

Carlos Eduardo Tinoco de Carvalho

Assessing the Reliability of OpenStreetMap for Cycling Analytics Through Map-Matching Anomalies

UMinho | 2026



Universidade do Minho
Escola de Engenharia

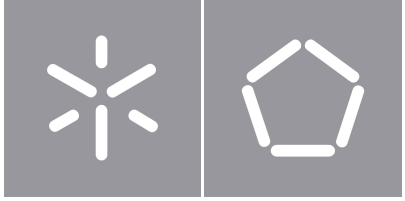
Carlos Eduardo Tinoco de Carvalho

Assessing the Reliability of OpenStreetMap for Cycling Analytics Through Map-Matching Anomalies

May 2026

Fundação para a Ciência e a Tecnologia

MIT Portugal



Universidade do Minho
Escola de Engenharia

Carlos Eduardo Tinoco de Carvalho

Assessing the Reliability of OpenStreetMap for Cycling Analytics Through Map-Matching Anomalies

Doctoral Thesis
Doctoral Programme in Information Systems and Technology

Work carried out under the supervision of
Professor Doctor Rui José
Professor Doctor Paulo Ribeiro

May 2026

COPYRIGHT AND TERMS OF USE OF THIS WORK BY A THIRD PARTY

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorisation conditions not provided for in the indicated licensing should contact the author through the RepositóriUM of the University of Minho.

License granted to users of this work



CC BY-NC-SA

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International

https://creativecommons.org/licenses/by-nc-sa/4.0/

Acknowledgements

First and foremost, I want to thank my supervisors, Professor Rui José and Professor Paulo Ribeiro, for their invaluable guidance throughout this work. Without their wisdom, I would not be able to complete this journey, a journey that began well before my enrollment in the doctoral programme. Thank you for motivating me to pursue academic research and to embark on the doctoral programme. Thank you for the countless lessons I learned during experiments, writing, and academic events. Above all, thank you for continually challenging me to grow, not only as a researcher but also as a person.

I also want to thank Professor Maria João Nicolau, who, during my master's dissertation, believed in me and invited me to serve as a teaching assistant. That opportunity sparked my interest in academia and in contributing to the education of future generations of students and engineers.

I extend my gratitude to MIT Portugal and FCT - Fundação para a Ciência e Tecnologia, I.P., for the PhD grant [1] that allowed me to develop this work.

To my friends, in particular, Hugo Peixoto, Ricardo Peixoto, and Ivo Silva, thank you for the moments of relaxation, the adventures, and the words of encouragement. These moments made this journey easier.

A special thanks to my Mother and Father for the education, values and unconditional love. Thank you for teaching me the values of perseverance and hard work, essential for such a demanding path, for being my safety net, for your sacrifices, and for providing me with everything I ever needed. I extend these acknowledgements to my extended family, especially my grandparents, who contributed immensely to me becoming who I am today.

Lastly, a special thanks to the love of my life and future wife, Mia, who has always been with me, in the good and the hard times, who always believed in me, and gave me the strength to never give up. Thank you for being who and how you are, for your way of seeing the world, for your boldness, kindness, and for the joy you bring to this world. I also extend my gratitude to Mia's family. Their way of living, rooted in love and unity, their family reunions, filled with warmth and good spirits... It's truly something special.

I am indeed a fortunate person.

[1] https://doi.org/10.54499/PRT/BD/152831/2021

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, May 2026



Carlos Eduardo Tinoco de Carvalho

DECLARATION ON THE USE OF ARTIFICIAL INTELLIGENCE (AI) TOOLS

Thereby declare having acted this academic work with integrity and confirm that I have not resorted to any form of undue use of AI tools, nor to the falsification or omission of information, sources, data, results or processes in any of the stages leading to its elaboration.

I further declare that any use of AI tools was done in a transparent and responsible manner, being duly noted in the work, and that I critically verified the accuracy of the content produced with their support, assuming full responsibility for what I submit.

I further declare that I have fully acknowledged the Code of Ethics and Conduct of the University of Minho, as well as the applicable standards of academic integrity and data protection.

