reserved only for cyclists[2].

[2] The red circled sign with the pedestrian indicates pedestrians are technically prohibited from the facility.

Figure 3: The off-street facility and roadway layout in Segment 1.

(2) Fioravanti (Figure 4) is along a one-way street that heads toward the city centre (southbound). Bus traffic is considerably less than in Segment #1 (5 per hour), but lateral parking is permitted on both sides of the roadway. The off-street bicycle facility in this section is bi-directional and is separated from a pedestrian zone by a painted stripe. Pedestrian volumes are modest.

Figure 4: The off-street facility and roadway layout in Segment 2.

(3) Matteotti (Figure 5) is adjacent to a four lane road close to the central train station; three lanes head southbound and one is reserved for northbound traffic. The character of the off-street bicycle facility in this section is similar to that in Segment #2. Being close to the central train station, bicycle, pedestrian and motorized volumes are high. On average, bus volumes are between 50 and 60 per hour, using all three southbound lanes.

Figure 5: The off-street facility and roadway layout in Segment 3.

1.3.2 Data collection

For each of the three segments, we studied cyclist travel on the off-street bicycle facility and on the immediately adjacent roadway (for a total of six environments). We chose seven working days (from 8:30 until 10:30) in April 2012 to be best representative of general cycling conditions. We collected data in two different phases. Our first phase counted the number of cyclists on each of the six different environments and timed how long each cyclist took to travel the 20 meter stretch. We tallied cyclists in all directions according to the environment they were cycling in (off-street versus roadway). This data provided us with baseline information about speed and also pointed to general patterns of use across the different facilities.

In the second phase we administered our own experiments to collect data relating disturbances and speed. One researcher cycled in each environment approximately 100 times. Employing the typology presented in Table 2, he used his own judgement to record all types of disturbance he encountered. For example, pedestrians encroaching onto the off-street bicycle facility were a disturbance; in the mixed traffic, cars or buses that forced him to share a lane or otherwise affected his travel were a disturbance . The researcher was asked to conform to typical cycling behaviour, and to keep it consistent (e.g., when to pass other cyclists). Furthermore, he was encouraged to apply consistency in how each type of disturbance was classified. The other researcher timed the cyclist on the 20 meter stretch. The research team did this approximately 100 times for each environment, totalling almost 600 coupled measures.

Off-street Bicycle Facility	**Roadway**
No pedestrians, bikes in same direction	No disturbances or two-wheeled vehicles
No pedestrians, bikes in opposite direction	1 car or more
1 – 3 pedestrians	1 bus or more
4 pedestrians or more	Heavy vehicles

Table 2: Classes of disturbances considered.

1.4. Results and analysis

1.4.1. Speed and use attributes

We first aimed to better understand baseline measures for each of the three environments and the relative characteristics. Segment #3 had the highest volume of cyclists, recording an average of 850 cyclists for the two-hour period; Segment #1 averaged 480 cyclists and Segment #2 had 240. These volumes are consistent with the intensity of land uses and activities around each segment (e.g., Segment #1 is adjacent to the central station and along a primary corridor headed into town; Segment #2 is in a more peripheral location and on a minor road). Using data collected from this phase we calculated average speeds for each environment; summary results are shown in Table 3. Average cycling speeds varied between 14.6 and 22 km/h, which generally agree with the values reported in Allen et al. (1998). Most studies with which we are familiar found bicycle speeds to be normally distributed; this varied by facility in our data and distributions for each are shown in Figure 6. A significance t-test was performed for the two speed samples, comparing off-street facility and mixed traffic data, and results revealed a statistically significant difference between the two samples for all the three segments.

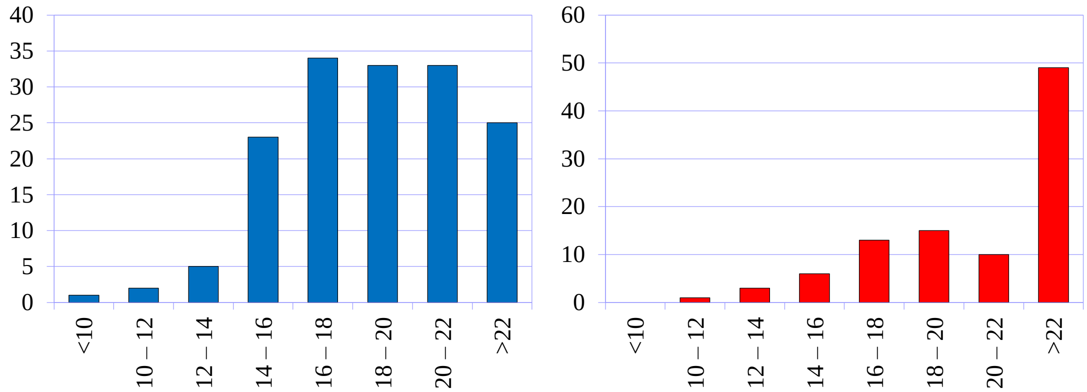
Across all segments, average cyclist speed was higher in mixed traffic facility than on the off-street bicycle facility. Overall, cyclists' speeds were highest in Segment #1 (18.9 km/h on the off-street bicycle facility and 22 km/h in mixed traffic), largely owing to pedestrians being technically forbidden in the off-street facility and cyclists often using the dedicated bus lanes in the roadway.

Figure 7 shows the proportion of cyclists' using the different environments. Segment #2 had the highest percentage of cyclists using the off-street facility (73%), followed by 58% in Segment #1. In contrast, less than half of the cyclists (47%) used the off-street facility in Segment #3. Even in Segment #1, where the off-street facility is not shared with other users, 42% of the cyclists chose to mix with traffic. Several factors contribute to these patterns and explaining them is clearly beyond the scope of our immediate study. Common sense suggests that variations and revealed preferences have to do with the width of the adjacent corridor for pedestrians (which lead to disturbances); other matters relate to discontinuities of the off-street facility further upstream (or downstream). Still other factors may be due to the cyclist's desire to link origins and destinations along the corridor; the specific facility might be out of the way of their needed route. But some of the variation is also influenced by intensity of the non-stationary disturbances (particularly pedestrians) in affecting speed along the routes. We now turn to quantifying the extent of these non-stationary disturbances in terms of decreased speed.

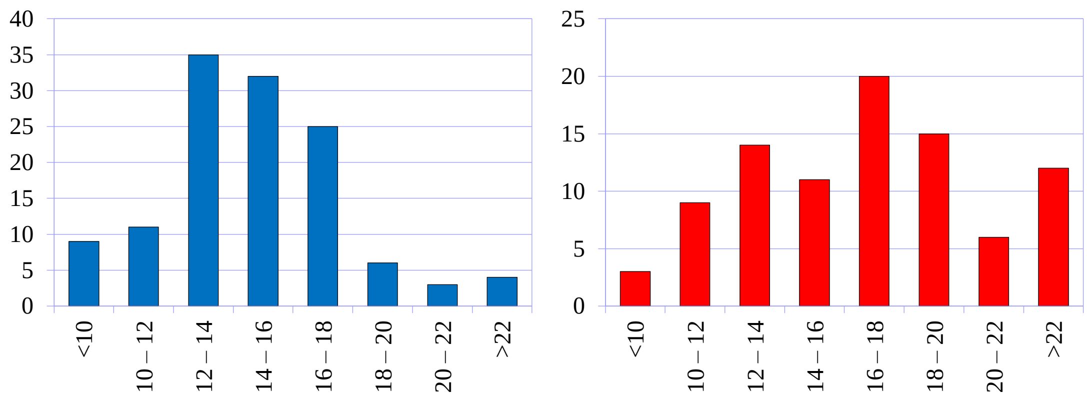
Segments	Average # of cyclists measured in 2 hrs	Statistic	Off-street Bicycle Facility	Roadway
(1)Ercolani	480	Mean speed (km/h)	18.90	22.00
		Standard deviation (km/h)	3.16	5.08
		Coeff. of Variation	0.168	0.231
(2)Fioravanti	240	Mean Speed (km/h)	14.60	16.8
		Standard deviation (km/h)	3.12	4.24
		Coeff. of Variation	0.213	0.252
(3)Matteotti	850	Mean Speed (km/h)	16.00	17.00
		Standard deviation (km/h)	2.97	4.39
		Coeff. of Variation	0.186	0.259

Table 3: Cyclists' volumes and average speeds for the three segments.

Segment 1 - Ercolani



Segment 2 - Fioravanti



Segment 3 - Matteotti

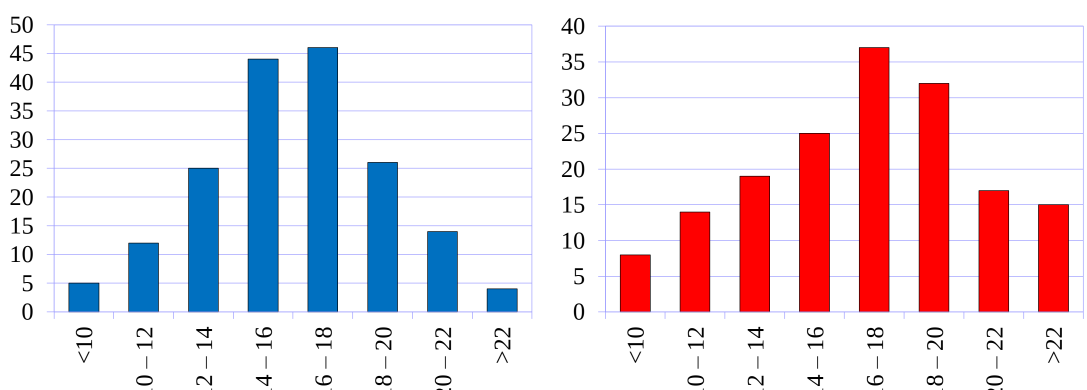


Figure 6: Cyclists' travel speed distributions: on the horizontal axes speed values are divided in ranges; the bars represent the frequency of speeds. In blue, the results for separated bicycle facilities, in red those for the roadway.

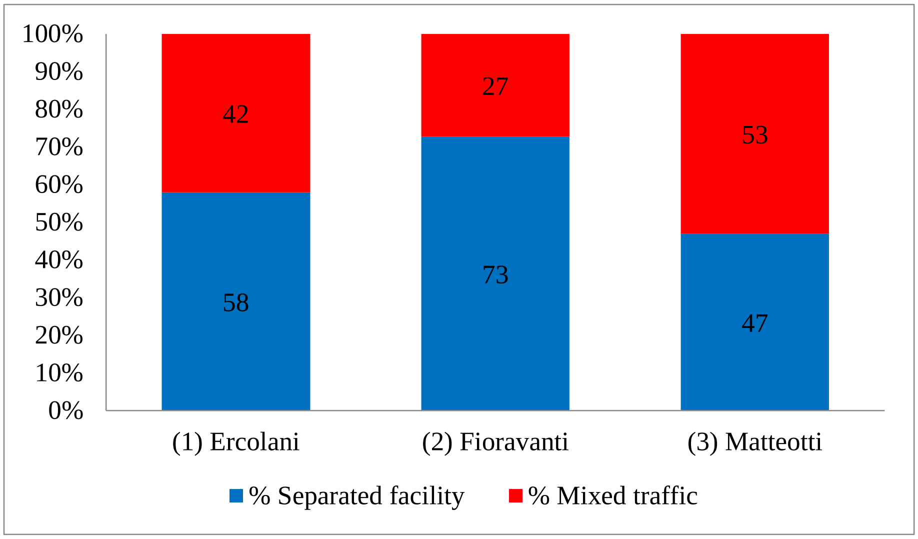


Figure 7: Cyclist use by facility.

1.4.2. Disturbance analysis

For each of the approximate 600 coupled observations, we calculated average speeds and examined them relative to the type of disturbance. For each recorded disturbance, we calculated the speed reduction compared to free flow (undisturbed) travel conditions. We also tallied the frequency of the different types of disturbances, reflected as a proportion of measurements where a disturbance was recorded. Finally, we calculated the average speed weighted by the frequency of each disturbance for both the off-street bicycle facility and the roadway. Results are shown in Tables 4 and 5, by segment.

Average travel speeds reflect the probability of each type of disturbance. The bottom row in the tables shows the frequency of measurements where at least one disturbance was registered — as a complement to the percentage of measurements where no disturbances were registered[3]. The cyclist-researcher experienced a disturbance approximately half the time. And, when the cyclist-researcher experienced no disturbances, his speeds generally agreed with the speeds recorded from phase 1 of the data collection. For Segments #1 and #3, disturbances were more frequent on the off-street bicycle facilities (54% and 90% of the time) than on the roadway (45% and 58%, respectively). The frequency of disturbances was more balanced across the facilities for Segment #2 (58% for separated and 62% for mixed).

The frequency by type of disturbance varied. Pedestrians were most common for the off-street facilities in Segments #2 and #3; for Segment #3, a cyclist would encounter a pedestrian in 90% of the observances. Cyclists in the opposite direction were most common in #1. For the roadway, cars were clearly the most common disturbance (24%, 57%, 47% of observances).

In each of the off-street facilities, pedestrians had the largest impact on cyclists' travel speed, affecting a 10 to 27% reduction. Other cyclists were second, slowing speeds by 5%, on average. The speed reduction was felt most acutely in Segment #3 — owing to a variety of factors — but overall pedestrian volumes were clearly one of them.

In mixed traffic, we noticed speed reductions from cars and trucks; disturbances from motorcycles were negligible. In Segment #1, buses led to speed reductions of 25%; in Segment #3, cars and buses led to speed reductions of 32% and 37%. The largest impact was associated with heavy vehicles, a 63% speed reduction. In general, cars, buses and heavy vehicles, have the greatest impact on reducing cyclist travel speed; pedestrians much less.

Furthermore, we considered weighted average — jointly analysing the average speed reduction weighted by frequency of disturbance. For Segments #1 and #2, weighted speed reductions are relatively minor, ranging between 2% and 14%. The weighted speed

[3] Percentages of observations with no disturbance and at least one disturbance sum to 100. Some of the observations, however, had more than one disturbance; therefore, the sum of the percentages of events for single classes of disturbance exceeds 100.

reductions we calculated in Segment #3 were considerable: 20% for the off-street facility and 40% for the mixed traffic. This segment is rich in disturbances and has notable speed reductions. Our analysis suggests cyclists might be making route choice trade-offs that are influenced by two factors: (1) frequent pedestrian disturbances on the off-street facility that produce moderate speed reductions, (2) relatively fewer disturbances on the roadway that have more severe speed reductions. Most cyclists prefer to travel in the mixed traffic — though robustly explaining this phenomena is beyond the scope of this data collection exercise.

Such factors suggest the attractiveness of a bicycle facility is impacted by jointly considering the frequency of disturbances *and* their effects. This joint calculus is a contributing factor in a cyclist's decision to use an off-street facility versus the roadway. In other words, even if they know that encountering a heavy vehicle might present a safety hazard, at least on segments we studied, the event could be perceived to be rare enough for them to choose the mixed traffic condition.

	(1)Ercolani		(2)Fioravanti		(3)Matteotti	
Disturbance type	**Average Travel Speed (km/h)**	**% Reduction**	**Average Travel Speed (km/h)**	**% Reduction**	**Average Travel Speed (km/h)**	**% Reduction**
No disturbance	16.7	-	15.2	-	15.6	-
% of events	45.9%		41.9%		9.6%	
Bikes same direction	16.0	4.2%	14.1	6.8%	15.5	0.14%
% of events	6.6%		13.3%		4.8%	
Bikes opp. direction	16.2	3.1%	14.4	5.0%	14.1	9.0%
% of events	29.5%		10.7%		1.2%	
1 – 3 pedestrians	14.5	13.1%	13.5	10.8%	12.3	20.9%
% of events	13.1%		33.3%		84.3%	
4 pedestrians or more	n.a.	n.a.	13.5	11.1%	11.3	27.2%
% of events	0.0%		1.3%		18.1%	
Average speed, weighted by disturbance						
	15.6	6.4%	**13.9**	8.8%	**12.5**	19.6%
% of events	54.1%		58.1%		90.4%	

Table 4: Cycling Speeds by Type of Disturbance (Off-street Bicycle Facilities).

	(1)Ercolani		(2)Fioravanti		(3)Matteotti	
Disturbance type	**Average Travel Speed (km/h)**	**Reduction%**	**Average Travel Speed (km/h)**	**Reduction%**	**Average Travel Speed (km/h)**	**Reduction%**
No disturbance	18.6	-	17.0	-	17.7	-
% of events	55.2%		37.5%		42.1%	
Two-wheeled veh.	18.4	1.1%	17.0	0.26%	17.3	2.2%
% of events	10.3%		3,1%		5,3%	
1 car or more	17.0	9.0%	16.8	1.15%	12.0	31.7%
% of events	24.1%		56.3%		47.4%	
1 bus or more	13.8	25.8%	n.a	n.a	11.1	37.2%
% of events	10.3%		0.0%		21.1%	
Heavy vehicles	n.a	n.a	14.9	12.3%	6.6	62.9%
% of events	0.0%		3.1%		2.6%	
Average speed, weighted by disturbance						
	16.0	14.1%	**16.7**	1.8%	**10.53**	40.5%
% of events	44.8%		62.5%		57.9%	

Table 5: Cycling Speeds by Type of Disturbance (Roadway).

1.5. Implications and future research

Aim to spur non-motorized travel, transport officials in cities are often challenged by finding available space in travel corridors for dedicated cycling facilities. This work informs these matters in several respects and also provides springboard to spur future research on the issue. We employed a data collection approach to quantify the speed effects of non-stationary disturbances — disturbances which often influence cyclists' route choice decisions. Relying on speed as just one of several performance measures, we measured the joint effects of disturbance frequency and disturbance impact.

We learned that motorized disturbances (particularly heavy vehicles) have the strongest impact on lowering cyclist's travel speed. However, these specific disturbances proved to be relatively infrequent relative to those encountered on off-street bicycle facilities (where pedestrians are more plentiful and contribute to slowing the speed of cyclists). To our knowledge, this is the first data collection effort and analysis to point to this trade-off. Furthermore, failing proper separation between users, we learned that the presence of pedestrians can reduce cyclists speed by up to 30 percent.

Myriad factors influence a cyclist's route choice, particularly the decision to use an off-street bicycle facility or ride in traffic. The probability of disturbances and their impact on travel speed are just two of them. Our results suggest that design elements of these facilities can play a role in affecting the frequency, type, and severity of disturbances. However, it is necessary for future work to better isolate this element (e.g., better projecting how wide sidewalks need to be given pedestrian volumes to avoid disturbing the bicycle facility). For example, Segment #2 had excess space for pedestrian travel and therefore little speed reductions from pedestrians; cyclists can avert pedestrians without significantly reducing their speed and, evidently, also pedestrian safety results enhanced. Furthermore, wider cycle-tracks (when warranted) benefit cyclist-cyclist interactions because it is easier to pass each other going in the same direction.

Our data collection effort represents a univariate population, which is an oversimplification. Different types of cyclists prefer different environments (Larsen and El-Geneidy 2011, Sener et al. 2009, Wilkinson et al. 1994). This type investigation could benefit from more robustly accounting for demographic, attitudinal or other behavioural data which was unfortunately unavailable for this effort. Furthermore, the type of detailed GPS data that is now being employed in other bicycle research applications (Broach et al. 2012, Hood et al. 2011, Menghini et al. 2011) could be adapted to focus specifically on issues we studied here. GPS technology could then be used to focus on the role of trip type and broader route choice characteristics.

A subsequent step would be to integrate these results into a bicycle route choice model where frequency of disturbance and its magnitude represent various attributes of the utility. In the current state of the research, the most advanced approach merely considers the proportion of the facility that is separated for bicycling (Broach et al. 2012, Hood et al. 2011, Menghini et al. 2010). Finally, this research documents the frequency and quantifies the intensity of disturbances on speed. As future research on this matter evolves, cities would benefit from learning about various thresholds of pedestrian volume and corresponding cycling speed impacts. This work lays the foundation to create a decay curve which could be used to predict travel speed vis-à-vis pedestrian volumes (similar to how speed decay curves of vehicle flow are used to determine the level of service of infrastructures). Planners would then be able to more robustly estimate cyclist speeds—and levels of service—for different pedestrian volumes.

1.6. Summary

When off-street bicycle facilities are poorly designed or placed in less than optimum locations (such as in Segment #3), their intended use is less than anticipated. By analysing the speed reductions from different types of disturbances we found that frequency of disturbance might weigh more heavily than the intensity of the disturbance. Furthermore, encountering pedestrians along an off-street bicycle facility lowers average speeds by as much as 30 percent, especially in situations where the bicycle facility fails to conform to minimum standards. In mixed traffic, the presence of motorized vehicles can produce even more severe speed reductions (i.e. in case of heavy vehicles), but these interactions proved to be less frequent.

Most policy officials know they need to heed caution when prescribing off-street bicycle facilities adjacent to areas with high pedestrian activity. An outstanding question is how many pedestrians present a problem and the corresponding speed reductions. For researchers, this work helps spearhead more robust investigations to establish, for example speed decay curves based on pedestrian volumes. For practitioners, our results draw attention to the need for physically separating pedestrians from cyclists in facility design. Or, where the context suggests less than optimum conditions for an off-street facility, it helps quantify when it might be more convenient to better accommodate bicycle travel in mixed traffic conditions.

1.7. References

[1] AASHTO. 2012. Guide for the Development of Bicycle Facilities. American Association of State Highway and Transportation Officials, Washington D.C.

[2] Abraham, J., McMillan, S., Brownlee A., Hunt, J.D., 2002, Investigation of cycling sensitivities. Presented at the 81st Annual Meeting of the Transportation Research Board, Washington, D.C.

[3] Allen, D. P., Rouphail, N. M., Hummer, J. E., Milazzo, J. S., 1998., Operational Analysis of Uninterrupted Bicycle Facilities. Transportation Research Record, 163, 29-36.

[4] Aultman-Hall, L., Hall, F.L., 1998, Ottawa-Carleton commuter cyclist on- and off-road incident rates. Accident Analysis and Prevention, 30(1), 29-43.

[5] Botma, H., 1995, Method to Determine Level of Service for Bicycle Paths and Pedestrian-Bicycle Paths. Transportation Research Record, 1502, 38-44.

[6] Broach, J., Dill, J.,Gliebe, J., 2012, Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A, 46, 1730-1740.

[7] CROW, 2007, Design manual For Bicycle Traffic. National Information and Technology Platform for Infrastructure, Traffic, Transport and Public Space, The Netherlands.

[8] Decreto Legislativo 30 aprile 1992, n. 285. Nuovo Codice della Strada.

[9] Decreto Ministeriale 30 novembre 1999, n. 557. Regolamento recante norme per la definizione delle caratteristiche tecniche delle piste ciclabili.

[10] European Economic and Social Committee, 2012, European Cycling Lexicon. EUR-OP. doi:10.2864/10239.

[11] Green, J. S., Hummer, J. E., Rouphail, N. M, 2003, Using the number of passing and meeting events to design pedestrian and bicycles-use paths. Presented at the 2nd Urban Street Symposium: Uptown, downtown, or small town: designing urban streets that work; July 28-30, 2003.

[12] Hood, J., Sall, E., Charlton, B., 2011, A GPS-based bicycle route choice model for San Francisco, California. Transportation Letters, 3(1), 63-75.

[13] Krizek, K.J., Roland, R.W., 2005, What is at the end of the road? Understanding Discontinuities of On-street Bicycle Lane in Urban Settings. Transportation Research Part D, 10, 55-68.

[14] Larsen, J., El-Geneidy, A., 2011, A travel behavior analysis of urban cycling facilities in Montreal Canada. Transportation Research Part D, 16, 172-177.

[15] Lusk, A.C., Furth, P.G., Morency, P., Miranda-Moreno, L.F., Willet, W.C., Dennerlein, J.T., 2011, Risk of Injury for Bicycling on Cycle Tracks versus in the Street. Injury Prevention, 17(2), 131-135.

[16] Lusk, A.C., Morency, P., Miranda-Moreno, L.F., Willett, W.C., Dennerlein, J.T., 2013, Bicycle guidelines and crash rates on cycle tracks in the United States. American Journal of Public Health, 103(7), 1240-1248.

[17] Lusk, A.C., Wen, X., Zhou, L., 2014, Gender and used/preferred differences of bicycle routes, parking, intersection signals, and bicycle type: Professional middle class preferences in Hangzhou, China. Journal of Transport and Health, 1(2), 124-133.

[18] Menghini, G., Carrasco, N., Schüssler, N., Axhausen, K.W. 2010. Route choice of cyclists in Zurich. Transportation Research Part A, 44, 754-765.

[19] NACTO, 2012, Urban Bikeway Design Guide. National Association of City Transportation Officials, Washington D.C.

[20] Phillips, R. O., Bjørnskau, T., Hagman, R., Sagberg, F., 2011, Reduction in car-bicycle conflict at a road–cycle path intersection: Evidence of road user adaptation?. Transportation Research Part F, 14, 87-95.

[21] Schepers, J.P., Kroeze, P.A., Sweers, W., Wüst, J.C., 2011, Road factors and bicycle - motor vehicle crashes at unsignalized intersection. Accident Analysis and Prevention, 42, 853-861.

[22] Schweizer, J., Rupi, F., 2014, Performance evaluation of extreme bicycle scenarios. Procedia of Social and Behavioral Sciences, 111, 508-517.

[23] Sener, I., Eluru, N., Bhat, C., 2009, An analysis of bicycle route choice preferences in Texas, US. Transportation, 36(5), 511-539.

[24] Sorton, A., Walsh, T., 1994, Stress level as a tool to evaluate urban and suburban bicycle compatibility. Transportation Research Record 1438, 17-24.

[25] Stinson, M.A., Bhat, C.R., 2003, An Analysis of commuter bicyclist route choice using a stated preference survey. Transportation Research Record, 1829, 107-115.

[26] Strauss, J., Miranda-Moreno, L.F., 2013, Spatial modelling of bicycle activity at signalized intersections. Journal of Transport and Land Use, 6(2), 47-58.

[27] Tilahun, N., Levinson, D.M., Krizek, J.K., 2007, Trails, lanes, or traffic: the value of different bicycle facilities using an adaptive stated preference survey. Transportation Research Part A, 41, 287.301.

[28] Transportation Research Board., 2010, Off-street pedestrian and bicycle facilities. In Highway Capacity Manual. Washington, D.C.

[29] Virkler, M., Balasubramanian, R., 1998, Flow Characteristics on Shared Hiking-Biking-Jogging Trails. Transportation Research Record, 1636, 43-46.

[30] Wilkinson, W., Clarke, A., Epperson, B., Knoblauch, R., 1994, The effects of bicycle accommodations on bicycle/motor vehicle safety and traffic operations. U.S. Department of Transportation, Federal Highway Administration.

Chapter 2: The use of GPS data for bicycle route choice modelling

2.1. Introduction

As we discussed in the previous sections, understanding the main factors cyclists consider when evaluating an available path between an OD pair, is crucial in order to build effective infrastructures. The aim, when investing in bicycle infrastructure, is to answer to the following questions: Where people want to cycle from? Where to? On which path? Which type of facility they would prefer to ride on? And in which environment? When determining which path they will take from a certain origin to a certain destination, cyclists consider different aspects of the trip: travel time, directness, comfort, safety, are those generally indicated by literature. The analysis of cyclists' choice behaviour has been investigated in recent years, through the estimation of route choice models.

One of the important issues in studying route-choice behaviour in cycling is the availability of data and its collection procedure. Information about actual chosen routes cannot be collected through conventional household travel-surveys. Thus, one reason for the lack of such data is that routes cannot be easily reported in the Computer-Assisted-Telephone-Interview (CATI) methods used for data collection. In the past decades, route choice models could be estimated mainly using *stated preference data.* In this case, for the data collection, the source of data would be surveys where users were asked to state which path they would choose between a certain origin and destination. Survey's participants would have to indicate the chosen path on a map or, in alternative, different options could be pre-selected by researchers and users would have to indicate the one they would consider the most convenient. This second option would better allow researchers to control the alternative choice set. Nevertheless, stated preference surveys have been largely criticized, as they do not capture the actual choices of users, but only their intentions; indeed, it has been verified that what people state they would do, does not always match with their actual, observed, behaviour.

A way to collect data that more reliably describe users' choices is through *revealed preference data*. Traditionally, information about trips were collected at the destination, asking users to declare (and describe) the route they have just chose to reach that specific destination. While this data collection method let researchers observe more reliable choices, two are the main limitations: the small size of samples, when obtained by surveys at a specific location, and the difficulty of observing the routes.

In recent years, these issues has been significantly overcome by the use of Global Positioning System (GPS)-based travel surveys now provides an approach to trace vehicle movements and, hence, collect data on the actual routes chosen by cyclists in their trips.

To date, the literature on empirical modelling of route choices using GPS traces are quite limited (see the work of Schussler and Axhausen 2009, Menghini et al. 2010 for the city of Zurich, Hood et al., 2010, with an application to the city of San Francisco, Broach et al, 2012, for Portland), but undoubtedly expanding. This can be explained considering different factors: for example, on one side the level of accuracy of GPS devices has increased, thanks to the investments on the involved technologies. On the other side, people habits and attitude towards the use of smartphone changed very rapidly, to the point that nowadays it is absolutely common for people to possess a smartphone, keep self-localization devices activated, declare their position in order to make it available for a burgeoning set of applications and services, register, monitor and share with people their position, trips and activities. Further, in parallel with these changes in technology and society, research on mobility followed: several of the methodological developments relevant to modelling route choices from GPS-based travel surveys are relatively recent.

In this context, the broad focus of this research is to combine data from GPS-based travel surveys and Geographic Information Systems (GIS)-based roadway network databases to estimate models for route-choice. For these purpose three are the main macro-areas of research that have seen a significant development in recent years, and they can be classified following the three basic steps for any route-choice modelling study using raw GPS-data:

1) Map-matching;

2) Choice Set Generation;

3) Route Choice Modelling.

The map-matching is the process of identifying the specific links of the roadway traversed by a vehicle by mapping the points from its GPS trace to an underlying road network database. This step is critical as it identifies the fundamental "choice" (i.e., the route) of interest. In other words, given the singular GPS points registered during one trip – could be one per second, or registered with higher headway, generally up to 20 seconds – the map-matching procedure provides the sequence of links, belonging to the network database, that most likely where used. This means that the network database plays a key-role in the accuracy of the results of the map-matching: the more complete is the network representation, more likely it is that the process provides the path that was actually taken for the single registered trip. Furthermore, map-matching procedure should be easily replied for different contexts, and almost completely automated. For this reason, another important issue is the efficiency of the algorithms implied in this process, their ability to examine a high number of points and links with a reasonable computational effort.

Once the chosen path has been identified, the next step is to determine which other paths were available to the same traveller for moving between the same origin-destination pair. This process is called choice-set generation. Contrary to the case of a stated choice experiment, in this case there is not a limited choice set built "a priori", or indicated by the respondent. For this reason, researchers (see for example Bovy and Fiorenzo-Catalano, 2007, Prato and Bekhor, 2007, Rieser-Schüssler et al. 2012, Halldórsdóttir et al. 2014) tried to develop algorithms and procedures to "generate" plausible alternatives, considering the network topology or a various set of network features. The challenge, for choice-set generation methodologies, is to build alternatives that are realistic, distinct but also – as for the map-matching step – to maximize the efficiency of the involved procedures, especially in terms of computational effort.

Comparing the chosen routes for all respondents and all origin-destination pairs, and the corresponding choice sets, route choice models can be estimated, to describe the aspects of route choice behaviour. Considering the research field of route choice models, the literature is burgeoning with theoretical approaches and applications. For example, for an extensive description of discrete choice models that can be suitable for short-term travel decisions, such as route choice decisions, see Ben Akiva and Bierlaire (1999). The various models applied to bicycle route choice models will be listed and described further in this Chapter, in paragraph 3.2.3. In general, it can be said that the main aspect to considerate, when dealing with bicycle route choice deduced by smartphone data are: (1) the overlapping of alternatives and (2) the repetition of choices made by the same respondent. More in detail, the urban context and the roads on which people usually cycle require a small scale approach, and this leads to consider alternatives – paths – that can be very similar and highly overlapped. Different approaches and model forms were proposed to overcome this issue. Further, route choice models estimated from travel diaries – which is generally the case with smartphone data, should account for trip repetitions, between the same origin-destination pair, made by the same user. It is very likely that those choices that are usual and repeated over the days – systematic trips – present different features if compared with the occasional ones.

Another aspect that differs between the various models proposed in literature is the choice of explanatory variables, made by the analyst. This aspect is, of course, connected with the availability of data. As highlighted in the previous Chapters of this thesis, and widely established by research on bicycle trip mode, many are the aspects that cyclists consider when choosing a path over another, and such variety of factors should be captured by route choice models. In the last decade, in parallel with the development of the technologies connected to trip detecting and recording through portable localization devices, also the level of detail and attributes incorporated in network database has significantly increased. Thanks to the diffusion of what has been defined “volunteered geographic information”, richer network database are available, compiled by volunteers, both in terms of completeness and in terms of attributes describing the network's elements. As we will see in the following Chapter of this thesis, the application to the two case studies has been made, in both cases, using network database coming from volunteer cyclists: for the Italian case study, the Open Street Map network has been freely obtained and used; for the Dutch case study, the network compiled by the volunteers of Fietsersbond (a Dutch cyclists union) has been used. The "crowd-sourcing" approach for network database – also applied to other relevant information, e.g. land-use – allow to have larger – both in terms of area and in terms of detail – and more descriptive networks. Thus, more accurate models can be estimated for bicycle route choice, for example including route attributes (such as travel time, numbers of turns, and number of intersections, pavement, presence and type of bicycle facility, environment, etc.), trip attributes (time of the day, day of the week, trip purpose), and traveller attributes (gender, age, experience etc).

The next paragraphs provide a review of the existing studies in the areas of map matching, choice-set generation and route choice modelling.

2.2. Literature Background

2.2.1. Map-matching

The map-matching process consists of all those operations that transform a stream of GPS points to a road-network database to identify the traversed links in the chosen route.

There are many map-matching algorithms proposed in literature, depending on the input data and on the application context. An extended review of map-matching algorithms applied to transportation research can be found it Quddus et al. In this review, map-matching algorithms have been categorized into three typologies: geometrical, topological and advanced. The geometrical approach is normally based on the distance between GPS points and network elements, edges or nodes (proximity-based algorithms): this approach is rather uncomplicated in terms of implementation, but the accuracy and the stability of such algorithms is rather scarce. For example, Spissu et al (2011) implemented a GIS-based procedure, using proximity as the main criteria for matching the GPS points to the network: a spatial join was used to match the GPS points to the corresponding routes, and afterwards a manual inspection corrected the matching errors. The method found routes only for 58% of the trips. Unmatched trips were due to missing GPS points, inconsistent activity data and missing links in the roadway network. Furthermore, the main drawback of this methodology is the need of a manual check of the results.

Topological approaches, other than the sole distance, also take into consideration the sequence of GPS points and the connectivity of network elements. For example, some studies have applied the “shortest-path procedure”. Du (2005) proposed a method that predicts the chosen route by determining the shortest path satisfying network topology such as link location, connectivity, one-ways, and allowable u-turns. The method was implemented in ArcGIS and examined against 674 trips collected on 18 known routes of Lexington, KY. For a known OD pair, approximately 95% of the routes were constructed entirely. However, high computational times and manual interventions were needed.

Finally, advanced map-matching procedures typically combine geometric and topological procedures with additional criteria, parameters, assumptions, probabilistic analysis, or optimization. For example, an advanced approach consists of taking into consideration the GPS measurement errors and constructing confidence regions around the points, based on data quality; then, only the edges within these regions are selected for evaluation. The evaluation can be based on a combination of different criteria, such as distance, speed or connectivity to the previously matched edges. The challenge for new and advanced map-matching procedures is to increase the accuracy while maintaining computational efficiency.

Marchal et al. (2005) proposed an algorithm that uses a multiple hypothesis technique (MHT).The MHT stores multiple paths during the process and in the end selects a path with the best score. The authors starts by calculating a “link score” by determining the distance of the link to the GPS point. The algorithm starts with finding a set of links that are closest to the first GPS point. For each link, a new path is created and links are inserted with their scores assigned to the respective paths. GPS points are processed in order, based on their timestamp, and a path score is updated by adding the score of the last added link to its previous score. When the end of a link is reached, a copy of the path is created for each outgoing link and then the path is removed from the set of paths. In the end, the path (in the set of paths) with the lowest score is selected as the traversed path. The algorithm could not produce continuous routes in most cases, which the authors suggest is due to irregularities GPS streams caused by tunnels, tree canopies, poor signal, and so forth. As a result, a sequence of paths was generated instead of a continuous route. The authors also added that algorithm is sensitive to outliers in GPS data.

Schüssler and Axhausen (2009) modified the original algorithm by Marchal et al. (2005) to overcome their limitation of not producing a continuous route. Additionally, a modified method was used to calculate a score. First, they subdivided each trip into continuous segments depending on the gaps in GPS streams. Afterwards, they created the trip segments by using the algorithm by Marchal et al. (2005). Then, a complete trip was obtained by connecting trip segments through a shortest path search with a treatment for low quality map matching results. The results showed a smaller number of matched routes in comparison to the total routes. Further investigation of the results showed three main reasons for such low numbers of matched routes: missing links in the roadway network, off-network travel, and u-turns.

In the next Chapter, a novel methodology to map-match the GPS traces collected by cyclitsts in Bologna (Italy) will be proposed. The map-matching procedure described in Chapter 4 applies some of the concepts introduced by Marchal et al. (2005), such as using GPS point proximity to a link to determine its probability of being chosen, and the idea of building the complete chosen path maximizing the link score – or, as it will be described, minimizing the link weight. The aim is to obtain an efficient algorithm, capable of matching a large sample of GPS points to high detail network, with a high accuracy. The accuracy of the map-matching methodology will be evaluated based on the difference in length between matched route and the polyline obtained by linking the GPS points' sequence.

What should be highlighted is that the application of map-matching methodologies that have been consolidated for car trips could present issues when applied to bicycle trips. Indeed, bicycle trips registered in urban contexts presents specific issues. Firstly, in dense urban streets the partial absence of GPS points in a trip is more likely, due to the loss of signal of the geo-localizing device. Secondly, cyclists do not necessarily cycle in roadways or on the provided bikeways, but they sometimes use pedestrian paths, prohibited ways, or they cross parks and squares. This higher degree of freedom characterizing bicycle trips makes GPS recorded data more difficult to match to the available road maps: in case of bicycle traces, the quality of the map-matching is particularly sensitive to the accuracy and details of the available road network model.

2.2.2. Alternative choice set generation

Once the chosen routes have been determined, the next step is to determine the set of alternative paths available for the same trip. Since the revealed preference surveys do not directly ask the respondents to provide information on the options available/considered, nor a limited set of alternatives are designed by the analysts and submitted to respondents, the choice sets have to be determined considering those paths that are topologically feasible with the respect to the network model. Clearly, it is unlikely that any traveler, at the moment of starting his trip, is aware of all the paths available between the desired origin-destination pair. For this reason, to consider as the choice set the complete set of possible paths would not be correct, as a “universal” choice set would contain a high number of unattractive and unrealistic routes that the traveler would never consider during the decision-making process. The inclusion of these unrealistic routes in a choice set would result in an incorrect model estimation, as well as representing an incredible burden in terms of computational effort. Therefore, a choice set must be limited to those paths that actually represent a realistic alternative for the traveler. By definition, the choice set consist of the collection of travel options perceived as available by individual travelers in satisfying their travel demand (Bovy, 2009). Another aspect that should be accounted for is the overlapping of alternatives: indeed, considering options that could be perceived as too similar by the traveler would result in errors in the modelling phase.

For what concern bicycle route choice applications, choice-set generation approaches proposed in literature can be classified into two categories:

1) Shortest path based methods;

2) Probabilistic methods.

Shortest-path based methods are the most popular and commonly used methods in the literature. Assuming a given generalized cost expression, this method searches for the minimum-cost path in the network between the origin and destination under examination. Over the years, researchers have introduced several variations to the basic approach of minimizing path's cost, proposing different applications which maintained a fair computational efficiency.

For example, "K-shortest path" algorithms extend the idea of calculating a single shortest path to determine k-shortest paths for a given generalized link cost function (Papinski and Scott, 2011, Spissu et al., 2011)

Another approach is to calculate multiple shortest paths considering multiple attributes for cost, called "labels" (from which the name "labelling" approach). In other words, the labelling approach consists in finding the best path considering more than one attribute at a time. Ben-Akiva et al. (1984), who proposed the approach, generated routes using labels that include time, distance, scenic, signals, capacity, hierarchical travel pattern, quality of pavement, commercial development, highway distance, and congestion. Routes with only two labels, time and distance, replicated 70% of the chosen routes and routes with all labels together replicated 90% of the chosen routes. The study found signals not being a significant factor and concluded that factors other than time and distance do play a significant role in route choice. Prato and Bekhor (2007) applied the same approach, but using only distance, free-flow time, delay and traffic lights.

Further, “link elimination” approaches (Frejinger and Bierlaire, 2007, Halldórsdóttir et al. 2014) also start from calculating the minimum-cost path, but subsequently they generate multiple alternatives by iteratively removing one or more links from it, and re-calculate the new best path. Schüssler et al. (2012) proposed a variation to the approach, called “breadth-first search link elimination” (BFS-LE). The algorithm searches for the minimum-cost path between origin and destination. Consecutively, the links of the shortest path are removed one by one and the shortest path for the resulting network is determined. Once all links of the original shortest path have been processed, the algorithm proceeds to the next level, where two links at a time are eliminated. The algorithm monitors the generated networks and only keeps unique and connected routes. The algorithm continues until the maximum number of unique routes has been generated, the time abort threshold is met or there are not more feasible routes between origin and destination (Halldórsdóttir et al., 2014). Also the BFS-LE can consider different cost attributes for the determination of the minimum-cost path; for example, Halldórsdóttir et al. (2014) considered four different cost functions, based on a combination of different attributes: road type (large or small), bicycle lane (segregated or not), and land use (scenic roads, or forest roads).

Differently to the deterministic methods, like those that have been just described, where an alternative either belongs to a choice set or not, probabilistic methods, consider intermediate availabilities by assigning perceived probabilities to routes. This set of approaches relies on the assumption that all routes connecting origin and destination belong to the choice set to some degree (Dhakar, 2013).

A fairly recent probabilistic approach, applied to bicycle route choice, is the “doubly stochastic generation function” (DSGF) (Bovy and Fiorenzo-Catalano, 2007, Hood et al. 2011, Halldórsdóttir et al., 2014). In the DSGF method, a shortest path search is carried out iteratively using an implementation of the Dijkstra´s algorithm on the network. At each iteration, costs for the paths are obtained by randomly drawing the cost of each link from a probability distribution and extracting attribute preferences for each traveller from another probability distribution. After each iteration, only unique routes not generated in previous ones are added to the route choice set. The algorithm stops when the choice set size has reached a pre-determined maximum or, alternatively, a pre-determined time threshold has been reached.

In general, almost every choice set generation method has been derived from approaches and algorithms that had already been tested for car trips. Once applied to bicycle trips, the most important variation made by researchers was to account for various cost attributes, not only time or length. This is due to the fact that, as we saw in the first Chapter of this thesis, since the first studies on bike as a travel mode, and the first attempts to determine a measure of the quality that cyclists attribute to the path they use – being them on the roadway or on exclusive facilities – it has been demonstrated that many different factors strongly influence cyclists perceptions and behavior, like perceived safety and comfort. These perceived safety and comfort cannot be explained only by the speed and directness of the path, but also from the presence of dedicated facilities, segregated lanes, intersections, signals, and environment.

2.2.3. Bicycle route choice modelling

The random utility discrete-choice models are the most commonly used approach for analyzing route-choice decisions (Ben Akiva and Bierlaire 1999). These models assume that the utility of an alternative consists of two components, one deterministic and one stochastic. In particular, the utility of alternative i in the choice set C_n , as perceived by the traveler $n,$ is given by:

$$U_{in} = V_{in} + \varepsilon_{in} \tag{1}$$

where V_{in} is the deterministic (observed) component and ε_{in} is the stochastic (unobserved) component.

For choice modeling, logit-based models are most commonly used. Among the family of logit models, the Multinomial Logit Model (MNL) is the simplest one. For the MNL model, the probability of choosing an alternative i in the choice set $C_n$is given by:

$$P(i/C_n) = \frac{e^{V_{in}}}{\sum_{j\in C_n} e^{V_{jn}}} \tag{2}$$

The MNL model is based on the assumption of Irrelevance of Independent Alternatives (IIA), and therefore does not consider the similarities between alternatives. This represents a limitation of the MNL model, which can affect the results when dealing with similar alternatives in the choice set. As we have already discussed in the previous paragraph, that is the case for route-choice models, even more for bicycle route-choice models, where the alternative paths can be significantly overlapped. On the other side, researchers have tried to maintain the computational benefits of the simple, closed-form MNL model, while proposing modifications to capture the similarities among paths.

Several models have been proposed in literature to overcome the limitations of the MNL model. This paragraph will present some of the most common models that proved to be suitable for bicycle route- choice behavior. For a deeper review of route-choice modelling research, one could refer to Prato (2009).

Methodologies that modify the deterministic part of the utility function include the C-logit and the Path Size Logit (PSL).