University of Minho, Braga, May 2026



Carlos Eduardo Tinoco de Carvalho

Resumo

Avaliação da fiabilidade do OpenStreetMap para analíticas de mobilidade ciclável através de anomalias de map-matching

Uma estratégia de plataforma para analítica de mobilidade ciclável requer um referencial para agregar diferentes tipos de dados em pontos comuns. O OpenStreetMap (OSM) posiciona-se como uma tecnologia promissora. Contudo, o conceito de rede ciclável é ambíguo, com os ciclistas a andar por caminhos não convencionais. Além disso, a comunidade OSM aplica diferentes práticas de mapeamento, influenciadas por diversos fatores. Essas ambiguidades reduzem a capacidade do OSM de representar caminhos feitos por ciclistas. Por este motivo, é importante perceber se o OSM pode ser usado como referencial para agregar dados de mobilidade ciclável e como facilitador de novas soluções.

Esta tese mostra que o OSM tem várias limitações para representar diferentes tipos de caminhos, especialmente os não oficiais e os menos visíveis. Mostrou também que sem uma maior adoção, o OSM talvez não atinja o nível necessário para permitir novas soluções para a mobilidade ciclável.

Uma contribuição principal é o conceito de anomalias de map-matching, uma discrepância sistemática entre o OSM e as rotas de ciclistas numa área específica. A tese propõe um método não supervisionado para identificar anomalias numa cidade agrupando falhas do map-matching.

Entrevistas com intervenientes principais revelam que as anomalias devem-se a limitações técnicas do OSM para representar a realidade e acomodar diferentes perfis de ciclistas e às práticas ambíguas de mapeamento. As anomalias podem ser vistas como infrações ou como ferramenta para melhorar a rede ciclável, revelando as linhas de desejo dos ciclistas. As anomalias podem conscientizar os interessados sobre problemas existentes e direcionar discussões. Resolver anomalias existentes aumenta o potencial do OSM como facilitador de novas soluções, o que, por sua vez, incentiva a adoção do OSM e cria um círculo virtuoso de melhoria. A solução ótima para cada caso vai necessitar de colaboração entre os vários intervenientes. No entanto, essa colaboração ainda está longe de ser realidade.

Palavras-chave: mobilidade ciclável, rede ciclável, OpenStreetMap, modelos de análise de dados

Abstract

Assessing the Reliability of OpenStreetMap for Cycling Analytics Through Map-Matching Anomalies

A major requirement for a platform strategy for cycling analytics is the usage of a reference system that can aggregate multiple sources of data around geographic concepts. For this purpose, OpenStreetMap (OSM) stands as a promising candidate. However, the concept of a cycle network is fuzzy, with cyclists often riding on unconventional paths. Moreover, the OSM community employs different mapping practices, influenced by various factors. These ambiguities pose a huge challenge to the ability of OSM to represent real cyclists' paths. Therefore, it is important to understand whether OSM can be used as a reference system for aggregating cycling mobility data and whether it can enable new cycling mobility solutions.

This work shows that OSM has several limitations in representing different types of cycle routes, especially unofficial and less visible paths. It also shows that without wider adoption, OSM may not be able to reach the level needed to facilitate new solutions for cycling mobility.

A main contribution is the concept of map-matching anomaly, a systematic discrepancy between OSM and the pattern of GPS traces crossing a specific area. This thesis also proposed an unsupervised method to identify map-matching anomalies in a city by clustering map-matching gaps.

Interviews with key stakeholders revealed that anomalies are mostly caused by OSM technical limitations in representing reality and accommodating different routing profiles and by ambiguous mapping practices. Anomalies can either be seen as traffic infractions or as a tool for continuous improvement of the cycling network by revealing cyclists' desire lines. Anomalies can raise awareness among stakeholders about existing problems and frame discussions. Addressing existing anomalies can improve the potential of OSM as an enabler of new models and applications, which in turn, encourages OSM adoption and creates a virtuous cycle of improvement. The optimal solution for each case will always require collaboration between stakeholders. However, this collaboration is still far from becoming a reality.

Keywords: cycling mobility; cycling road network; OpenStreetMap, mobility models

Contents

1 Introduction **1**

1.1 Research Context 1

1.1.1 Cycling Analytics: Current Development Paradigm 2

1.1.2 A Solution Based on a Digital Platform Strategy 3

1.1.3 Research Challenges 4

1.2 The Challenge of a Common Reference System 5

1.2.1 The OpenStreetMap Project 6

1.2.2 Bicycle Route Ambiguities 7

1.3 Objectives 8

1.4 Thesis Structure 10

1.5 Research Publications 10

1.5.1 Contributions per Chapter 11

2 State of the Art **13**

2.1 Bikeability 14

2.2 Bike-Friendliness Indexes 19

2.3 Cycling Analytics Platform 26

2.4 Cyclists' Route Preferences 35

2.5 OSM Road Network Completeness 38

2.6 OSM Representation Politics 41

2.7 Map-Matching 44

2.8 Summary 46

3 Cycling Analytics Models Decomposition **48**

3.1 Motivation 48

3.2 Methodology 48

3.2.1 Model Selection . . . 49
3.3 Results . . . 49
3.4 Key Properties for a Platform Strategy for Cycling Analytics . . . 52
3.5 Summary . . . 56