C-logit model, firstly developed by Cascetta et al. (1996), introduced a term called "commonality factor" in the deterministic part of the utility that measures the physical overlap of a route with other routes in the choice set. The commonality factor (CF) reduces the utility of a route due to its similarity with other routes. The probability of choosing an alternative i in the choice set C_n is given by:

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

(3).

Cascetta et al. (1996) propose the following specification for the commonality factor:

$$CF_{in} = \beta_{CF} ln \sum_{j \epsilon C_n} \left(\frac{L_{ij}}{\sqrt{L_i L_j}} \right)^{\gamma} \tag{4}$$

where L_{ij} is the length of links common to paths i and j, and L_i and L_j are the overall length of paths i and j, respectively. β_{CF} is a coefficient to be estimated. The parameter γ may be estimated or fixed to a convenient value, often 1 or 2. Note that the commonality factor of an alternative is not one of its attributes; in other words, it can be viewed as a measure of how the alternative is perceived within a choice set.

Ben-Akiva and Bierlaire (1999) proposed the Path-Size Logit (PSL) model and measured the similarity using a Path-Size term in the deterministic component. The Path-Size indicates the fraction of the path that constitutes a "full" alternative. Also for the PSL model the expression of the probability of choosing route k within the alternative paths reflects the simple Logit structure:

$$P(i/C_n) = \frac{e^{V_{in} + lnPS_{in}}}{\sum_{j \in C_n} e^{V_{jn} + lnPS_{jn}}} \tag{5}$$

where the Path-Size factor is defined as:

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \epsilon C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}} \tag{6}$$

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

Even though C-Logit and PSL have similar functional forms, each model gives a different interpretation with respect to the correction term introduced within the utility function.

The commonality factor reduces the utility of a path because of its similarity with respect to other routes, while the path size indicates the fraction of the path that constitutes a “full” alternative. Accordingly, a unique path has a size equal to one and N duplicate paths share the size $1/N$ (Prato 2009).

Another adaptation of the Logit model that account for similarities in the stochastic part of the utility (error correlations) while maintaining a closed-form formula for probabilities is called Mixed Logit (Ben-Akiva and Bolduc, 1996; Train, 2003). The defining characteristic of the Mixed Logit model (also called “Logit Kernel”) is that the unobserved factors can be decomposed into a part that contains correlation and heteroscedasticity, and another part that is i.i.d. extreme value. Mixed Logit is a highly flexible model that can approximate any random utility model (Train 2003). It resolves the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Unlike Probit, it is not restricted to normal distributions.
The most straightforward use Mixed Logit model is based on random coefficients. The probability for an individual n of choosing route i has the same form of the standard Logit, but it is conditional on the distribution of the coefficients :

$$P'(i) = \int \frac{e^{V_{in}(\beta)}}{\sum_{j \in C_n} e^{V_{jn}(\beta)}} f(\beta) d\beta \tag{7}$$

where $f(\beta)$ is a density function. In other words, Mixed Logit probabilities are the integral of standard logit probabilities over a density of parameters. Different distributions for the coefficient β used in the literature include uniform, normal, log-normal, and gamma distributions.
Further, a Mixed Logit model can be used without a random-coefficients interpretation, but simply representing error components that create correlations among the utilities for different alternatives.

The model formulations just described have been recently applied to bicycle route choice, using GPS traces as observations (for example by Menghini et al. 2010, Broach et al. 2011, Hood et al. 2011, Frejinger and Bierlaire 2007). As explanatory variables, the following were used: length, turns, proportion of path with bicycle facilities, slope and traffic lights.

2.3. References

[31] Ben-Akiva, M.E. and Bierlaire, M., 1999, Discrete choice methods and their applications to short-term travel decisions. Handbook of Transportation Science, pp. 5 -34.

[32] Ben-Akiva, M.E., Bergman, M.J., Daly, A.J., Ramaswamy, R., 1984, Modeling inter-urban route choice behaviour, Proceedings of the 9th International Symposium on Transportation and Traffic Theory. VNU Science Press, Utrecht, The Netherlands, 1984, 299-330.

[33] Ben-Akiva, M.E., Bolduc, D., 1996, Multinomial probit with a logit kernel and a general parametric specification of the covariance structure. Working Paper, Massachusetts Institute of Technology, Cambridge, USA.

[34] Bovy, P.H.L., 2009, On modelling route choice sets in transportation networks: a synthesis. Transport Reviews, 29(1), 43-68.

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

[36] Broach, J., Dill, J., Gliebe, J., 2012, Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A, 46, 1730- 1740.

[37] Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., 1996, A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In: Lesort, J.B. (Ed.), Proceedings of the Thirteenth

International Symposium on Transportation and Traffic Theory. Pergamon, Lyon, France, 697-711.

[38] Cascetta, E., and A. Papola, 2001, Random utility models with implicit availability perception of choice travel for the simulation of travel demand. Transportation Research part C, 9(4), 249-263.

[39] Dhakar, N. S., 2013, Route-choice modeling using GPS-based travel surveys. PhD Dissertation submitted at the University of Florida.

[40] Du, J., 2005, Investigating route choice and driving behavior using GPS-collected data. PhD Dissertation submitted at the University of Connecticut.

[41] Frejinger, E., Bierlaire, M., 2007, Capturing correlation with subnetworks in route choice models. Transportation Research Part B, 41, 363-378.

[42] Frejinger, E., Bierlaire, M., Ben-Akiva, M.E., 2009, Sampling of alternatives for route choice modeling. Transportation Research Part B, 43(10), 984-994.

[43] Griffin, T., Y. Huang, and S. Seals, 2011, Routing-based map matching for extracting routes from GPS trajectories. Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications, 24, 1-6.

[44] Halldórsdóttir, K., Rieser-Schüssler, N., Axhausen, K.W., Nielsen, O.A. and Prato, C.G., 2014, Efficiency of choice set generation methods for bicycle routes. European Journal of Transport and Infrastructure Research, 14 (4), 332-348.

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

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

[47] Papinski, D., Scott, D.M., 2011, Modelling home-to-work route choice decisions using GPS data: a comparison of two approaches for generating choice Sets. Presented at the Transport Research Board Annual Meeting, Washington D.C.

[48] Prato, C. G., 2009, Route choice modeling: past, present and future research directions, Journal of Choice Modelling, 2(1), 65-100.

[49] Prato, C.G., Bekhor, S., 2007, Modelling route choice behavior: how relevant is the composition of choice set? Transportation Research Record, 2003, 64-73.

[50] Quddus, M. A., Ochieng, W. Y., Noland, R. B., 2007, Current map-matching algorithms for transport applications: State-of-the-art and future research directions. Transportation Research Part C, 15(5), 312- 328.

[51] Rieser-Schüssler, N., Balmer, M. Axhausen, K.W., 2012, Route choice sets for very high-resolution data. Transportmetrica, 9, 825-845.

[52] Schüssler, N., Axhausen, K.W., 2009, Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT) – Working paper.

[53] Spissu, E., Meloni, I., Sanjust, B., 2011, Behavioral analysis of choice of daily route with data from global positioning system. Transportation Research Record: Journal of the Transportation Research Board, 2230, 96-103.

[54] Train, K.E, 2003, Discrete Choice Methods with Simulation, Cambridge University Press.

Chapter 3: Application to the first case study: the European Cycling Challenge in Bologna (Italy)

3.1. The GPS trips database

The GPS points available for this work have been obtained from the data collection made by SRM Bologna Srl and the Administration of Bologna, in the framework of the European Project CIVITAS Mimosa. The city of Bologna participated in the European Cycling Challenge (www.europeancyclingchalleg.org), along with other 11 cities in Europe, and 270 cyclists registered for participating to the challenge. Of these 270 registered cyclists, 150 were actually active for the whole data collection period. Incentives, such prizes, were given to the most active city and the most active users. The “scores” obtained by the participant cities are shown in Figure 8.

Participants were asked to register their bicycle trips using the application Endomondo (https://www.endomondo.com), during the month of May 2013. As can be seen in Figure 9, the application, once opened, allows users to declare the start and end of a trip, as well as the mode, and provides some information about the distance travelled, duration and speed.

The data was provided by SRM Bologna Srl in the form of a table in .csv format, where each row represented a GPS point recorded, and an identification number of the trip was also present. For each GPS point the date, time, longitude, latitude and altitude were recorded. The original sample contained over 1,050,000 GPS points, covering the entire road network of Bologna, and belonging to 9571 separated trips. Unfortunately, no identification number for users was provided, so was not possible to automatically individuate trip repetitions, nor to correlate trips with personal features of cyclists (such age, gender or bike experience).

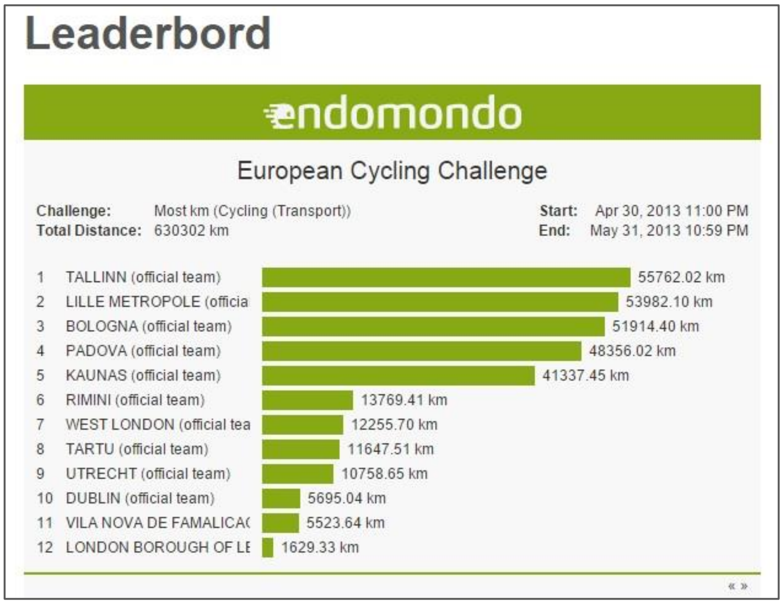


Figure 8: Results of the European Cycling Challenge for 2013 (Source: http://www.europeancyclingchallenge.eu/ecc2013/).

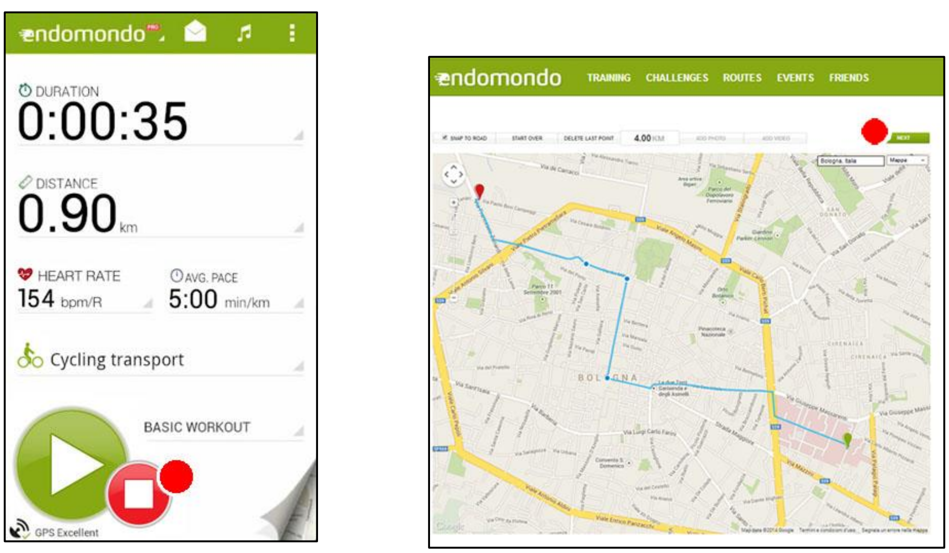


Figure 9: The Endomondo app, used for the data collection (Source: http://www.europeancyclingchallenge.eu/ecc2013/).

3.2. The network database

For the application of the algorithm to a real network, an Open Street Map network of Bologna has been used (Figure 10). The network consists of 10435 links with a total length of 1628.2 km of which 105 km are reserved roads for bicycles, see Table 6 and Figure 11. The Open Street Map data contains a wide variety of link types, and is compiled by volunteers. It represent a rich database, even though there are of course some links missing, and some with unknown type description.

Figure 10: Open Street Map network database of Bologna.

Link type	km	% of total
Cycleway	105.00	6%
Pedestrian	289.74	18%
Motorway	137.77	8%
Primary	61.91	4%
Secondary	86.07	5%
Tertiary	152.29	9%
Residential	437.30	27%
Service	217.01	13%
Construction	0.57	0%
Other	13.80	1%
Unclassified	126.69	8%

Table 6: Types of link in the Open Street Map of Bologna, in kilometers and percentage.

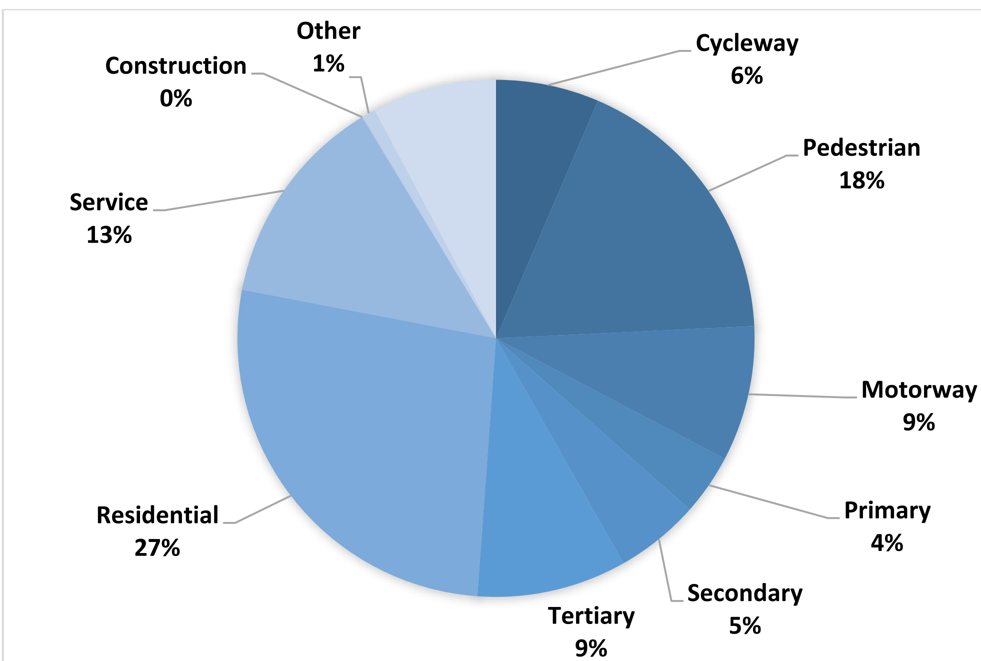


Figure 11: Types of link in the Open Street Map of Bologna.

3.3. The map-matching methodology

Journal Paper 2: The paper excerpted from this section of the study is first-authored by Dr. Joerg Schweizer, co-authored by the candidate, and co-authored by Professor Federico Rupi, and is published as:

Schweizer, J., Bernardi, S., Rupi, F. (2016). Map-matching algorithm applied to bicycle GPS traces in Bologna. ITE Intelligent Transport Systems, In Press.

A preliminary version of this paper was presented at the XXI Congress of Transport Professors and Researchers (SIDT 2015) in Turin (IT), in September 2015. The paper was then updated, rewritten and published from the ITE Intelligent Transport Systems, available online at http://ietdl.org/t/4Yyj9b.

3.3.1. Introduction

The use of GPS data in transport modeling has attracted considerable research interest in recent years. There has been a constant evolution in the use and analyses of geo-referenced data, e.g. demand models and route-choice models for cycling have been calibrated with geo-referenced topological characteristics and RP data from GPS recording (Aultman-Hall et al. 1997, Broach et al. 2012). But only the recent diffusion of smartphones has brought the data collection to an unprecedented quality level: on the one hand, citizens began to record their own commuter trips with the GPS functionality of their cell phones and to upload the traces on central databases. On the other hand, volunteers helped to expand and refine the geo-referenced maps, on open data platforms. For instance, the Open Street Map network provides a large number of link attributes which allow in depth analysis regarding the infrastructure usage and the cyclist behavior.

The combination of bicycle transport supply and transport demand data constitutes very valuable information for planners, because it allows to analyze the usage of the network and to identify specific criticalities. Such analysis can help the city to plan for an effective city-wide cycling network or to prioritize interventions and extensions of the present bicycle paths. In a broader context, the identified travel patterns of the population could be cross-valued with socio-economic and land-use geographic data in order to generate a virtual population and activity type models.

In order to make GPS data usable for transport modeling applications, a key processing operation is called map-matching, consisting of algorithms to identify the traveled road segments, called edges, based on a network model. When evaluating the performance of a map-matching procedure, most researchers refer to accuracy, meaning the percentage of correctly identified road segments. Nevertheless, due to the increase of data collection size and network scale, also the need to evaluate a map-matching procedure in terms of computational efficiency arises (Schüssler and Axhausen 2009). The challenge is to make algorithms faster, without losing accuracy.

Furthermore, the application of map-matching methodologies to bicycle trips registered in urban contexts presents specific issues. Firstly, in dense urban streets the partial absence of GPS points in a trip is more likely, due to the loss of signal of the geo-localizing device. The GPS devices only record data when certain requirements are met, in order to ensure the accuracy of the collected data (Harvey et al. 2008). Due to the blockage of buildings, mountains or vegetation, signals from GPS satellites can experience fading or delays (Costa 2011).This means performance can be degraded and the availability of high accuracy GPS-based location reduced. These problems may be particularly critical for urban canyon environments (Wang 2007). Secondly, cyclists do not necessarily cycle in roadways or on the provided bikeways, but they sometimes use pedestrian paths, prohibited ways, or they cross parks and squares. This higher degree of freedom characterizing bicycle trips makes GPS recorded data more difficult to match to the available road maps: in case of bicycle traces, the quality of the map-matching is particularly sensitive to the accuracy and details of the available road network model (Quddus et al. 2007).

In the present work an advanced map-matching algorithm is presented, based on maximizing the likelihood that a route is identical to the real route from where the GPS trace has been sampled. So-called buffers which encircle edges are used to determine the probability of finding GPS points near edges. For this reason the proposed method shall be called a buffer-based method. This approach takes into account all GPS points measured from the start to the end of the trip. Furthermore, the procedure makes use of network edge attributes to estimate the route in case of incomplete GPS data. Furthermore, the algorithm can identify whether the cyclist used a reserved bikeway, where available. The presented methodology has several important improvement in functionality and precision over previously-seen maximum likelihood methods (Marchal et al. 2005, Schüssler and Axhausen 2009).

The proposed map-matching algorithm has been applied to identify the routes of volunteer cyclists in Bologna, who recorded their GPS traces by means of smartphones.
This Chapter section is organized as follows: Paragraph 4.3.2 consists of a synthetic classification of map-matching methodologies which constitute the state of the art. In paragraph 4.3.3 a detailed description of the novel map-matching algorithm is provided. Paragraph 4.3.4 illustrates the application of the methodology to a real case study, i.e. to the GPS traces recorded by cyclists in Bologna: first, the context, the points and the network maps available are described; later in the paragraph, the criteria adopted to evaluate the results are described, a sensitivity analysis is carried out and a comparison with a reference topology-based algorithm is provided, to better enlighten the strength and the limitations of our work. In paragraph 4.3.5 some observations about the map-matching procedure applications are made and later in the Chapter, in section 4.4, a statistical description of the traces obtained from the map-matching process is provided.

3.3.2. Map-matching methods

The purpose of map-matching algorithms is to use GPS data and spatial road network maps to provide a positioning output which identifies the correct sequence of road edges (i.e. link) on which a vehicle travelled.
There are many map-matching algorithms proposed in literature, depending on the input data and on the application context. For an extended review of map-matching algorithms used in transportation research see Quddus et al. (2007). Map-matching algorithms have been categorized into three typologies: geometrical, topological and advanced. The geometrical approach is normally based on the distance between GPS points and network elements, edges or nodes (proximity-based algorithms): this approach is rather uncomplicated in terms of implementation, but the accuracy and the stability of such algorithms is rather scarce. In case of sparse GPS data, though, geometrical procedures can represent the only option (Xu et al. 2007). Topological approaches, other than the sole

distance, also take into consideration the sequence of GPS points and the connectivity of network elements. Different procedures have been tested (Velaga et al. 2009, Dalumpines and Scott 2011, Hudson et al. 2012, Quddus and Washington 2015). Most of them consist of a combination of geometrical approaches: first, the initial edge – or node – is found based on some geometrical criteria; then a set of likely edges is defined, for example using a buffer around the GPS trace; finally, the route is developed by choosing the most likely edge out of the set, again using different geometrical criteria. In terms of performance, topological approaches are generally better than geometrical ones, for both accuracy and computational efficiency. The main reason is that the topological approach evaluates only a reduced number of edges for each GPS point.

Finally, advanced map-matching procedures typically combine geometric and topological procedures with additional criteria, parameters, assumptions, probabilistic analysis, or optimization (Pyo et al. 2001, Marchal et al. 2005, Newsom and Krumm 2009, Schüssler and Axhausen 2009). For example, an advanced approach consists of taking into consideration the GPS measurement errors and constructing confidence regions around the points, based on data quality; then, only the edges within these regions are selected for evaluation. The evaluation can be based on a combination of different criteria, such as distance, speed or connectivity to the previously matched edges. The challenge for new and advanced map-matching procedures is to increase the accuracy while maintaining computational efficiency.

3.3.3. The buffer-based map-matching procedure

Each recorded GPS trace is a sequence of N points $P_j = \left(x_j, y_j, t_j\right), j = 1 \dots N$, where x_j, y_j represent the position and t_j the recording time of point P_j. The considered road network is given by the directed graph $G(A, V)$, where A is the set of edges and V is the set of vertexes or nodes. The task of map-matching is to identify the route R that correspond to the road links chosen by the cyclist who recorded the trace. Route $R = [a_1, a_2, \dots, a_M], a_i \in A$, is a sequence of M consecutive edges.

The basic concept of this map-matching algorithm is to define buffers B_a of width W around each network edge $a \in A$ and to count how many GPS points are contained in each edge buffer; the matched route R is the one that maximizes the number of GPS points, contained in the buffers of its edges. In practice, the number of points residing in buffer B_a is used to determine the weight attribute w_a for all $a \in A$. Then a Dijkstra algorithm (Dijkstra 1959) is used to find the route R from the origin edge to the destination edge that minimizes the sum of edge weights $\sum_{a \in R} w_a$.

In the following the notation $P_j \in B_a$ is used to express the condition that the respective point coordinates (x_j, y_j) must be within the boundaries of buffer B_a. As edge buffers often overlap, a single GPS point may be found in multiple edge buffers, for example at junctions, as illustrated in Fig.12. For this reason, a probability measure is used to count the GPS points in the edge buffer: if the GPS point P_j is located within m_j different edge buffers B_a, then the probability to find point P_j in either of these buffers is $p_j = {}^{1}/_{m_j}$. Consequently, the count of GPS points associated with edge a becomes the weighted sum $\sum_{P_j \in B_a} \frac{1}{m_j}$.

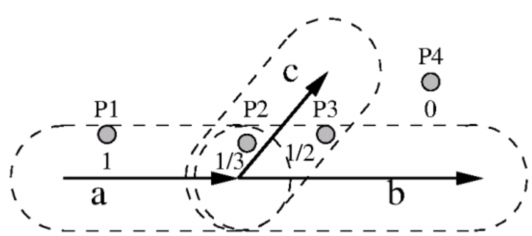


Figure 12: Illustration of edge weight and match error determination. In (a) four GPS points (P1-P4)and three network edges (bold arrows, labeled a,b,c) are shown together with their respective buffers(dashed line) . P1 contributes with $p_1 = 1$ to the edge weight of edge a, P2 contributes with $p_2 = {}^{1}/_{3}$ to edge weights of a,b,c, P3 contributes with $p_3 = {}^{1}/_{2}$ to edge weights of b,c, while P4 does not contribute to any weight .

As previously mentioned, the GPS traces may be interrupted and some edge buffers along the route may not contain any GPS points and a stable map-matching procedure should account for this. A further objective has been to "simulate" a cyclist preference for bikeways, when a bikeway edge exist and is adjacent to a roadway edge. The proposed method foresees to add the component $c_a\, L_a$ to the edge weight, which is proportional to the edge geometrical length L_a, while factor c_a is given by

$$c_a = \begin{cases} C_B C_L & if\ a\ \in A_B \\ C_L & otherwise \end{cases}.$$

Parameter C_L and C_B are called length factor and bike factor, respectively. Both parameters are positive, with values less than unity. $A_B \subset A$ is the set of network edges reserved for bicycles. This length-proportional element is expected to achieve two effects: (i) the routing algorithm will choose the minimum distance in the absence of GPS points, and (ii) routes containing reserved ways for bicycles will be preferred by the routing algorithm[4]. The final edge weight is determined by:

$$w_a = c_a L_a - \sum_{P_j \in B_a} \frac{1}{m_j} \tag{8}$$

Note that the presence of GPS points in edge buffers $\left(1/m_j\right)$ are counted negatively, while the edge length proportional part $(c_a L_a)$ is counted positively because the routing algorithms will *minimize* the route costs, which is the sum of its edge weights. Let R_{a_1,a_2} denote the minimum cost route with consecutive network edges between start edge a_1 and destination edge a_2 , then the associated minimum route cost becomes:

[4] It has been observed, especially for those urban contexts where the bicycle network is fragmented, poorly designed or mixed with pedestrian, that a cyclist preference for bikeway links cannot be taken for granted: the percentage of cyclists choosing to ride on the roadway, mixing with motorized traffic, rather than on a bicycle facility, can be relevant (Bernardi and Rupi 2015, Bernardi et al. 2016). Nevertheless, for the aim of the map-matching algorithm, we believe this assumption is reasonable and the use of the bike factor an advantage, as it can be eventually adjusted to observed attitudes in a specific context.

$$G_{a_1,a_2} = \sum_{\forall a \in R_{a_1,a_2}} w_a$$

(9).

In order to find the initial and destination edges of the GPS trace, the set O which contains potential start edges and the set D which contains potential destination edges are created. In practice, O is the set with all network edges whose edge buffers contain the first GPS point P_0 and D is the set with all network edges whose edge buffers that contain the final GPS point P_N, hence

$$O = \{a: P_0 \in B_a\}, \quad D = \{a: P_N \in B_a\} \tag{10}$$

The final matched route $R_{a_1^*,a_2^*}$ can be identified by calculating all minimum cost routs R_{a_1,a_2} with costs W_{a_1,a_2} between all start edges in O and all destination edges in D, and by selecting the route with the minimum cost $W_{a_1^*,a_2^*}$.

The flowchart of the map-matching algorithm is straight forward (see Fig. 13) and can be divided into 4 steps – repeated for each trace:

Step 1: Assign the weight $w_a, \ a \in A$ to each edge of the network according to Eq. (8).

Step 2: Identify the set of potential start edges O and the set of potential destination edges D, defined in Eq. (10).

Step 3: Determine the route/cost set $S = \{(R_{a_1,a_2}, \ W_{a_1,a_2}): \forall a_1 \ \in \mathrm{O}, \forall a_2 \in \mathrm{D}\}$, which are the minimum cost routes between all edge pairs in set O and D, using the Dijkstra algorithm for minimum cost path search (Dijkstra 1959).

Step 4: The matched route becomes $R = R_{a_1^*,a_2^*}$ if $W_{a_1^*,a_2^*}$ is the minimum cost in route set S.

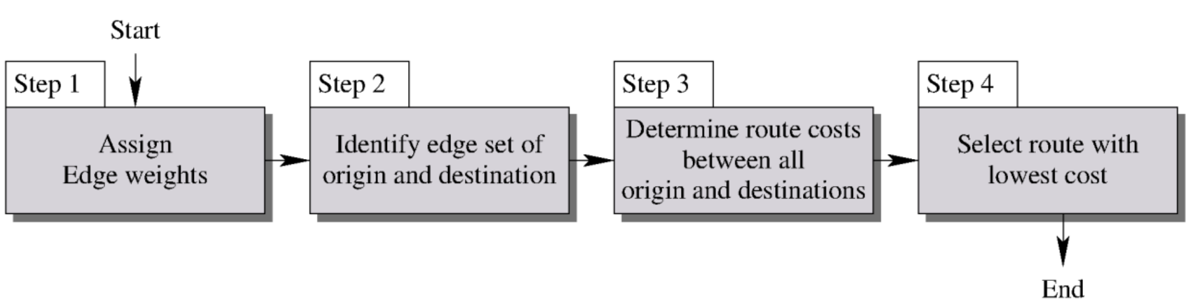


Figure 13: Simplified flowchart of algorithm.

3.3.4. Application of the map-matching algorithm to the case study of Bologna

For the application of the algorithm to a real network, an Open Street Map network of Bologna has been used which has been converted into a SUMO transport network in XML format (Krajzewicz et al. 2012). The unidirectional road network consists of 8484 edges with a total length of 660.61 km of which 72.67 km are reserved roads for bicycles, see Fig. 14. The Open Street Map data contains a large number of road attributes, such as access restriction, level, maximum speed, lanes, width etc., which are also imported by the SUMO network.

The GPS points available for this work have been extracted from a data collection, made by SRM Bologna Srl and the Administration of Bologna, in the framework of the European Project CIVITAS Mimosa. The city of Bologna participated in the European Cycling Challenge (www.europeancyclingchalleg.org) and hundreds of cyclists registered their bicycle trips, using the application Endomondo (https://www.endomondo.com), during the month of May 2013. For each GPS point the date, time, longitude, latitude and altitude were recorded. The original sample contained over 1,050,000 GPS points, covering the entire road network of Bologna.

No identification of the users has been provided. All traces with less than 10 total GPS points, a total duration of less than 30s and a total distance of less than 300 m have been eliminated in an automated pre-processing phase. In addition, traces with no GPS point inside the boundaries of the study area have also been eliminated. Overall, the eliminated traces prior to the map-matching correspond to 37% of all 9500 GPS traces provided by the complete database.

All algorithms have been implemented in the Python programming language, using excessively the array oriented Numpy package.

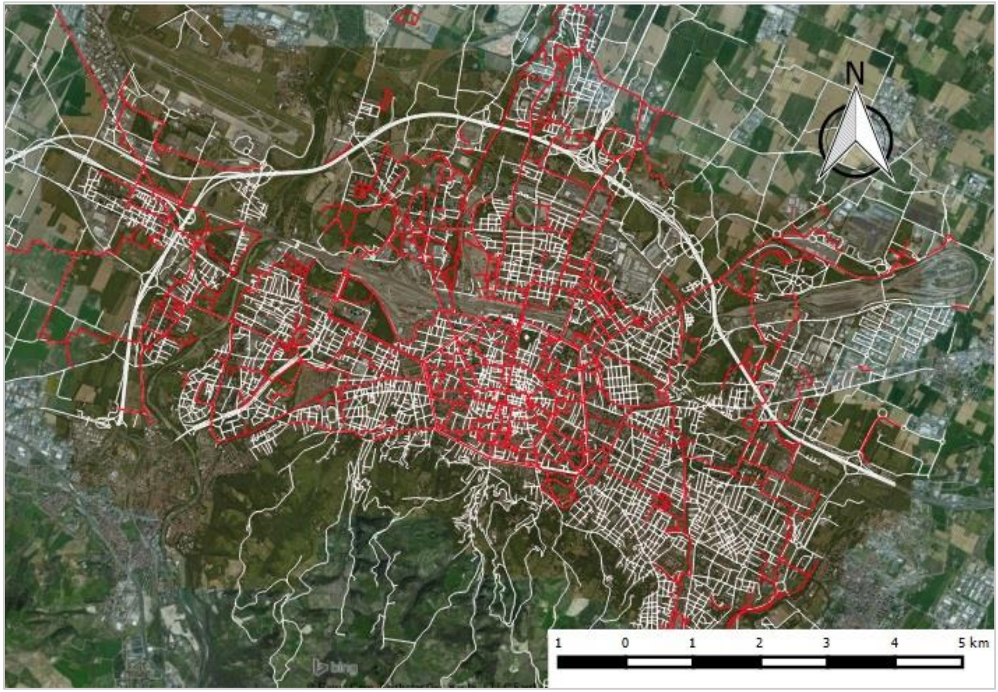


Figure 14: Road network (white) and bicycle network (red) of Bologna.

Evaluation criteria

The most common quality measure is the average distance error per point between trace in a sequence of N points $P_1, \ldots, P_N$, and the matched route R:

$$E = \frac{1}{N}\sum_{i=0}^{N} d_{min}(P_i, \varphi(R)) \tag{11}$$

where $\varphi(R)$ represent the geometrical shapes of the route edges, which is most commonly a multi-line or spline interpolation of consecutive points. The concept of the minimum distance is illustrated in Fig.15. However, as previously mentioned, the true routes made by the cyclists are generally unknown, which means the distance error is false if route R has been matched incorrectly.

In order to evaluate the credibility of the map-matching results, the length index I_L has been introduced. The length index is determined by dividing the length of the matched route, by the line-interpolated length of the GPS trace. More precisely, the length index I_L can be expressed by the fraction:

$$I_L = \frac{\sum_{\forall a \in R} L_a}{\sum_{j=1}^{N-1} \overline{P_{j+1}, P_j}}$$

(12)

where $\overline{P_{j+1}, P_j}$ is the Euclidian distance between points P_{j+1} and P_j. The reason behind I_L, is that the disturbances of the GPS points are of low noise (almost an offset) and do not significantly alter the length of the GPS trace.

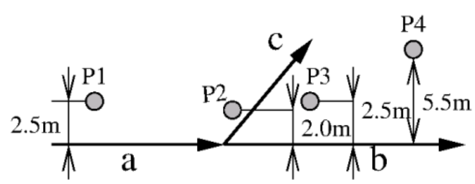


Figure 15: The GPS points and edges from (a) are shown, indicating the minimum geometric distances d_{min} between points P1…P4 and route R=[a,b]. In this specific case, the cumulative distances over both edges equals $2.5m + 2.0m + 2.5m + 5.5m = 12.5m$. The distance error becomes $E = 12.5/4$ m per point.

This means that I_L should be close to unity, otherwise the matched route contains detours ($I_L > 1$) or shortcuts ($I_L < 1$) with respect to the real route. Surely, the length index is not perfect in a sense that $I_L = 1$ does not necessarily mean a correct match: if the matched route contained shortcuts which compensate for detours in other parts of the route, then the length index would remain at one. Nevertheless, the investigated map-matching algorithms appear to systematically over- or under-estimate the true route length. The length index has also been a useful indicator check whether the different parameters of the match algorithm produce reasonable results.

Beside the previous indicators, it is important to monitor the share of route edges whose buffers contain GPS points. The hit index I_H is determined by dividing the number of route edges whose buffers contain at least one GPS point by the total number of edges in the route, hence:

$$I_H = \frac{N_H}{M} \tag{13}$$

where N_H is the number of non-empty buffers $B_a, \forall a \in R$ and M is the total number of edges in the route R. A low hit index means that a large part of the map-matching has been found by minimizing the route length (with edge weight $w_a = c_a L_a$) rather than by following GPS points.

The computing speed of map-matching algorithms can be quantified by dividing the computing time to match one trace T_M by the total length of the matched route. This specific match time $T_s = T_M (\sum_{\forall a \in R} L_a)^{-1}$, given in units of s/m, allows a fair comparison of different map-matching algorithms.

Sensitivity analysis

The buffer-based map-matching algorithm has two main parameters: the length constant C_L and the buffer width W. Sensitivity analyses of the bike constant C_B have not been considered in this work. If C_B has a positive value below 0.5 then links with reserved bikeway have almost always been part of the path, even in presence of a parallel ordinary road edge.

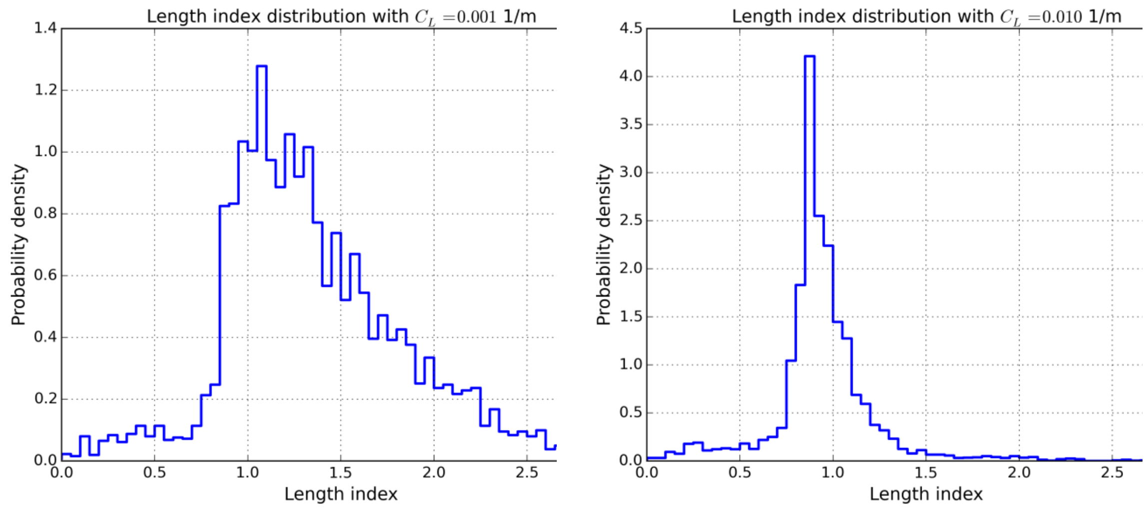
The constant C_L should be positive, but small as the GPS points count should be dominant in the determination of the edge weights (see Eq. 8). The evaluation results are shown in Tab. 7 and Fig. 16. With an increasing C_L, the hit index remains almost constant, the length index decreases, while the distance error increases. This behavior is expected, as the map-matching for low C_L is dominated by GPS points, and detours are not "punished" by a length proportional edge weight. The effect is that for low C_L, the matched routes often enter side-roads and return immediately. This unrealistic behavior can be seen in an increased share of high index length in Fig 16 (a). As C_L increases such effects vanish and the length index is more concentrated around unity Fig 16 (b). From Fig 16 it can be seen that the highest probability is at a value of $I_L = 0.9$ instead of one. This is most likely due to the disturbances of the GPS signal, which slightly extends the length of the poly-line made of GPS points.

In practice, matched traces with a too high length index or a too low hit index would be eliminated.

The buffer width Wshould be large enough to accommodate most GPS points in edge buffers. From observations of the Endomondo GPS traces we have seen that most points do not deviate more than 30m from the road. The evaluation results, varying the buffer width, with the same traces from the Endomondo database are shown in Tab. 8.

Length constant C_L [1/m]	Avg. matched road length per trip [m]	Avg. matched bikeway length per trip [m]	Avg. hit index I_H	Avg. length index I_L	Avg. distance error per point E [m/pt]
0.001	4371.5845	874.7982	0.8204	1.4733	56.7835
0.002	3906.2891	806.1498	0.8258	1.3233	56.0675
0.003	3585.1108	750.7404	0.8261	1.2160	56.7529
0.004	3377.0456	718.9130	0.8258	1.1495	57.1653
0.005	3228.6794	706.5380	0.8235	1.1050	58.1911
0.006	3137.4078	696.7614	0.8218	1.0768	58.8187
0.007	3072.7043	691.4565	0.8198	1.0544	59.7312
0.008	3028.0098	694.9206	0.8178	1.0402	60.4933
0.009	2989.6779	699.5974	0.8138	1.0279	62.1227
0.010	2956.4651	701.4289	0.8110	1.0183	63.9251

Table 7: Sensitivity analyses for length constant C_L. Values are averaged over 6500 traces from the Endomondo database.



(a) (b)

Figure 16: Histogram of length index I_L for (a) $C_L = 0.001$ and (b) $C_L = 0.01$.

Buffer width W [m]	Avg. hit index I_H	Avg. length index I_L	Avg. distance error per point E [m/pnt]
10.0	63.6297	1.0133	60.57
20.0	77.0484	1.0212	61.42
30.0	81.1366	1.0191	63.43
40.0	82.9902	1.0112	64.24
50.0	84.1820	1.0130	66.03
60.0	85.2123	1.0178	67.91

Table 8: Sensitivity analyses for buffer width W. Values are averaged over 6500 traces from the Endomondo database.

The main observation is that the hit index increases with an increasing buffer width, which is expected, because more GPS points will reside in edge buffers as they become larger. The downside is that the distance error increases as well with the buffer width. The reason is that with larger buffers, more buffer-overlaps at junctions occur, which increases ambiguity and the likelihood of a false route selection.

Performance comparison

In this section, the performance of the proposed algorithm is assessed with realistic parameters: $C_L = 0.01/m$, $C_B = 0.5$ and buffer width $W = 30m$. In addition, only the matched routes with a length index in the interval $0.8 < I_L < 1.2$ have been considered for evaluation. As demonstrated in Fig.16, this condition is satisfied by the vast majority of the traces. In addition, valid traces had to show a hit index $I_H > 0.8$. The chosen thresholds values for I_L and I_H are a compromise between obtaining quality results and thenumber of traces to considered valid. Clearly, a too severe limit on the quality would drastically reduce the number of valid traces.

In order to compare the performance of the proposed algorithm, the map-matching method from Marchal et. al (2005) has been implemented. This algorithm has been devised for similar purposes as the present algorithm, such as the processing of large quantities of GPS traces. The main principle of this topology-based algorithm is to find the route that minimizes the distance error E. The algorithm is iterative, as it starts with the first GPS point of the trace and N_R edges which show the smallest minimum distances to the point. These N_R initial edges constitute the origins of routes that “follow” the successive GPS points. While iterating over the GPS points, the routes are branched, according to the network topology, but only N_R routes with the lowest cumulative distance error are kept for a successive iteration. After the evaluation of the final GPS point, the route with the lowest distance error is selected out of the N_R final routes. This algorithm, which is described in detail in Marchal et al. 2005, has been implemented and tested with different values for N_R. For the comparison we use $N_R = 10$, as distance errors improve little by increasing N_R.

A comparison of the proposed and the topology-based algorithm is shown in table 3. Both algorithm used the same traces from the Endomondo database and have been applied using the same network model. However, the topology-based algorithm requires that the GPS traces are without significant gaps. For this reason, each trace has been split in two if it contained successive GPS points with a time difference greater than 30s, or a distance greater than 300m. Due to this procedure, the buffer based algorithm matched 3249 traces, while the topology-based algorithm matched 6786 shorter traces.

The distance error is asymmetricly distributed, similar to a log-likelyhood function, with a peak at 5m/pnt and 7m/pnt for the buffer and topology based algorithm, respectively. The average and standart deviation of the distance error of both matching algorithms are in the range of GPS recording precision.

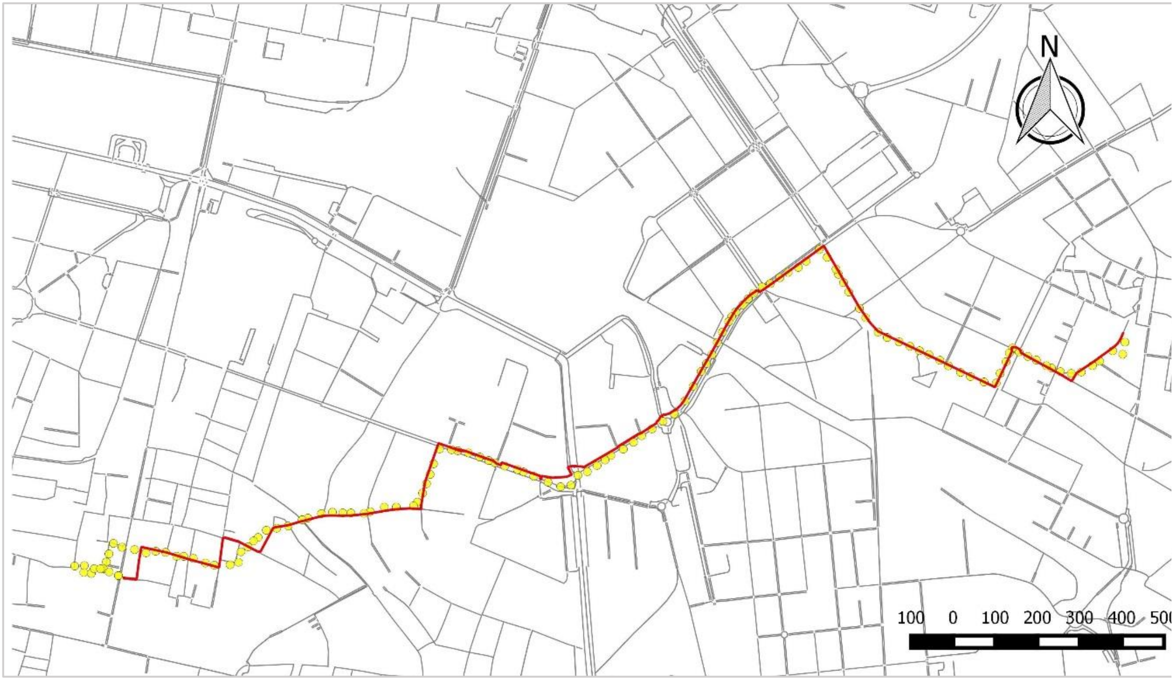
	Buffer-based algorithm	Topology-based algorithm
Average and standard deviation of length index I_L	0.9719 ± 0.0882	1.8887 ± 0.3521
Average and standard deviation of distance error per point E [m/pnt]	12.1363 ± 5.3859	15.5376 ± 8.4848
Average match time per matched meter T_s [ms/m]	1.6628	0.2499
Computation time for entire dataset on an AMD Athlone 3GHz, 1MB cash, single core processor [h]	16.6	2.5

Table 9: Comparison of present and topology-based algorithm on the same database and network.