4 The Gaps Between Cycling Traces and OSM Road Network 58
4.1 Motivation . . . 58
4.2 Methodology . . . 59
4.2.1 Route Processing - Map-Matching . . . 60
4.2.2 Identification of Map-Matching Gaps . . . 61
4.3 Results . . . 62
4.3.1 Gap Categories . . . 62
4.4 The Limitations of the OSM Road Network Model . . . 67
4.5 Summary . . . 68

5 A Supervised Method to Automatically Detect and Classify Map-Matching Gaps in Cycling GPS Traces 69
5.1 Motivation . . . 69
5.2 Methodology . . . 69
5.2.1 Mobility Data Acquisition and Preparation . . . 70
5.2.2 Road Network Data Acquisition . . . 70
5.2.3 Map-matching . . . 70
5.2.4 Ground-Truth Data Acquisition . . . 71
5.2.5 Generation of Similarity Measures . . . 71
5.2.6 Classification models . . . 73
5.3 Results . . . 74
5.3.1 Analysis of the Map-Matching Process . . . 74
5.3.2 Binary Classification for Map-Matching Gap Detection . . . 77
5.3.3 Multi-class Classification for Map-Matching Gap Classification . . . 79
5.3.4 Error Indicator Method Assessment . . . 81
5.4 Summary . . . 84

6 An Unsupervised Method to Identify Anomalies in Road Network Models for Urban Cycling 85

6.1 Motivation . . . 85
6.2 Methodology . . . 87
6.3 Process and Results . . . 89
6.3.1 Map-Matching . . . 89
6.3.2 Gap Extraction . . . 90
6.3.3 Anomaly Identification . . . 92
6.3.4 Anomaly Analysis . . . 97
6.4 Overcoming Anomalies to Improve Cycling . . . 99
6.4.1 Workflow Analysis . . . 100
6.5 Summary . . . 101

7 Stakeholders' Perspectives on the Gaps Between Cycling Traces and Their Representation on OSM 103

7.1 Motivation . . . 103
7.2 Research Methods . . . 103
7.2.1 Gap Cases . . . 104
7.2.2 Focus Groups and Interviews . . . 106
7.2.3 Qualitative Analysis . . . 109
7.3 Limitations of OSM Technology and Mapping Practices . . . 110
7.3.1 The Model is Only a Simplification of the World . . . 110
7.3.2 Reduced Awareness of Cycling Traces . . . 111
7.3.3 Representation Granularity . . . 111
7.3.4 Tagging Consensus . . . 111
7.3.5 Route Selection is Circumstantial . . . 112
7.3.6 Mapping vs Routing . . . 112
7.4 The Politics of Route Representation . . . 113
7.4.1 Public Authorities and OSM Awareness . . . 113
7.4.2 Circumstantial Interpretations of Routes . . . 114
7.4.3 Private vs Public . . . 115
7.4.4 Anomalies as Infractions . . . 115
7.4.5 Anomalies as a Tool for Continuous Improvement of the Cycling Network . . . 116
7.4.6 Accountability vs Informed Decision . . . 117
7.5 Implications in the OSM's Role as an Enabling Technology . . . 119

7.5.1 Implications of Technical Uncertainty 119
7.5.2 Implications of Conflicting Perspectives Between Stakeholders 120
7.6 Summary 121

8 Conclusions **123**

8.1 Contributions 123
8.2 Limitations 130
8.3 Future Work 130

List of Figures

1 Cyclers Navigate Route Planner (Cyclers, 2019) 28
2 Pedalheat® - Urban Cycling Trips Heatmap (KAPPO, 2016) 30
3 Ride Report Analyze Tab (Murray, 2021) 31

4 Map view of the original bike trace recorded in Braga. 60
5 Map-matching of a GPS trace performed on a sidewalk along a one-way road. 62
6 Missing Segment (ID3) 63
7 Missing Intersection (ID15) 64
8 One Way (ID13) 65
9 Competing Road (ID1) 65
10 Complex Crossing (ID10) 66
11 Semi-public Road (ID4) 66