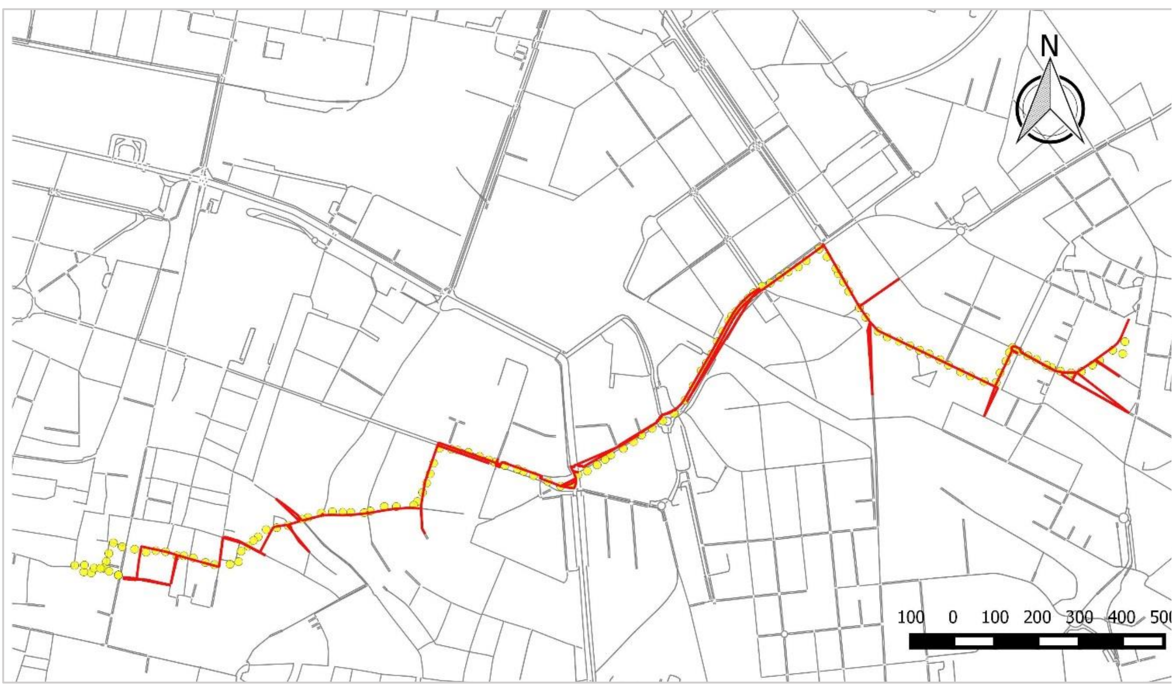
Regarding the length index, it is surprising that the topology-based algorithm produces a value of $I_L = 1.9$ instead of a value close to unity, despite the low distance errors. Note that I_L has a straight standart deviation interval for the buffer-based algorithm because all traces outside $0.8 < I_L < 1.2$ have been eliminated. While for the topology based algorithm, no such selection on I_L has been performed, because doing so, would have eliminated almost all matched traces (see averages and deviation of I_L in Tab. 9). This phenomenon shall be investigated further as it reveals a strength of the buffer-based map-matching algorithm: looking at the matched route of a particular trace one can observe, that the route identified by topology-based algorithm makes frequently “detours” into lateral roads. Figure 17 shows an example route, matched by the buffer-based and the topology-based algorithm.

The correct identification of routes at junctions appears to be an intrinsic problem for the considered topology-based match algorithm. It is beyond the scope of this paper to investigate the cause of this deviations, but it is fairly easy to find a rather typical example where a route, passing by the deviation into a latheral road, actually minimizes the distance error (see in Fig. 18).

Regarding the computing time, Marchal’s topology-based algorithm has been almost 7 times faster than the buffer based approach on the same computer. However, this computing time does not include the time required to rejoin the GPS traces with gaps (those which have been split previously). It has been suggested to perform rejoining with minimum distance routing (Marchal et al. 2005). With the buffer-based algorithm, the vast majority of computing time is needed to determine the edge weights w_a, because for each of the *N* GPS points, one needs to verified whether it resides in any of the edge buffers.

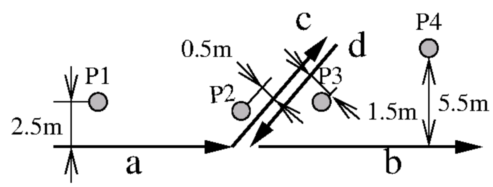


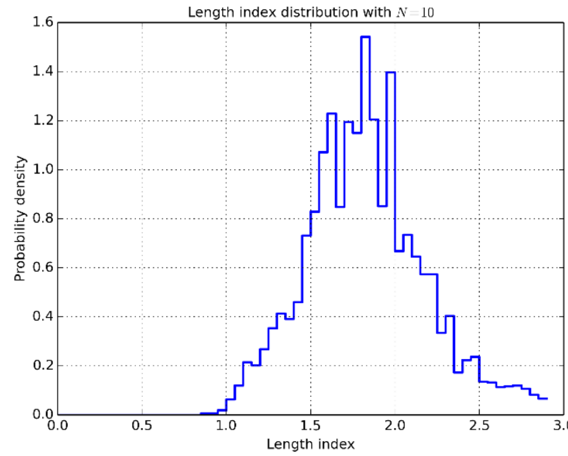
(a)



(b)

Figure 17: Map-matching algorithms of an example trace. Road network of a fraction of Bologna (gray), matched trace (red), GPS points (yellow). (a) Buffer-based algorithm (b) topology-based algorithm.





(a) (b)

Figure 18: Investigating detours at junctionsof the topology-based algorithm. (a) In this example, the "deviating" route $\boldsymbol{R} = [\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{b}]$ has a cumulative distance error of $\mathbf{10}\boldsymbol{m}$, while the straight rout $\boldsymbol{R} = [\boldsymbol{a}, \boldsymbol{b}]$e produces a greater cumulative distance error of $\mathbf{12.5}\boldsymbol{m}$ (see Fig. 12) . (b) Histogram with probability density of the length index using the topology-based map-matching, mean value is 1.9.

Furthermore, because an edge is made of one or several line segments, for computational reasons each edge buffer is generally composed of several linear buffers, each corrisponding to a line segment. This results in a complexity of order grater than $O(N \cdot |A| \cdot N_S)$, where N_S is the average number of line segments per edge. This means the computation time of the buffer based algorithm grows linearly with the number of GPS points and network edges while the routing between start and destination edges is relatively fast, as Dijkstra's algorithm has an order of $O(|A| + |V| \log \frac{|A|}{|V|} \log |V|)$ (Mehlhorn and Sanders 2008). Instead, the computation time of the topology-based algorithm depends on the number of GPS points *N* per trace and on the number of competing routes N_R, but not on the number of edges in the network $|A|$. For this reason, the topology-based algorithm is more scalable with respect to the network size, compared with the buffer-based algorithm.

Limitations

The buffer-based map-matching algorithm works well for most traces with the parameters $C_L = 0.01/m$, $C_B = 0.5$ and $W = 30m$. However, some of the matched routes have been significantly longer than the multi-line interpolation of the GPS points ($I_L > 1$). In most cases the reason is an imperfect network: the map-matching algorithm will always find a route between start and destination edge, as long as both are connected through the network. If, during the trip, some of the edges have no GPS points in their buffers, then the Dijkstra algorithms will bridge the missing pieces with the shortest path, no matter how many edges are not covered by GPS points. Unfortunately, this can also become a source of misinterpretation. In fact, the biggest routing errors occur, when edges are connected in reality, but are represented as disconnected in the network model. If a GPS trace runs over such disconnected edges, then the router tries to find the shortest connected alternative route. Because the routed alternative path can be seen as a “deviation” from the real path, the matched route becomes much longer than the real route. In case the trace of cyclists go through parks and footpath that are not present in the network model, the map-matching will also try to find the shortest route around the missing edges, considering only the edges that are present in the network. This means that, especially for bicycle trips, the network model used for map-matching must be as detailed as possible, including local and service roads as well as “unofficial” paths through parks. Another source of errors are strong offsets, which are able to “move” the GPS trace to neighbouring streets in a dense urban street grid. Such disturbances will inevitably lead to matching errors. Another limitation of the buffer-based map-matching algorithm is the processing time using very large networks, such as metropolitan areas or entire regions.

3.3.5. Summary

In this work a novel, buffer-based map-matching algorithm has been devised and applied to GPS traces collected by cyclists, and the road network of Bologna, Italy. Approximately

3000 (55%) traces could be reliably matched to the road network. The creation of buffers around network edges and the determination of the probability of GPS points inside the buffer, in combination with length – and type – specific edge attribues, has shown promising results, in particular in dense street grids and for different road types.

The reliability of the matched routes has been verified by comparing the length of the matched route to the length of the GPS trace. The ratio of both length measures has been termed length index, which should be approximately one for a good match. Sensitivity analyses for all important algorithm parameters have been conducted. The resulting distance errors are in the range of the GPS recording precision.

The algorithm generally satisfies the set objectives: the computing time is 1.6 milliseconds per meter, a 3km trace can be matched in less than 5s; the reliability of the routes has been verified even in presence of complex street grids and junctions; the “closing” of gaps caused by GPS signals blackouts is part of the algorithm itself and does not need to be performed externally; the proposed algorythm does also use edges with reserved bikeways, when available.

The performance of the proposed buffer-based matching algorithm has been compared to a topology-based algorithm, using the same traces, network and computer. The topology-based algorithm has 7 times faster execution times, but showed a length index of 1.9 due to unrealistic detours at junctions. Furthermore the investigated topology-based algorithm requires continuous GPS points, while gaps in the trace must be closed by an external algorithm.

Matching errors can also occur due to imperfections in the connectivity of the network. The level of detail in modeling cycling networks should follow the increasing level of detail that GPS data provides to analysts. There is still ample room for improvement of map-matching algorithms. In particular, an alternative definition of the edge weights could further improve matching reliability.

3.4. Descriptive analysis of the results

As it has been described in the previous sections of this Chapter, the database provided by the European Cycling Challenge for 2013 in the city of Bologna contained originally 9500 trips – approximately. The trips were separated and given a unique identification number, considering the start point and the end point as declared by users (no trip detection algorithm was necessary). Nevertheless, a filtering phase has been necessary prior the map-matching process. All traces with less than 10 total GPS points, a total duration of less than 30s and a total distance of less than 300 m have been eliminated in an automated pre-processing phase. In addition, traces with no GPS point inside the boundaries of the study area have also been eliminated. Overall, the eliminated traces prior to the map-matching correspond to 37% of all 9500 GPS traces provided by the complete database.

Subsequently, the map-matching procedure, as described in detail in the previous paragraph, has been applied to the filtered trip database. This process allowed to obtain approximately 3000 trips matched to the network, i.e. constituted by a sequence of links belonging to the network database.

A descriptive analysis of the trips resulting from the map-matching process has been performed.

First, we wanted to answer the question: when were the trips made? Figure 19 shows the distribution of the trips by day of the week. As it can be observed, the majority of the trips were made from Monday to Friday.

Figure 20 shows the distribution by hour of day of the trips. Two peak periods can be clearly identified: the morning peak from 7 to 9 am (with a strong single-hour peak at 8 am) and an evening peak from 5 to 7 pm.

Further, we analyzed trip durations (Figure 21): the majority of trips lasted less than 30 minutes, significant statistics are shown in Table 10. In the database, trips that lasted longer than 2 hours were 104, i.e. the 4%.

The distribution of trip length is shown in Figure 22, and significant statistics in Table 11.

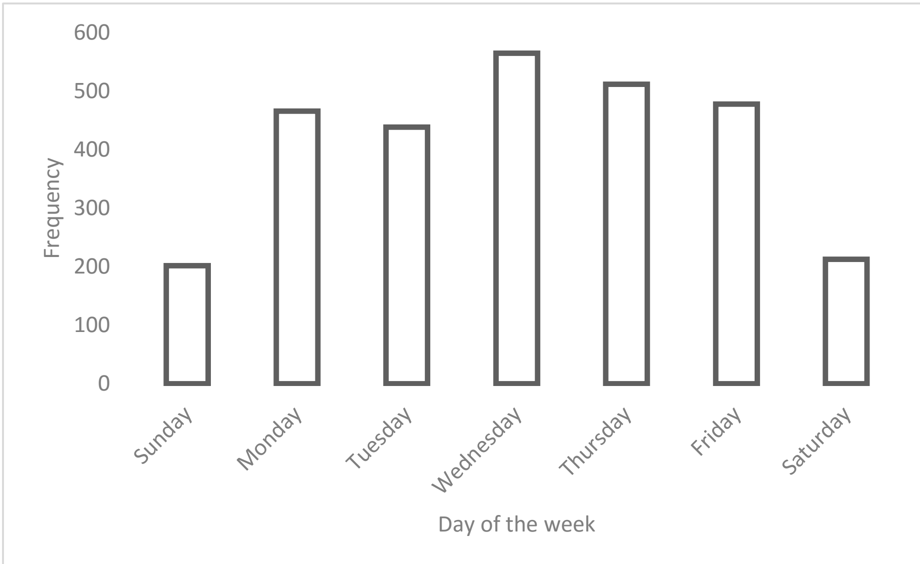


Figure 19: Distribution of trips by day of the week.

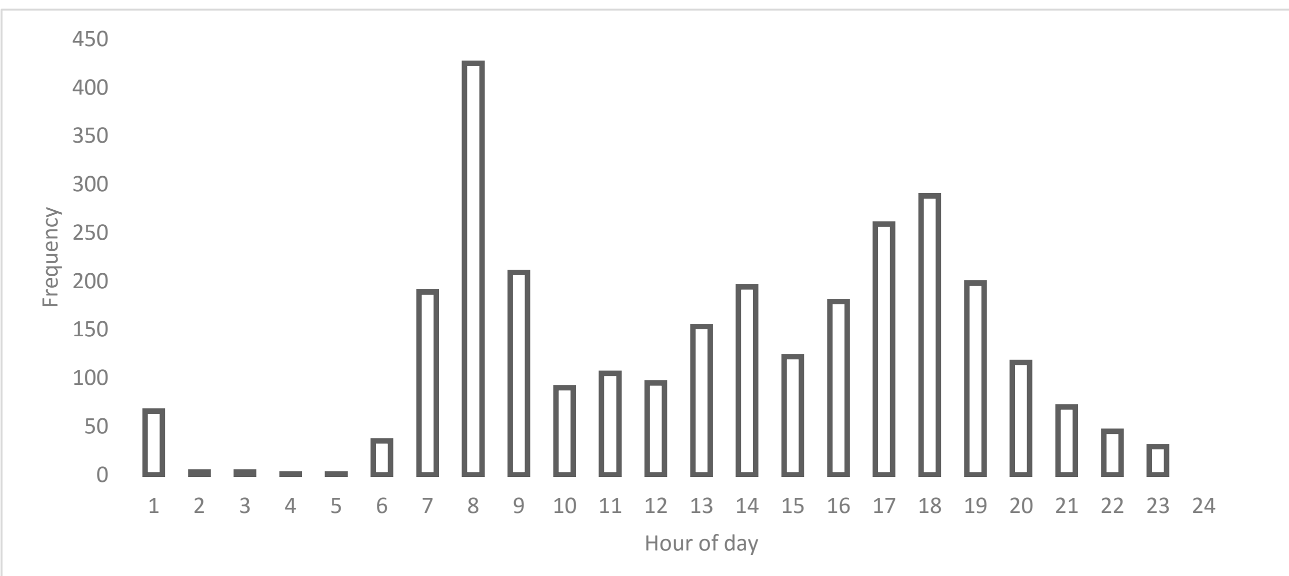


Figure 20: Distribution of trips by hour of day.

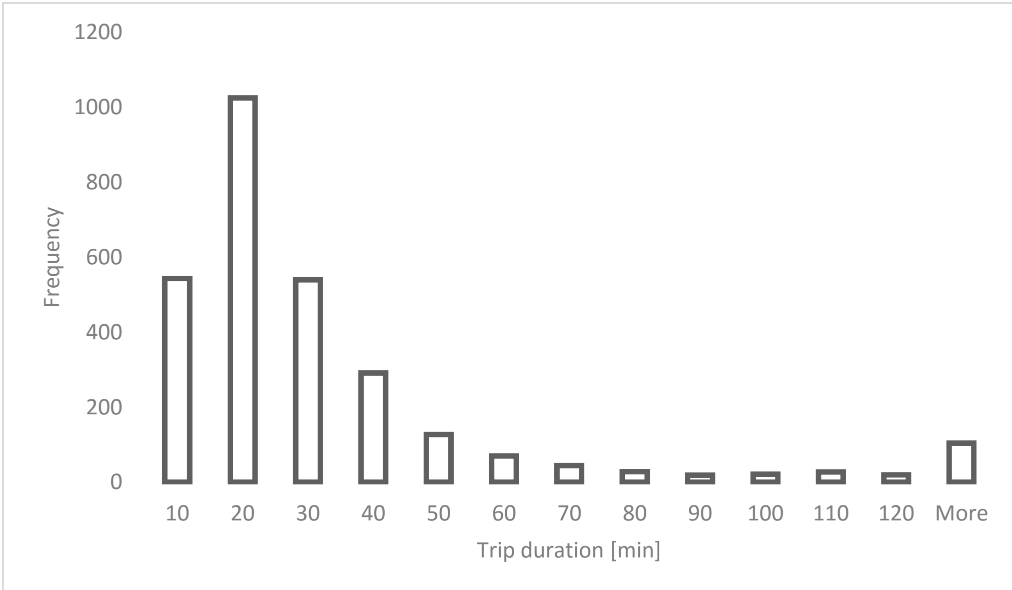


Figure 21: Distribution of trips durations.

Statistic	Minutes
Mean	32
Standard Deviation	52
Median	19
Mode	15
95th percentile	102

Table 10: Statistics for trip duration.

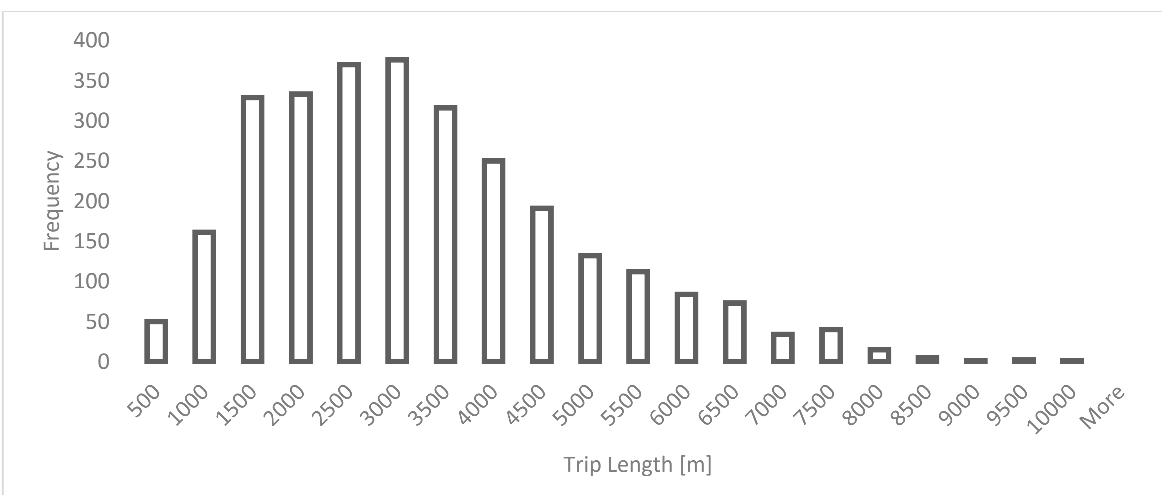


Figure 22: Distribution of trips length.

Statistic	Kilometers
Mean	3.00
Standard Deviation	2.74
Median	2.29
Mode	1.62
95th percentile	6.10

Table 11: Statistics for trip lenght.

From this first analysis we can deduct that the filtering and map-matching phases produced reasonable results. The purpose of trip was not declared by cyclists, but the distributions of trips by day of the week and hour of day suggest that a majority of trips were commuting trips. Nevertheless, we are aware that the database also contains trips made for errands, or for leisure: the latter are the most likely trip purpose for what concern longer trips, those with duration longer than the average.

Furthermore, the trips matched to the network database can provide another precious information, answering the question: which type of link did cyclists ride on? Indeed, the Open Street Map network database, as previously described, provides a classification by type of link. Table 12 shows the percentage by trip, on average, of each road typology, compared with the percentage of the same road typology offered by the network of Bologna. This comparison provides an indication regarding cyclists' preferences in terms of road typology. As can be observed, the most cycled roads are residential roads, which are also the most common in the network. Tertiary roads follow, proving to be preferred in cyclists' travels even more they do not represent a big part of the city's network. For what concerns cycleways, their usage results very low If compared to their presence in the network, and the same can be said for pedestrian paths. Finally, another interesting note should be made regarding unknown links (tagged "unclassified"): even though they represent the 8% of the total length of links in the network, they are not very frequently cycled by the participants to this study. This could confirm what has often been hypothesized, i.e. that for volunteered geographic information and crowd-sourced network database, the missing information are frequently located in remote parts of the territory, far from urban centers, and in general belongs for the major part to “unused” areas of the network.

Link type	% in chosen routes	% in network
Cycleway	2%	6%
Pedestrian	3%	18%
Motorway	2%	8%
Primary	12%	4%
Secondary	5%	5%
Tertiary	17%	9%
Residential	49%	27%
Service	1%	13%
Unclassified	3%	8%

Table 12: Percentage of type on link, comparing the chosen routes with the network.

Finally, a comparison between chosen routes and shortest routes should be made. Shortest routes have been calculated between all origin-destination pairs identified by the chosen routes database, using SUMO, considering length as the only parameter of link cost. For each trip, the path chosen by the traveler and the shortest option available were compared in terms of length, the difference in kilometers was calculated, as well as the difference in percentage. Such difference between chosen route and shortest route, in percentage, can be referred to as “detour”. Basically, it represent how much further than the “optimum” path cyclists travel.

We made the assumption that if the chosen trip length corresponded to the shortest length between the same OD pair, or it was longer by no more than 5%, then I could be said the cyclist chose the shortest path for travelling between that specific OD pair. Given this assumption, Figure 23 shows the percentage of choice for the shortest path, versus other options.

The distribution of the detour is shown in Figure 24 and 25, as well as significant statistics in Table 13.

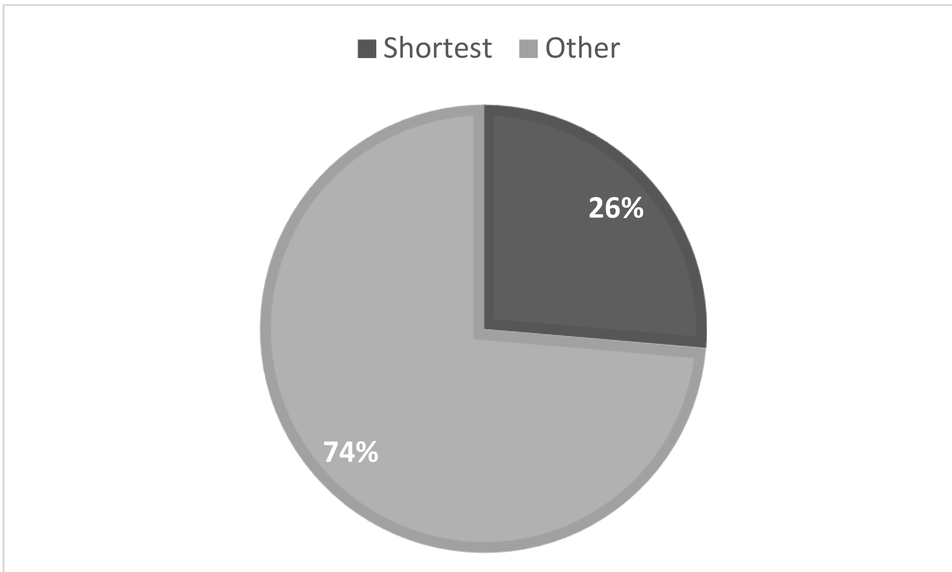


Figure 23: Percentage of choice, divided by shortest route or other.

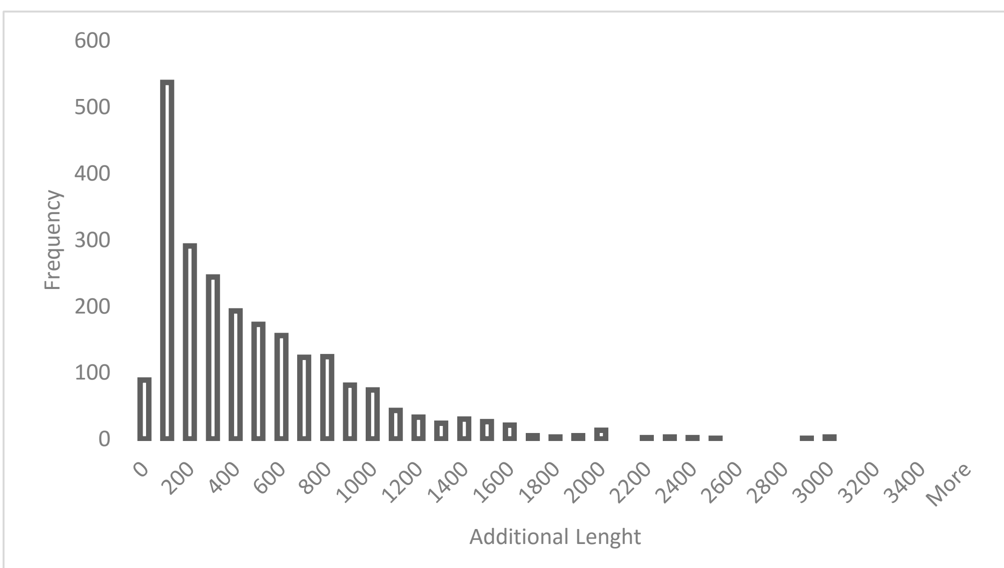


Figure 24: Distribution of the additional length (detour) in meters.

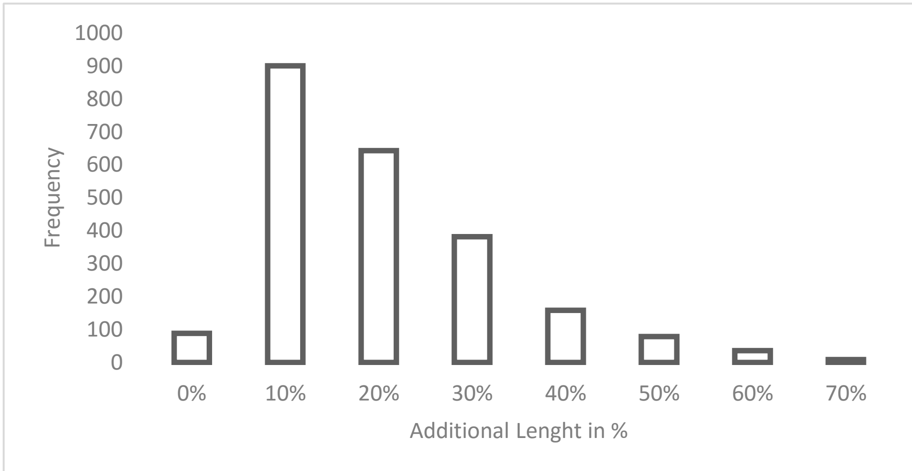


Figure 25: Distribution of the additional length (detour) percentage over the total trip length.

Statistic	Detour [m	Detour [%]
Mean	421.99	15%
Standard Deviation	295.19	13%
Median	0.00	0%
Mode	425.70	22%
95th percentile	1308.53	42%

Table 13: Statistics for detour.

It should be said, though, that many could be the explanatory factors for the detour: cyclists could choose to cycle longer than the optimum – in percentage – in order to cycle on specific bicycle facilities (but, given the percentage of kilometers cycled on cycleways, that does not appear to be the case). In alternative, the purpose of the trip would be a precious information to include in this specific aspect of the analysis: to be more accurate, a distinction between commuting trips and leisure trips should be made. Nevertheless, if we look at the distribution of percentage of detour along the week, and along the day (Figures

26 and 27), detour tends to increase and to variate more for Saturdays and Sundays, as well as for non-peak hours of the day (10 am to 4 pm).

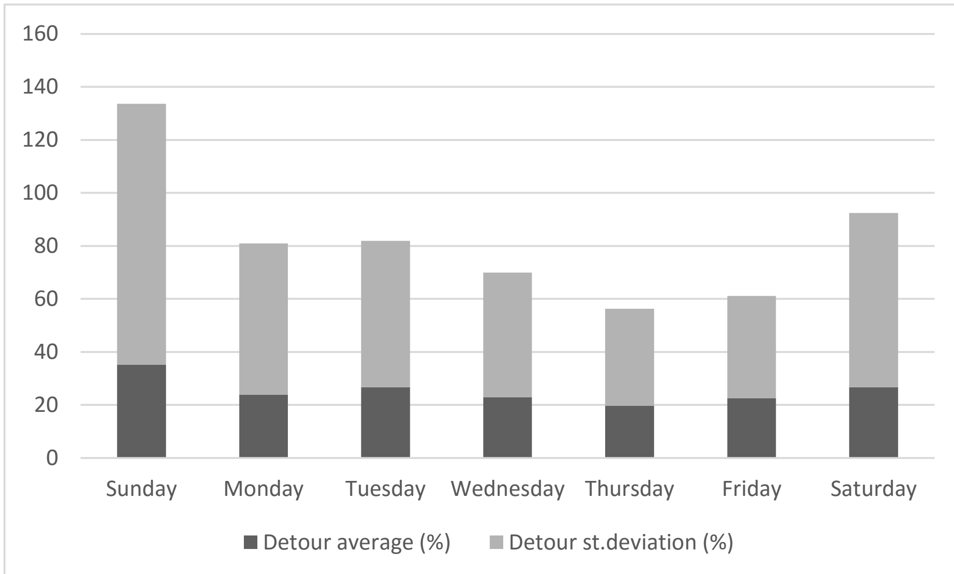


Figure 26: Average detour (in %) by day of the week, and corresponding standard deviation.

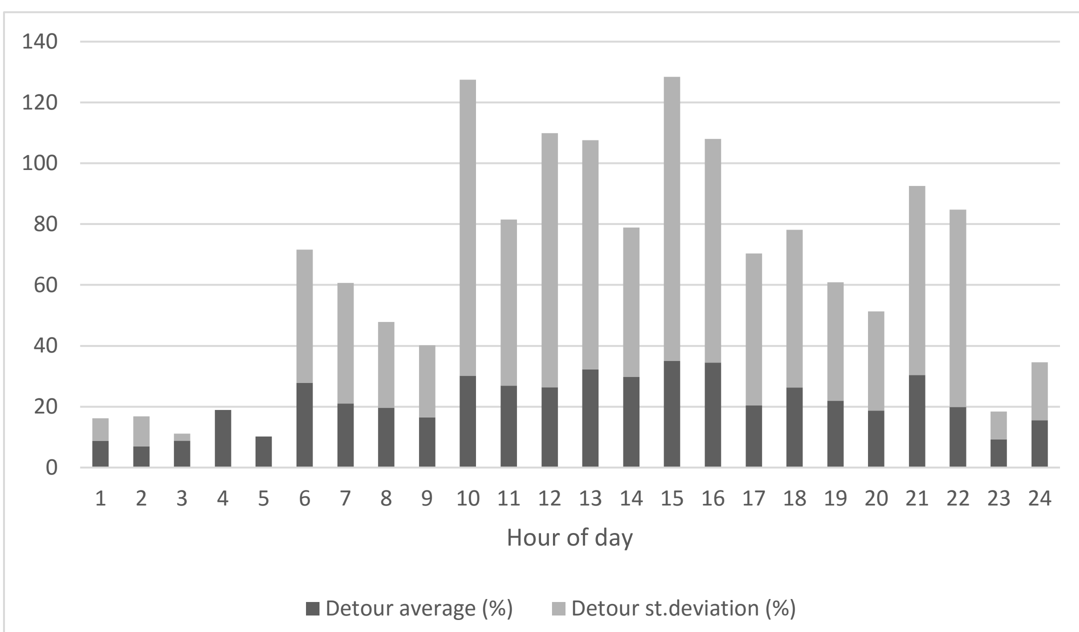


Figure 27: Average detour (in %) by hour of day, and corresponding standard deviation.

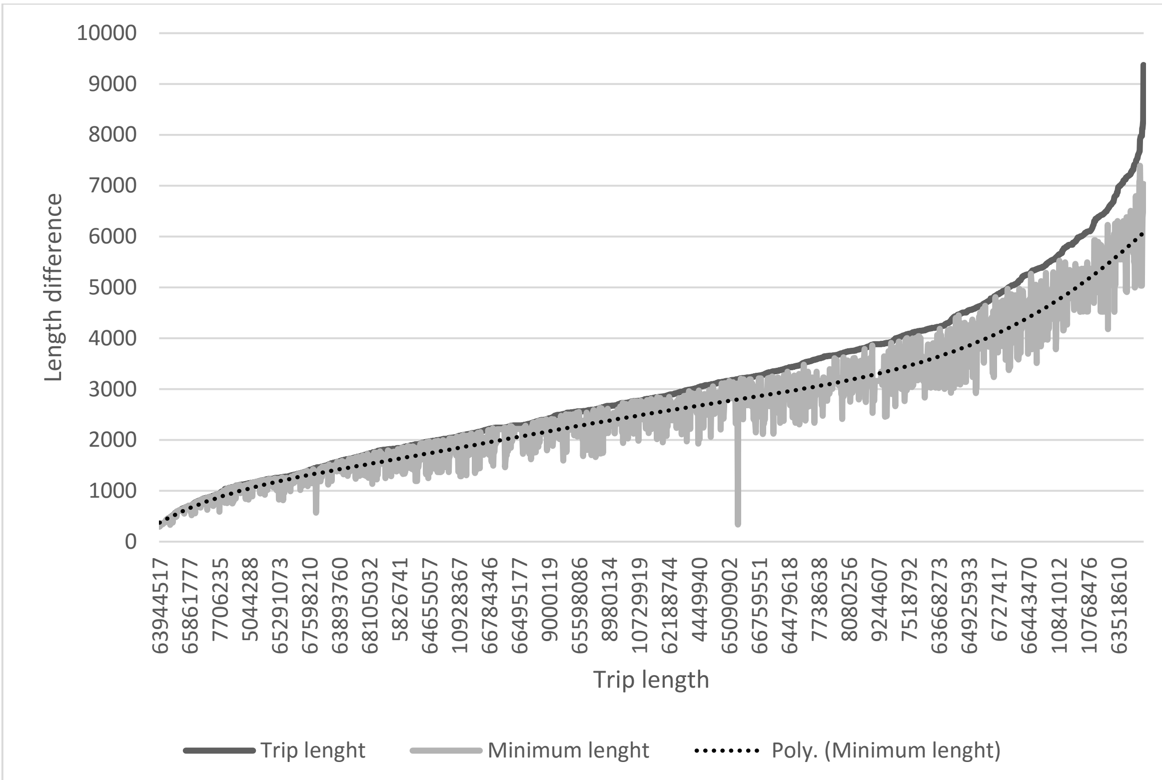


Figure 28: Difference of trip length and minimum length (shortest path).

Finally, Figure 28 shows that the difference between trip length and the corresponding shortest path's length, increases as the trip length increases. This result has also been found in other studies (e.g. by Dill and Gliebe 2008), and it can be explained considering that longer trips are more likely to be trips made for leisure of physical exercise purposes, and in these circumstances it results fairly reasonable the assumption that cyclists do not tend to "optimize" their trip, as minimizing travel time is not on the main objective for them.

3.5. References

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

[56] Dill, J., Gliebe, J., 2008. Understanding and Measuring Bicycling Behavior: a Focus on Travel Time and Route Choice. Oregon Transportation Research and Education Consortium, Portland, OR.

[57] Aultman-Hall, L., Hall, F., Baetz, B. (1997), Analysis of Bicycle Commuter Routes Using Geographic Information Systems: Implications for Bicycle Planning. Transportation Research Record, 1578 (1), 102 – 110.

[58] Schüssler, N., Axhausen, K.W. (2009). Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT) – Working paper. Swiss Federal Institute of Technology Zurich.

[59] Harvey, F., Krizek, K.J., Collins, R., (2008). Using GPS data to assess bicycle commuters route choice. Proceedings of the Transportation Research Board 87th annual meeting.

[60] Costa, E. (2011). Simulation of the Effects of Different Urban Environments on GPS Performance Using Digital Elevation Models and Building Databases. IEEE Transactions on Intelligent Transportation Systems, 12(3), 819 – 829.

[61] Wang, J. H., Gao, Y. (2007). High-Sensitivity GPS Data Classification Based on Signal Degradation Conditions. IEEE Transactions on Vehicular Technology, 56(2), 566 – 574.

[62] Quddus, M. A., Ochieng, W. Y., Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the-art and future research directions. Transportation Research Part C, 15(5), 312 – 328.

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

[64] Xu, L., Minglu, L., Wei, S., Minyou, W. (2007). A practical map-matching algorithm for GPS-based vehicular networks in Shanghai urban area. IET Conference Publications, 533 CP, pp. 454 – 457.

[65] Dalumpines, R., Scott, D.M., 2011, GIS-based map-matching: development and demonstration of a post-processing map-matching algorithm for transportation research. Lecture Notes in Geoinformation and Cartography, 101-120.

[66] Hudson, J., Duthie, J., Rathod, Y., Larsen, K., Meyer, J., 2012, Using Smartphones to Collect Bicycle Travel Data in Texas. University Transportation Center for Mobility. http://ntl.bts.gov/lib/45000/45700/45731/Hudson_11-35-69.pdf

[67] Velaga, N. R., Quddus, M.A., Bristow, A.L., 2009, Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems. Transportation Research Part C, 17, 672- 683.

[68] Quddus, M. A., Washington, S., 2015, Shortest path and vehicle trajectory aided map-matching for low frequency GPS data. Transportation Research Part C, 55, 328-339.

[69] Newsom, P., Krumm, J., 2009, Hidden Markov Map Matching Through Noise and Sparseness. ACM GIS 2009. Seattle, WA.

[70] Pyo, J., Shin, D., Sung, T., 2001, Development of a map matching method using the multiple hypothesis technique. IEEE Proceedings on Intelligent Transportation Systems, 23-27.

[71] Dijkstra, E.W., 1959, A note on two problems in connexion with graph. Numerische Mathematik, 1 (1), 269-271.

[72] Bernardi, S., Krizek, K.J., Rupi, F. , 2016, Quantifying the role of disturbances and speeds on separated bicycle facilities. Journal of Transport and Land Use, 9(2), *In Press*.

[73] Bernardi, S., Rupi, F, 2015, An Analysis of Bicycle Travel Speed and Disturbances on Off-street and On-street Facilities. Transportation Research Procedia, 5, 82-94.

[74] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker., 2012, Recent Development and Applications of SUMO - Simulation of Urban MObility". International Journal on Advances in Systems and Measurements, 5 (3&4), 128- 138.

[75] Mehlhorn, Kurt; Sanders, Peter, 2008, Algorithms and Data Structures: The Basic Toolbox. Springer.

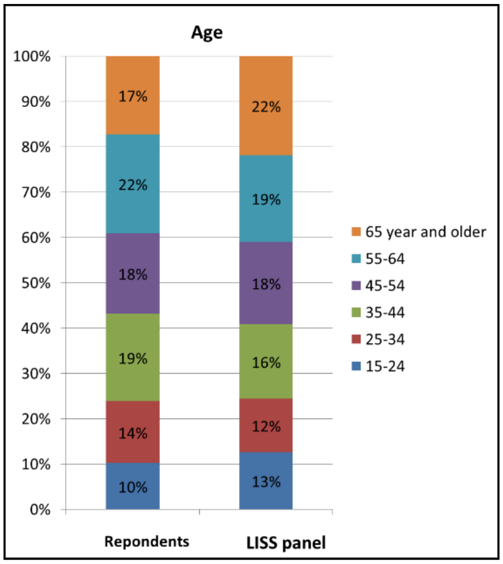
Chapter 4: Application to the second case study: the Mobile Mobility Panel in the Netherlands

4.1. Description of the database

4.1.1. The GPS trips database

The available database consisted of over 50,000 trips (28 million GPS points) recorded all over the Netherlands, for a period of four weeks between April and May 2014. About 600 respondents were recruited from an existing online panel, the Mobile Mobility Panel, providing a true representation of the Dutch population. The respondents were asked to register all their trips through a smartphone application specifically elaborated for this data collection, called MoveSmarter (for IPhone and Android). A true representation has been achieved also by distributing smartphones to those respondents who did not own a smartphone with a suitable GPS positioning device. For instance, also an older segment of population was included in the panel (Fig. 29).

During the monitoring period, respondents also participated in a web-based prompted recall survey to check and revise trip characteristics, delete or add trips if necessary (Fig. 30). For each registered trip, also information such as origin, destination, mode and purpose has been checked by users and registered; this has been useful to "correct" the automatic trip split and mode detection operations following the data collection stage.



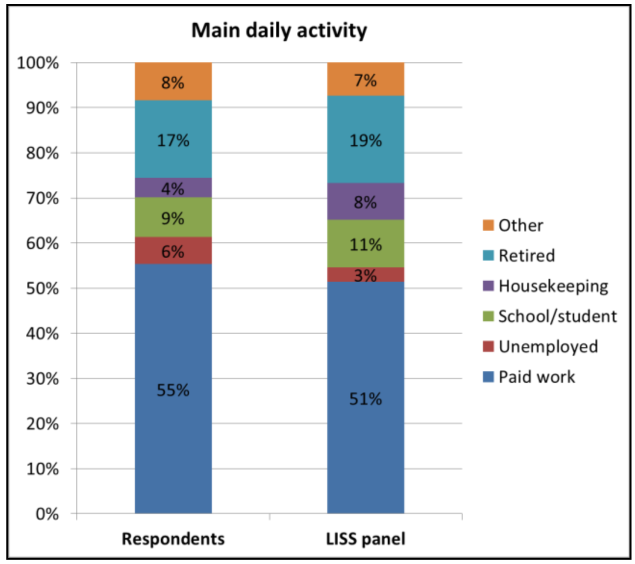


Figure 29: Age and activity statistics of the population sample of the Dutch Mobile Mobility panel (Source: Geurs et al. 2015).

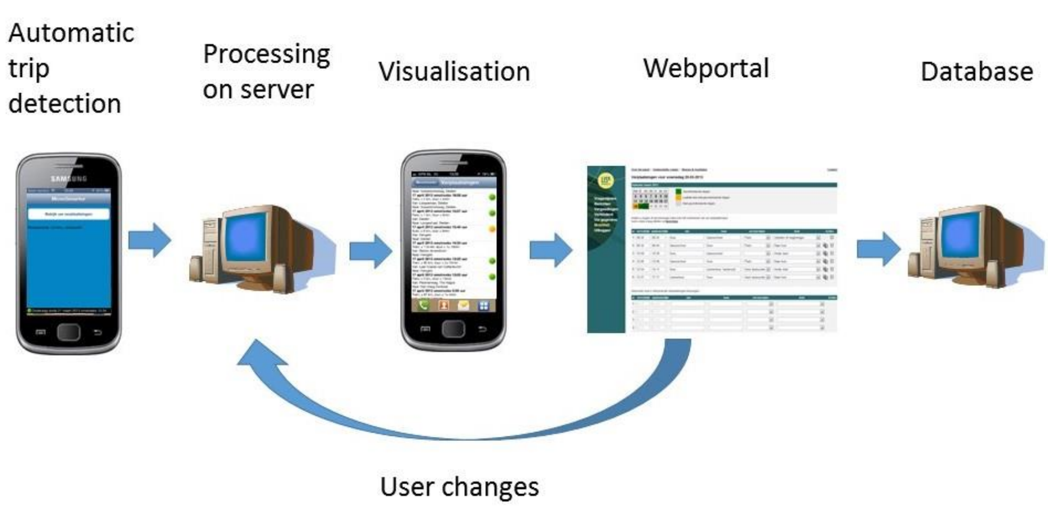


Figure 30: Schematic representation of the data collection process (Source: Geurs et al. 2015).

For more information about the data collection and the automated trip detection see Geurs et al. (2015). To the author's knowledge, this is the first application of a smartphone app as a measurement tool for travel surveys for such a large representative sample and over such a long period of time. Furthermore, a rich set of socio-economical information was available for each participant: sex, age, family composition, home location, activity type and location, income etc.

For the application described in this paper, we selected only the trips made by bike, so the size of the database used for this study is approximately 4,000 trips registered by 280 users.

4.1.2. The network database

The network database available for this application is the bicycle network provided by Fiersersbond, a Dutch cyclist union which realized different types of bicycle-related study and initiative in the Dutch territory and, thanks to volunteers, annually updates a representation of the bicycle network throughout the whole country, including roadways (only those accessible for cyclists, i.e. motorways are not included) and all types of bicycle facility available. A wide range of attributes are associated to each link, including geometrical features (length, width), type of facility, type of pavement, quality, beauty of the context, illumination etc. If compared with the Open Street Map database for the Netherlands, the amount of missing information is lower that for the Open Street Map network, both in terms of missing links or connection, and in terms of attributes (percentage of unknown/blank of approximately 20%, while this percentage reaches 30% in the Open Street Map network database for the Netherlands). Overall, the Fietsersbond network consists of more than 1,5 million links, for a total length of over 180,000 km (Figure 31). It should be said, though, that also the Open Street Map network database for the Netherland was taken into consideration; first, as will be mentioned in the following paragraph, the Open Street Map network was used for the map matching. Furthermore, it was also used in order to compare the descriptive analysis performed for the Dutch traces database and the Italian one – where only the attributes of link type as classified by Open Street Map volunteers were available. To attach the attributes of the Fietsersbond network database to the Open Street Map links, an ArcGIS model based on the spatial join tool has been designed for the purpose. Basically, the model searched iteratively, for every chosen link belonging to the Open Street Map network, the corresponding link belonging to the Fiestersbond network. The correspondence between two links belonging to the two different network datasets was based not only on proximity, but also on other topological criteria.

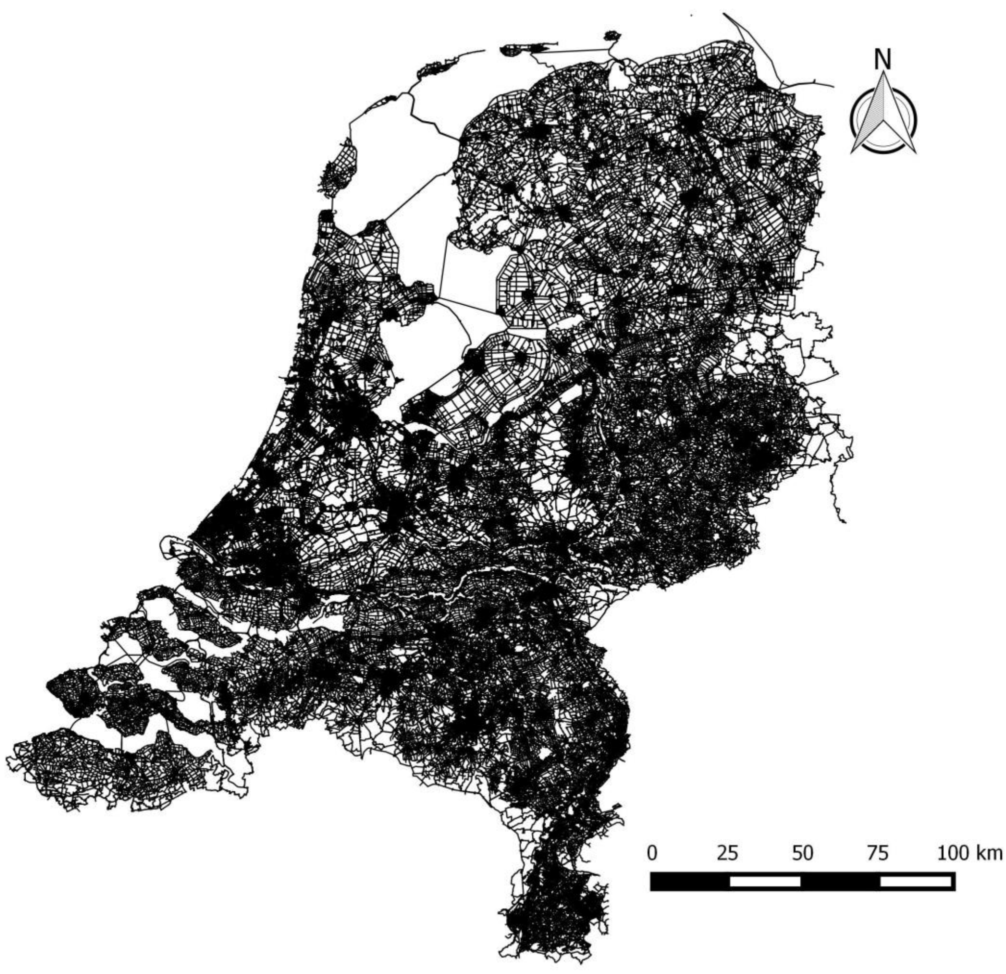


Figure 31: The network database for the whole Dutch territory.

4.1.3. GPS Post-Processing

GPS data as recorded by smartphone applications need a considering amount of post-processing operation, in order to become available for further analysis of revealed preferences. In general, the post-process consists of individual stages accounting for data filtering, trip and activities detection, mode stage determination, mode identification, map-matching (Schuessler and Axhausen, 2009). In this case, such operations were made by Mobidot Srl. As for the data filtering, the trip and the mode detection stages, details on methodologies and results are available in Thomas et al. (2015). As already pointed out in the data collection methodology description, participants also filled out a web-based travel diary, so that trips features such as origin, destination, time, purpose and mode as automatically detected could be verified by the information provided by participants. Finally, the GPS points were matched to the network using the map matching algorithm described in Marchal et al (2004). The map-matching was made by Mobidot Srl as part of the post-processing phase, using the Open Street Map network. For this study the trips were provided as already matched to the network, i.e. in the form of one text file (.csv) for each trip, containing the sequence of chosen links.

4.2. Descriptive analysis of the results

4.2.1. Statistics

Figure 32 shows the percentages of bike trips made in the monitoring period by the participants, grouped by purpose: the trips made for work and study related reasons constitute a 12% of the total, while a 25% of trips were made for leisure (including in the category physical exercise and enjoying the view, as these were other purposes the application proposed to participants. Shopping trips were the 11% of the total, while other small number of trips were labeled by users with other available purpose categories that can fall into the class of ”errands” (visiting people, bring people somewhere etc.). A significant percentage of trips were categorized as “going home”, and it is very likely that part of these include going home from work; unfortunately it is not possible for all these trips to determine where participants went home from, as only for a few number of these trips participants provided a label for the origin location.

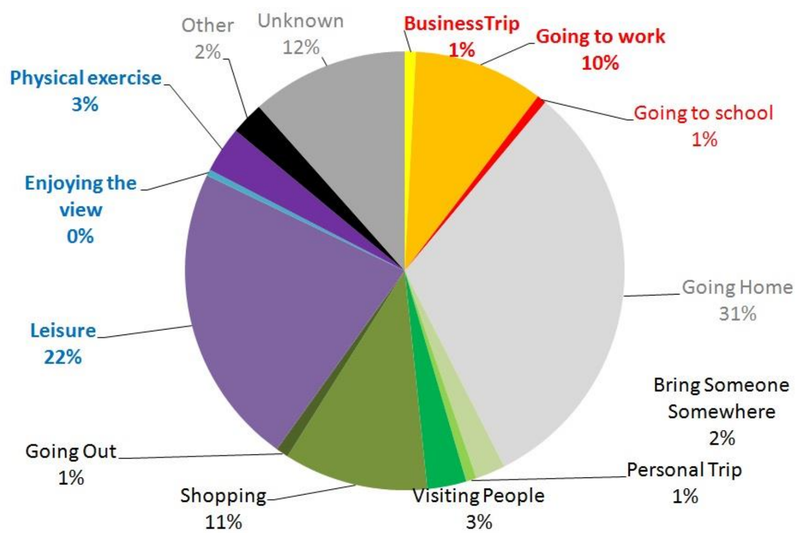


Figure 32: Percentages of bike trips by purpose.

Regarding the number of trips made by user, the majority of participants made up until 30 bike over the four weeks of data collection (Fig. 33): the sample of respondents includes systematic cyclists, commuters who cycle everyday to work and less frequent cyclists. Compared to other similar data collected in different contexts – e.g. the Italian database analyzed in the previous Chapter – the number of bike trips per person are higher, but this could be expected if one consider the high percentage of trips that the Dutch population make by bike.

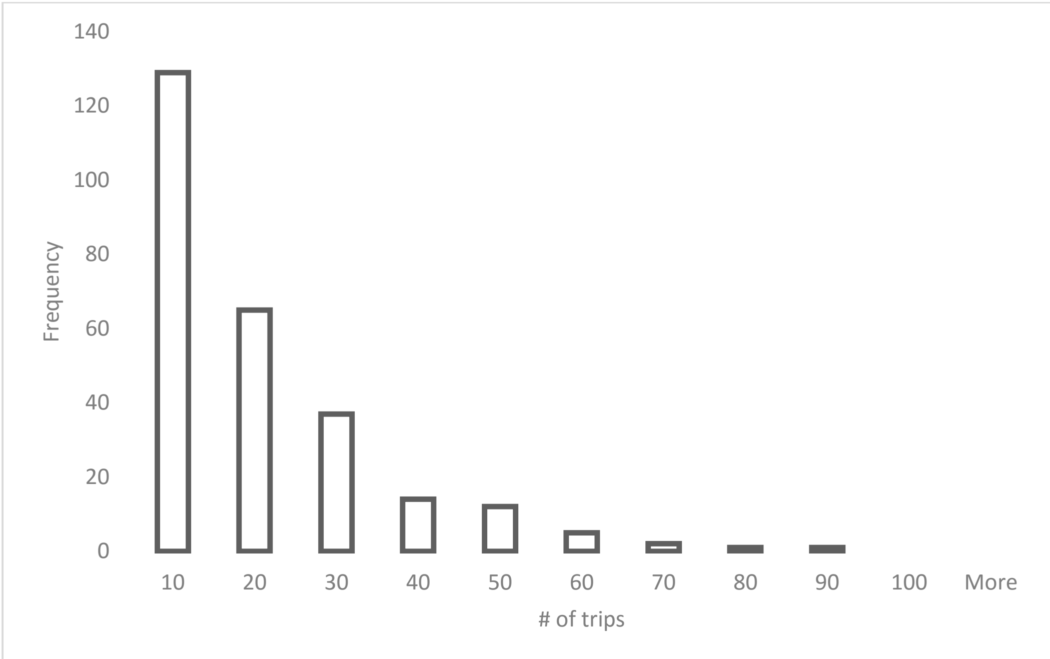


Figure 33: Number of trips per users.

Another interesting statistic to describe the cycling habits of a population is the distribution of trip lengths as a function of a measure of impedance to travel (typically measured in the form of distance) (Taylor 1975; Luoma et al. 1993), i.e. the distance-decay curve. As the name implies, distance has a decaying effect on the likelihood of cycling between two locations.

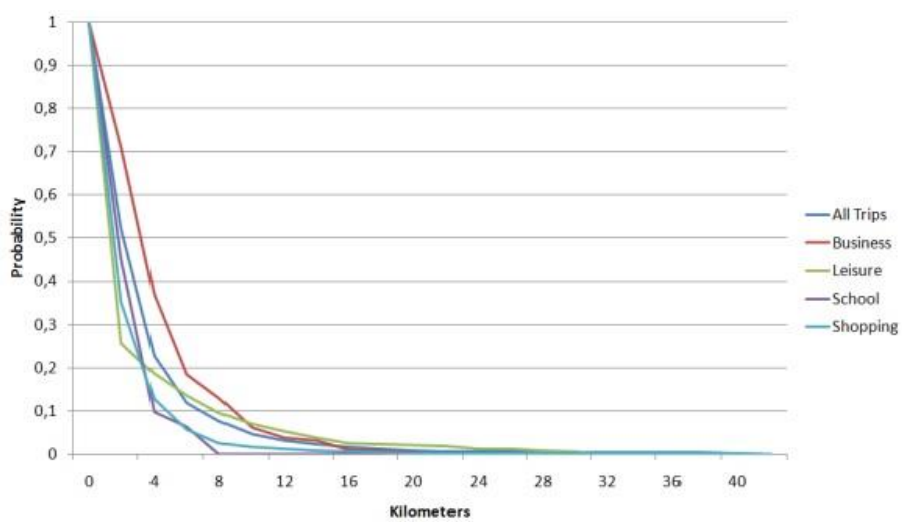


Figure 34: Distance-decay curves for different trip purposes.

The relationship between distance and probability to cycle is affected by a vast number of factors, connected to both users and to the cycling network, but also to the urban context, and the land use. It is fair to expect that the shape of the function vary by the purpose of the trips (Krizek et al. 2007). The curves obtained for this application are shown in Figure 34.

Regarding cyclist preferences, Table 13 and 14 shows the cycled kilometers grouped by type of link, whether we considered the link typologies of the Open Street Map network (Tab. 13), or those of the Fietsersbond network (Tab. 14).

Link type	Km cycled on type of link	%	Km in the network database	%
Cycleway	7450.73	37.40%	33017.7	14.4%
Unclassified	5452.48	27.40%	77221.4	33.7%
Residential/Tertiary	3936.39	19.80%	34544.4	15.1%
Footway/Pedestrian	1048.73	5.30%	17257.7	7.5%
Secondary	1049.27	5.30%	8140.0	3.6%
Path	397.72	2.00%	11806.5	5.2%
Primary	392.04	2.00%	6433.5	2.8%
Service	140.12	0.70%	13152.7	5.7%
Motorway	24.36	0.10%	6855.3	3.0%
Planned/Under construction	4.13	0.02%	615.3	0.27%

Table 13: Kilometers cycled by link type (OpenStreetMap Network).

The first table gives an interesting result in quantifying the use of bicycle facilities (labeled as “cycleway”) by the population sample. Such a high percentage of trip length made on cycleways is in line with the well-known bicycle-oriented travelling habits of the Dutch context, as well as with the high number of bicycle facilities provided in the Dutch cities: it is fair to assume that a high-level offer of bicycle facilities translates in a high number of cyclists using them. Another result to underline is the high percentage of trips made on links that are labelled as "unclassified". As already mentioned previously, the detail of the network to which GPS points are matched is a key factor for every analysis coming from smartphone travel data, especially in terms of attributes describing the network's links. The higher the number of "unclassified" links, the higher the loss of information for a specific trip.

Link type	Km cycled per type of link	%	Km in the network database	%
Normal roadway	6205.4	36.7%	89880.5	48.3%
Path along road	3360.6	20.6%	22827.2	12.3%
Bike lane	1890.0	11.2%	7143.5	3.8%
Exclusive bike path	1982.7	11.7%	14627.6	7.9%
Service road	478.1	2.8%	2963.3	2%
Bicycle boulevard[5]	134.3	0.8%	172.6	0.1%
Pedestrian	119.5	0.7%	1836.9	1%
Unknown	802.7	4.7%	13337.2	7.2%
Blank	1948.7	11.5%	32633.7	17.5%

Table 14: Kilometers cycled by link type (Fietsersbond Network).

As we mentioned in the previous paragraph, the network provided by the Fietsersbond Dutch Cyclist Union is richer in those attributes describing the perceived quality of bicycle

[5] With the expression "bicycle boulevard" the author indicates a road that have been optimized for bicycle traffic: this means the road is open to motorized vehicle but their posted speed is reduced (normally to 30 km/h), and signals indicate that the right-of-way belongs to bicycle. Such type of link is not so frequent in the Italian context, while it is diffused in the Netherlands.

paths. In Table 15, Table 16 and Table 17 we illustrated the kilometers cycled by respondents, grouped in terms of link quality, link beauty and traffic nuisance perceived on links.

Link quality	Km cycled per quality class	%	Km in the network database	%
Good	9593.8	56.7%	87856.4	47.2%
Fair	4019.7	23.8%	41726.8	22.4%
Low	270.4	1.6%	5253.9	2.8%
Unknown	952.1	5.63%	15600.3	8.4%
Blank	2086.7	12.3%	35639.85	19.1%

Table 15: Kilometers cycled by link quality level (Fietsersbond Network).

Link beauty	Km cycled per beauty class	%	Km in the network database	%
Neutral	8317.0	49.1%	77578.1	41.7%
Beautiful	4935.0	29.2%	51179.4	27.5%
Picturesque	330.1	1.95%	2707.9	1.5%
Ugly/boring	303.6	1.79%	3459.1	1.9%
Very ugly	6.9	0.1%	94.1	0.05%
Unknown	944.2	5.6%	15419.1	8.3%
Blank	2086.7	12.3%	35639.9	19.2%

Table 16: Kilometers cycled by link beauty level (Fietsersbond Network).

Link traffic nuisance	Km cycled per nuisance class	%	Km in the network database	%
Little	6795.0	44.9%	77720.2	41.8%
Reasonable	3049.0	20.4%	23776.9	12.8%
Very little	1831.0	10.8%	30347.4	16.3%
Much	798.3	5.4%	2963.6	1.6%
Very much	117.2	0.5%	237.4	0.13%
Unknown	3879.0	5.6%	15392.1	8.3%
Blank	1098.4	12.3%	35639.9	19.2%

Table 17: Kilometers cycled by traffic nuisance level (Fietsersbond Network).

4.2.2. Traveled distance versus shortest distance

A first evaluation of cyclists' behavior consists in the comparison between the paths that result to be chosen by them and the minimum cost path in terms of length. This gives a measure of how cyclists' trips were "optimized". Shortest paths have been calculated on the network model using ArcGIS, considering length, and excluding highways and motorways, as they do not represent an available option for cycling.

Comparing the chosen routes with the shortest paths, on average, by trip, cyclists cycled 1.37 km longer than the minimum cost path, 15% in percentage of trip length. Figure 35 and 36 show the distributions of the difference between chosen routes and shortest paths length, in meters and as a percentage of trip length.

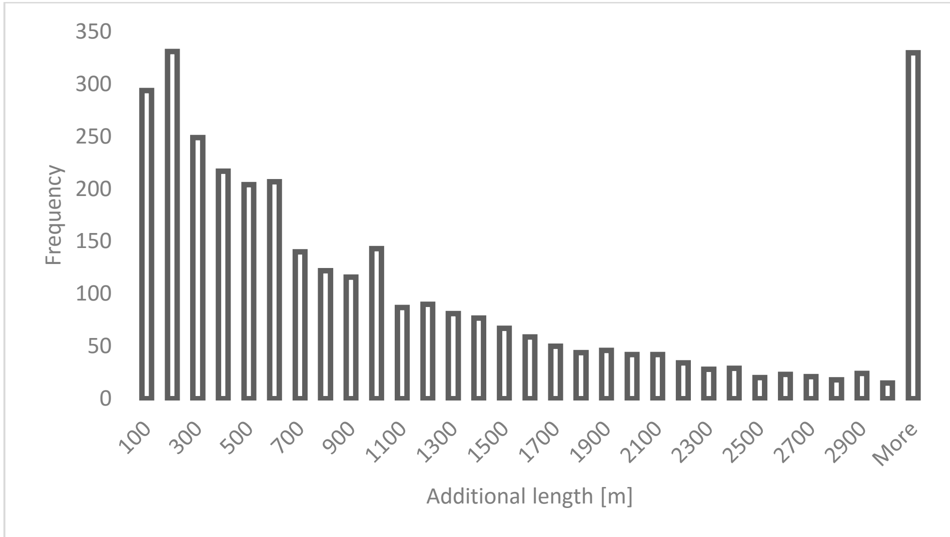


Figure 35: Distribution of the difference between chosen routes and calculated shortest routes.

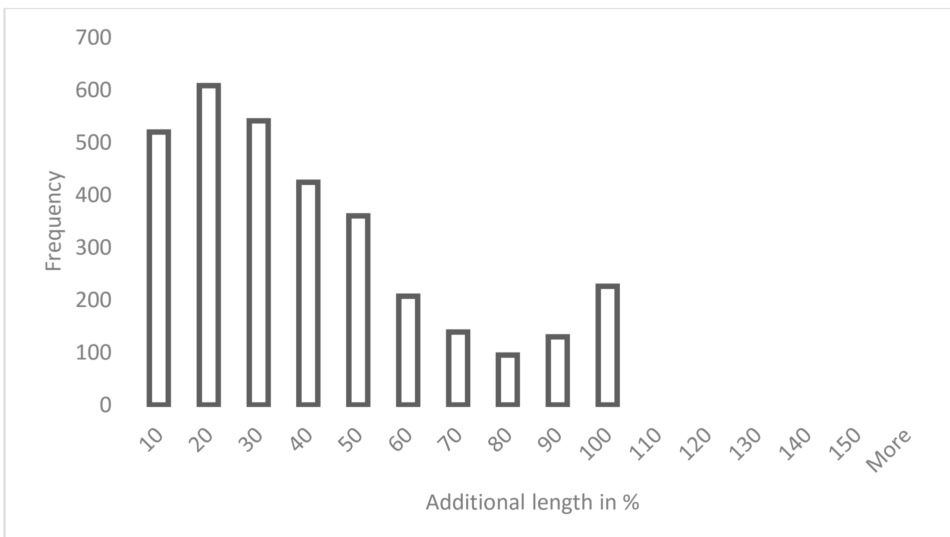


Figure 36: Distribution of the difference between chosen routes and calculated shortest routes, in percentage of trip length.

The “optimization” in terms of length of cyclists’ trips has also been categorized by purpose: one should expect that, for work trips, people tend to choose a shorter path as they would do when they are travelling for leisure, or for exercise. Nevertheless, this did not prove to be true for this specific study case, as the percentage of trips done on the shortest path varies around 50% and does not show significant deviation for specific trip purposes.

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

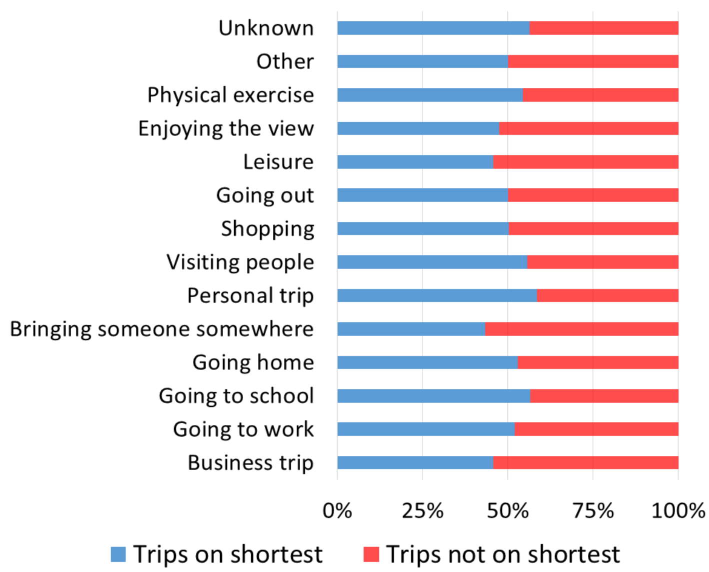


Figure 37: Percentages of trips on shortest path by purpose.

4.3. Choice set generation

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

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

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

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

Table 18: Alternatives and their choice rates.

4.4. Model estimation

4.4.1. Discrete Choice Modelling Framework

When dealing with route choice, the independence of irrelevant alternatives (IIA) property of the Multinomial Logit (MNL) model makes it inappropriate for estimating discrete choices among similar alternatives, such as different paths available for a cyclist between an OD pair in an urban context. In such case, an overlapping path may not be perceived as a distinct alternative. For bicycle route choice modelling this problem has been overcome by introducing a similarity measure in the utility function, as done by Ben-Akiva and Bierlaire (1999) for the so-called Path Size Logit model. In the route choice context, indeed, a path often contains links that are common for several paths. Hence, the size of a path with one or more links in common with another may be less than one. Ben-Akiva and Bierlaire has included a path-size factor PS_{in}in the utility function of a path, i.e.:

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \epsilon C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}} \tag{14}$$

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

Thus the corrected utility function is:

$$U_{in} = \mathrm{ASC} + \beta \cdot x_{in} + \varepsilon_{in} \tag{15}$$

where x_{in} is the vector of route attributes, β is the vector of parameters to estimate and ε_{in} are the error components.

As for the error component, we referred to the formulation of the Mixed Logit model (Train, 2002), in order to account for correlation in unobserved factors over time.

The model thusly obtained for the probability of choosing path i is:

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

(16).

4.4.2. Binomial logit model

The first model we estimated was a binomial mixed logit model, where the two alternatives are represented by the shortest route (Alt0) and the chosen route, when this differed from the shortest.

We made the assumption that if the chosen trip length corresponded to the shortest length between the same OD pair, or it was longer by no more than 15%, then it could be said the cyclist chose the shortest path for travelling between that specific OD pair. Given this assumption, Figure 38 shows the percentage of choice for the shortest path, versus other options.

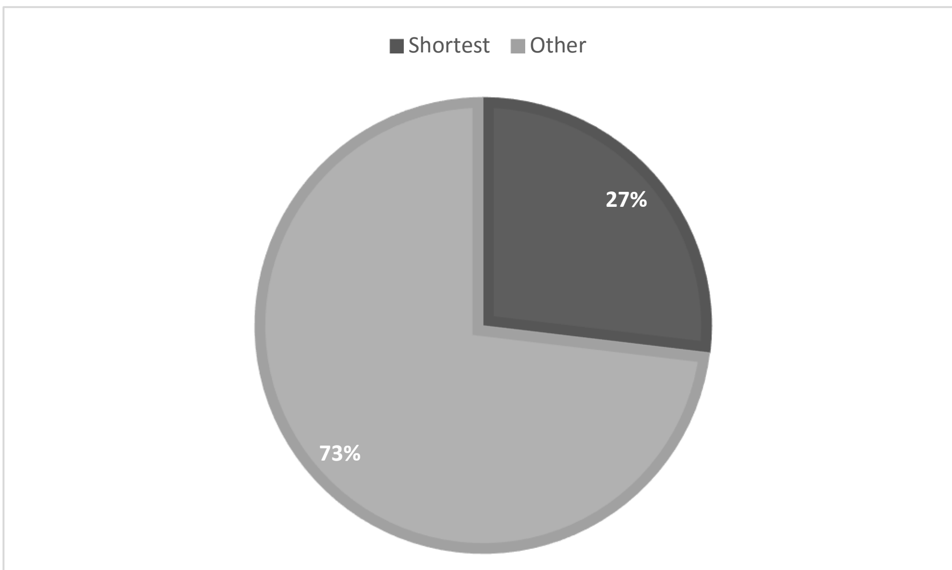


Figure 38: Percentage of choice, divided by shortest route or other.

The attributes considered were the percentage of link types in the trip (keeping roadway as a reference, and considering the different types of bicycle links), the number of traffic signals, percentage of good quality links, beautiful links, and low traffic nuisance links in the trip. As personal attributes age and sex were considered, and finally trip purpose. For all the parameters, the shortest path alternative was considered as reference. The parameters were estimated in BIOGEME (Bierlaire 2003).

Not all estimated parameters resulted significant; for the significant parameters, results of the estimation are shown in Table 19.

For what concerns the different types of link, only the presence of bike lanes, bicycle boulevards and service roads reached the level of significance. Given the negative signs of their parameters, all three of these link type proved to be avoided by cyclists (in relation to roadway links). Further, the presence of traffic signals resulted to be positively affecting the choice of longer paths, with respect to the shortest alternative.

Utility parameter	Value	T-test
$\boldsymbol{ASC_{short}}$	1	--(fixed)
$\boldsymbol{ASC_{non\_short}}$	-0.79	-4.82
$\boldsymbol{\beta_{male}}$ (dummy)	-0.68	-3.24
$\boldsymbol{\beta_{signals}}$	0.07	3.98
$\boldsymbol{\beta}_{bike\_lane}$	-1.13	-2.04
$\boldsymbol{\beta}_{bike\_boulevard}$	-4.87	-2.40
$\boldsymbol{\beta_{service}}$	-1.88	-2.09
$\boldsymbol{\beta_{good\_quality}}$	6.01	1.52
$\boldsymbol{\beta_{fair\_quality}}$	6.78	1.72
$\boldsymbol{\beta_{low\_quality}}$	7.81	1.9
$\boldsymbol{\zeta_{panel\_shortest}}$	0.79	3.09
$\boldsymbol{\zeta_{panel\_other}}$	-0.77	-2.96
Number of draws	200	
Number of estimated parameters:	25	
Number of observations:	2798	
Initial log-likelihood:	-1939.00	
Final log-likelihood:	-929.40	
Rho-square:	0.520	

Table 19: Results from the binomial mixed logit model estimate.

Referring to the personal characteristics of cyclists, age did not prove to be significant, while sex did: using sex as a dummy variable, and keeping “female” as a reference, male cyclists result to perceive longer alternatives as less attractive, if compared to the shortest available option.

As for the “quality-levels” of links, the parameters for the three levels (good, fair and low) resulted significant and positive. A positive sign could be expected for the proportion of good-quality and fair-quality links in the longer paths (alternative “other”), on the contrary a negative sign could be expected for the proportion of low-quality links. Nevertheless, this can be explained if we refer to Table 15: in the descriptive analysis, we already highlighted that low-quality links represented only the 1.6% of the chosen links, so the choice-sample for low-quality links is too small to result in something representative. More in general, it could be said that for this context, where the vast majority of links in the network database are of good quality, quality itself is not actually determining cyclists’ preferences.

4.4.3. Mixed multinomial Path-Size logit model

For the multinomial Path-Size model, the choice sets described in paragraph 5.3 were used. The Path Size formulation expressed in Eq. 14 was considered. The attributes considered were the percentage of link types in the trip, the number of traffic signals, percentage of good quality links, beautiful links, and low traffic nuisance links in the trip. As personal attributes age and sex were considered, and finally trip purpose. Specific parameters for the 5 alternatives available were estimated, again using BIOGEME (Bierlaire 2003).

Not all estimated parameters resulted significant; for the significant parameters, results of the estimation are shown in Table 20.

For what concerns the different types of link, only the presence of an exclusive bicycle facility resulted to be somewhat significant, even if the t-test value is still very low. The positive sign of the parameter’s value indicates cyclists demonstrated a certain preference for cycling longer when in presence of dedicated bicycle facilities. As seen for the binomial logit model, the presence of traffic signals results to be positively affecting the choice of longer paths, with respect to the shortest alternative. Referring to the personal characteristics of cyclists, age did not prove to be significant, while sex did: using sex as a dummy variable, and keeping “female” as a reference, male cyclists result to perceive longer path alternatives as less attractive. Nevertheless, the corresponding parameter failed to reach significance.

Utility parameter	Value	T-test
$\boldsymbol{ASC_{alt1}}$	2.74	4.13
$\boldsymbol{ASC_{alt2}}$	2.92	4.39
$\boldsymbol{ASC_{alt3}}$	2.81	4.19
$\boldsymbol{ASC_{alt4}}$	2.47	3.54
$\boldsymbol{\beta}_{Male}\ (dummy)$	-0.26	-1.07
$\boldsymbol{\beta_{exclusive\_path}}$	0.92	1.36
$\boldsymbol{\beta_{signals}}$	0.06	1.56
$\boldsymbol{\beta}_{PathSize\_Alt1}$	0.54	3.29
$\boldsymbol{\beta}_{PathSize\_Alt2}$	0.55	3.51
$\boldsymbol{\beta}_{PathSize\_Alt3}$	0.49	3.34
$\boldsymbol{\beta}_{PathSize\_Alt4}$	0.31	2.38
$\boldsymbol{\beta}_{Leisure}\ (dummy)$	1.01	3.48
$\boldsymbol{\zeta_{panel\_shortest}}$	0.943	5.75
Number of draws	200	
Number of estimated parameters:	38	
Number of observations:	2798	
Initial log-likelihood:	-1797.26	
Final log-likelihood:	-889.87	
Rho-square:	0.505	

Table 20: Results from the mixed multinomial Path-Size logit model estimate.

The positive sign of the Path-Size factor, as calculated for all of the alternatives (except for the shortest path that is the reference) indicates that if a route is less similar to the alternatives, its chances of getting chosen will be high. This positive effect of the Path-Size factor was also reported in several past studies on route choice models (e.g. Prato and Bekhor, 2006; Prato and Bekhor, 2007).

Among all possible trip purposes tested, only “leisure” resulted significantly influencing cyclists’ route choices, with a significant and positive parameter: this means that for leisure trips cyclists choose paths longer than the shortest available option, which represent a consistent and expected result.

Finally, the error component expressing the correlation between choices made by the same user (ζ_{panel}) only resulted significant for the shortest-path alternative, and positive: this means that users that repeated multiple times the same choice (same trip, i.e. same origin-destination pair) showed a tendency to choose the shortest path available. Such a result is somewhat expected and intuitive, as frequent cyclists are more likely to optimize their trips, especially when these trips are repeated such as commuting trips.

4.5. References

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

[77] Aultman-Hall, L., Hall, F., Baetz, B., 1997, Analysis of Bicycle Commuter Routes Using Geographic Information Systems: Implications for Bicycle Planning. Transportation Research Record, 1578 (1), 102-110.

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

[79] Ben-Akiva, M.E. and Bierlaire, M., 1999, Discrete Choice Methods and Their Applications to Short-Term Travel Decisions. Handbook of Transportation Science, pp. 5-34.

[80] Bierlaire, M., 2003, BIOGEME: A free package for the estimation of discrete choice models, Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

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

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

[83] Broach, J., Dill, J. and Gliebe, J., 2012, Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A, 46, 1730-1740.

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

[85] Dhakar, N., Srinivasan, S., 2014, Route choice modeling using GPS-based travel surveys. Transportation Research Record, 2413, 65-73.

[86] Geurs, K.T., Thomas, T., Bijlsma, M., Douhou, S., 2015, Automatic trip and mode detection with MoveSmarter: first results from the Dutch Mobile Mobility Panel. Transportation Research Procedia, 11, 247-262.

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

[88] Howard, C. and Burns, E.K., 2001, Cycling to Work in Phoenix: Route Choice, Travel Behavior, and Commuter Characteristics. Transportation Research Record, 1773, 39-46.

[89] Hunt, J.D. and Abraham, J.E., 2007, Influences on bicycle use. Transportation, 34, 453-470.

[90] Krizek, K.J., 2006, Two Approaches to Valuing Some of Bicycle Facilities' Presumed Benefits. Journal of the American Planning Association, 72(3), 309-320.

[91] Landis, B.W., V.R. Vattikuti, and M.T. Brannick, 1997, Real-Time Human Perceptions: Toward a Bicycle Level of Service, Transportation Research Record, 1578, 119-126.

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

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

[94] Ortúzar, J.D., Willumsen, L.G., 2011, Modelling Transport, 4th Edition, Wiley Editors.

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

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

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

[98] Schüssler, N. and Axhausen, K.W., 2009, Processing GPS raw data without additional information. Transportation Research Record, 2105, 28-36.

[99] Sener, I.N., Eluru, N. and Bhat, C.R., 2009, An analysis of bicycle route choice preferences in Texas, US. Transportation, 36, 511-539.

[100] Stinson, M.A. and Bhat, C.R., 2003, Commuter Bicyclist Route Choice: Analysis Using a Stated Preference Survey. Transportation Research Record, 1828, 107-115.

[101] Thomas, T. K.T. Geurs, J. Koolwaaij, M. Bijlsma, 2015, Automatic trip and mode detection with MoveSmarter: Results from the Dutch Mobile Mobility Panel. Paper for NECTAR 2015 conference. Ann Arbor, Michigan, USA, June 14-16.

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

[103] Train, K.E, 2003, Discrete Choice Methods with Simulation, Cambridge University Press.

[104] Winters, M., Teschke, K., Grant, M., Setton, E.M., Brauer, M., 2010, How far out of the way will we travel? Built environment influences on route selection for bicycle and car travel. Transportation Research Record, 2190, 1-10.

Concluding remarks

The satisfaction cyclists experience in their daily trips on a city's road network lies in a combination of factors that researchers tried to investigate and quantify through a wide variety of data and assessment methodologies. Undoubtedly, understanding cyclists' behavior in terms of route choice is seen as a way to build a solid framework of knowledge to refer to, when investing public resources for developing – or enhancing – a bicycle network for a city. If planners and decision-makers knows what makes a bicycle facility, or a road, more likely to attract cyclists, they own an instrument to better integrate cycling in a city as a convincing mobility option.

As research has been demonstrating in the last decades, cyclists consider a variety of factors when evaluating the quality of their cycling trips in an urban context. Cyclists' travel decisions are often motivated by perceived travel time – and speed – as well as by perceived safety. Both travel time and safety are influenced by the geometric design of the facilities (turning radii, slope, lane width) and functional features (proximity to cars, speed of adjacent cars, proximity of car parking stalls, presence of dedicated facility). Moreover, other factors intervene, less easy to quantify and measure, as land-use, pleasantness of the environment, or perceived quality of a bicycle context. This latter, for instance, can be determined by pavement conditions, illumination, interaction with motorized and non-motorized traffic components.

Within this research field, many are the typologies of data that analysts can rely to. The first studies tried to develop synthetic indicators of "bikeability", or stress level, or level of service, based on experimental measurements or stated preference surveys. The first are mainly based on measuring the three fundamental traffic quantities – speed, volumes and density. The idea is to determine threshold of such quantities that define a scale of performance for a given road or bicycle facility.

Stated preference surveys has been largely used for understanding cyclists' preferences, as they constitute a relatively cheap data collection method, and they provide a controlled experimental environment.

Another consistent part of the literature on the subject has relied on revealed preference surveys, where analyst asked participants to describe their travel choices based on actual trips they have recently completed. In order to use this more reliable source of data, though, more information on the network is needed, as well as specific methodologies for data analysis.

Recently, research on cyclists' preferences has grown thanks to the spread of the Global Positioning Systems (GPS) recording devices, incorporated in smartphones. On one side, the level of accuracy of GPS devices has increased, thanks to the investments on the involved technologies. On the other side, people habits and attitude towards the use of smartphone changed very rapidly, to the point that nowadays it is absolutely common for people to possess a smartphone, keep self-localization devices activated, declare their position in order to make it available for a burgeoning set of applications and services, register, monitor and share with people their position, trips and activities. Data availability for investigating cyclists travel choices and habits has seen an unprecedented enhancement.

Furthermore, in parallel with these changes in technology and society, research followed: several of the methodological developments relevant to modelling route choice are relatively recent, and have been applied to GPS-based travel surveys.

This work aimed at understanding which are the most critical aspects of cycling in an urban context, and individuating what makes a path more attractive of another for cyclists. To do so, we started to explore some experimental data from Bologna, an Italian medium-sized city that has one of the highest bicycle split rate, considering the Italian context. The first measurements of usage rates and speed patterns on cycleway and road segments of the network of Bologna indicated that cyclists often cycle on the road, mixing with motorized traffic, even if a bicycle facility next to the roadway is provided. This in mainly due to the frequent interruptions that a cyclist encounter along a dedicated bicycle path, and to the presence of disturbances. One of the main factors influencing cyclists' choices are indeed his perception of speed and safety.

The research then moved towards the use of GPS smartphone-based datasets. Such data provides a broader point of view on cyclists' behaviour, as well as more information on the cycling patterns of a much wider sample of cyclists on a much wider portion of the network. The benefits of the availability of such rich travel datasets comes, of course, with the burden of a much more complicated process of data elaboration and analysis. The main processing methodologies for modelling route choice from GPS traces recorded from smartphone have been described, and can generally fall into three main categories: (1) filtering and map-matching procedures, (2) choice set generation and (3) route-choice modelling.

The map-matching process consists of all those operations that transform a stream of GPS points to a road-network database, in order to identify the traversed links in the chosen route. In general, the challenge for new and advanced map-matching procedures is to increase the accuracy while maintaining computational efficiency.

In this work, a novel map-matching methodology have been proposed and applied to a real case-study, i.e. the GPS traces recorded by cyclists in the city of Bologna. Approximately 3000 (55%) traces could be reliably matched to the road network. The creation of buffers around network links and the determination of the probability of GPS points inside the buffer, in combination with length – and type – specific edge attributes, has shown promising results, especially in dense street grids and for different road types. In particular, when a cycleway link is adjacent to a road link, the algorithm allowed to match trips giving preference to the cycleway link. This is particularly useful for the dense network in urban areas, even if we saw from the first part of the study that such an assumption can lead to overestimate usage rates for bicycle facilities.

The trips obtained from the map-matching process have been analysed in terms of length distribution, time distribution, and travel duration.

Further, the composition of chosen routes in terms of link type has been analysed: results confirm the observations derived from the first part of the study, i.e. that cyclists chose very unfrequently to cycle on dedicated bicycle facilities, while they tended to choose road links. This is undoubtedly related to the scarcity of bicycle facilities (only the 6% of the total links in the Open Street Map network), but also to a preference for directness and time savings over safety.

Subsequently, chosen routes has been compared to the shortest paths calculated between the same origin-destination pairs. This comparison showed that only the 26% of trips were made choosing the shortest option available, while the majority of trips contained significant detours (i.e. additional length cycled). The data available did not allow to relate this result with trip purpose, as it was unknown, but the distribution of detour along the days of the week and the hours of day suggest that the detour is grater for weekends and non-peak hours, thus it seems to characterize more non-systematic trips.

In the final part of this work, a second case study has been considered: the GPS traces recorded in the framework of the Mobile Mobility Panel, throughout the whole territory of the Netherlands. Approximately 280 bicycle users were asked to register all their trips through a smartphone application specifically elaborated for this data collection, called MoveSmarter. Furthermore, the recorded trips have been corrected by the users themselves, who checked and corrected their travel information daily, on a web web-based prompted recall survey. Finally, the bike trips had already been matched to the Open Street Map network, but also a second network database was available, i.e. the Fietsersbond network database. This is a specific bicycle network built and updated by a Dutch cyclists union, which reaches a surprising level of detail both in terms of link number and in terms of information attached to its links.

From a first descriptive analysis of the Dutch GPS traces, one can clearly spot some main differences with the Italian case study, given by the different contexts considered: trip lengths and trip distribution over time shows a population sample that is much more used to cycle, for longer distances, and more frequently. Furthermore, when considering the composition of chosen routes in terms of link type, the usage of cycleway links is much more frequent. Of course, this was an expected result, as the percentage of dedicated facilities in the Dutch network is much higher than in the Italian one.

The level of detail of the datasets involved allowed to test for bicycle route choice models: a binomial logit model and a multinomial logit model have been estimated. Path-size logit model formulation has proved to be the best model formulation for bicycle route choice, as it allows to consider the overlapping of alternatives that characterizes bike trips. Furthermore, we referred to the mixed logit model formulation to introduce an error component that was capable of capturing the influence of trip repetitions. Indeed, many GPS traces from the database consisted of the same trip (origin-destination pair) repeated by the same user throughout the monitoring period. The estimation results have shown that for repeated trips, the shortest route option tend to be chosen more, suggesting that frequent cyclists, on systematic trips, tend to optimize their trip and to prefer shortest routes. Nevertheless, in general, the percentage of choice of the shortest route is completely comparable with the Italian one, i.e. around 25%.

Unexpectedly, the available information regarding quality, beauty and traffic nuisance for the network links, did not result in significant parameters in the modelling. Nevertheless, from the descriptive analysis of the GPS trips, we saw that most trips are made on links that present a good-to-very-good level of quality and beauty, and a low-to-very-low level of traffic nuisance. This is probably explained keeping into consideration that the general level of quality of bicycle links in the Dutch context, especially in urban environments, is rather high.

Another important aspect to highlight is the choice set generation methodology. Given the high number of trips, frequently repeated, in the Dutch trip datasets, we tried to identify the alternatives that compose the choice set based on the observed routes. Choice sets obtained in this way are certainly smaller than those obtained by topological or probabilistic algorithms we described in the literature review section (Chapter 3), but they do represent a more realistic set. In order to build the alternatives, these trips were then grouped by length, and 4 groups of registered trips – for each user and each OD pair –

where created, which represented the four alternatives available to cyclists. The attributes of each alternative has been obtained as an average of all the recorded trips belonging to the group. For each user and OD pair, one additional alternative has been introduced, represented by the shortest path between the OD pair. Thus, each choice set has a minimum size of 1 (i.e. the shortest path, calculated for all OD pair) and a maximum size of 5 alternatives (4 generated from the observed routes plus the shortest path).
As future development, it would be interesting to consider other criteria, other than the sole length, to group the observed routes and create the alternative sets; for example, groups could be made based on the proportion of good-quality links in the path, or based on the number of traffic signals. A cluster analysis on multiple parameters could be performed, in order to define the alternatives based on actual differences in observed routes. Alternatively, the choice set could be enriched by adding to the shortest route, the safest route, or the one made entirely by separated cycleway links, or the most scenic, etc., as calculated by any GIS routing tool.

Finally, this work shed light on an important aspect regarding the use of GPS traces coming from smartphone: such data truly represent a rich source of detailed information on cyclists' behavior and preferences, but in order to be used it needs just as much detail for what concerns the network database. As we saw in the map-matching application to the Italian case study, the accuracy of the process is greatly affected by the completeness of the network model, and this is even truer for bicycle trips, as cyclists frequently move on "unofficial" links, such as pavements or park paths. Moreover, in order to estimate accurate models, the analysts must include in the explanatory variable set a wide number of features, sometimes not only describing the links themselves, but the bike environment. If, as we saw, many are the factors influencing bicycle route choice, then greater in the effort to capture these factors and properly describe them through georeferenced information. In this sense, the diffusion of crowd-sourced information, and volunteered geographic information, can represent a positive future development for the research field.