12 Examples of common map-matching gaps. 76
13 Correlation between features 78
14 Confusion matrix for the RF model with no feature engineering. 81

15 Examples of three map-matching anomalies. 86
16 Heatmap with the spatial distribution of the GPS traces in the Bicification dataset. . . . 87
17 A four-step methodology for anomaly identification and analysis. 88
18 Example of two different map-matching anomalies. 91
19 Anomaly Identification results. 94
20 The distribution of the number of gaps per anomaly. 95
21 Anomaly Segmentation Results: The number of gap trace patterns per anomaly. 97
22 Example of three patterns emerging from segmenting an anomaly. 97
23 Different levels of analysis: a) global, b) anomaly, c) pattern, and d) gap. 100

24 Examples of six map-matching gaps. 105
25 Dimensions emerging from the qualitative analysis. 110

List of Tables

1 Bike-friendliness indexes assessed in this review 25
2 Analytics Platforms Reviewed 34
3 Mobility models analysed in this study. 49
4 Possible horizontal features identified in the analysis and their dependencies. 50
5 Common situations for the occurrences of map-matching gaps. 59
6 Identified gap categories 63
7 General stats: Map-matching gaps 75
8 Category occurrences per city. 76
9 Binary Classification Results - Short Version. 79
10 Train Dataset for the iteration: "-SMT-U". 80
11 Multi-class classification Results - Short Version. 80
12 Error Indicator Performance. 82
13 Error Indicator recall for the five most common categories. 83
14 Results of applying the data-cleaning procedure. 89
15 Results of applying the map-matching procedure. 90
16 Results of applying the gap extraction procedure. 92
17 Results of applying DBSCAN to the eligible map-matching gaps. 93
18 Evaluation of the anomalies. 98
19 Binary Classification Results 147
20 Multi-class classification Results 150

Acronyms

AADT Average Annual Daily Traffic.

AD Average Distance.

ADE Average Distance Error.

ANOVA Analysis of Variance.

API Application Programming Interface.

BFI Bike Friendly Index.

BiRNN Bidirectional Recurrent Neural Network.

BNA Bicycle Network Analysis.

DEM Digital Elevation Model.

DT Digital Transformation.

DTW Dynamic Time Warping.

EI Error Indicator.

FD Fréchet Distance.

FN False Negatives.

FP False Positives.

GIS Geographic Information System.

GNSS Global Navigation Satellite System.

GPS Global Positioning System.

ITS Intelligent Transportation System.

KNN K-Nearest Neighbors.

KPI Key Performance Indicator.

LAUs Local Administrative Units.

LI Length Index.

LR Logistic Regression.

LSOA Lower Layer Super Output Area.

LTS Level of Traffic Stress.

MaaS Mobility as a Service.

MDS Mobility Data Specification.

ML Map-matched Length .

MSOA Medium Layer Super Output Area.

NB Naive Bayes.

OD Origin-Destination.

OEG Organised Editing Guidelines.

OSM OpenStreetMap.

OWP Optimal Warping Path.

PCA Principal Component Analysis.

PCT Propensity to Cycle Tool.

PMM Partial Map-Matching.

POIs Points of Interest.

REST Representational State Transfer.

RF Random Forest.

SMOTE Synthetic Minority Over-sampling Technique.

STD Stardardised.

SVC Support Vector Classifier.

TL Trace Length.

TN True Negatives.

TP True Positives.

VGI Volunteered Geographic Information.

XGBoost Extreme Gradient Boosting.

Chapter 1

Introduction

1.1 Research Context

Municipalities increasingly recognise the role of bicycles as a key element in urban mobility and as a tool to cope with net-zero emission mandates (Eguiluz et al., 2022). As a result, municipalities around the globe are implementing policies and engaging in strategic investments to redesign public spaces and introduce innovative mobility schemes to promote a shift from conventional mobility modes to cycling (Bulc, 2016).

Digital transformation plays a crucial role in this mobility transition. The concept of Smart Cycling has emerged as the shared, real-time, and collaborative application of data, communications, and services to optimise individual and collective movement across the urban environments (Stratta, 2018). It promises to grant cyclists similar possibilities and privileges that drivers already have (Nikolaeva et al., 2019), and fosters new alliances by bringing together traditional cycling stakeholders, communities, organisations, and industries. In this context, digital transformation is allowing the emergence of new business models and innovative mobility services, such as bicycle sharing services and smartphone applications that provide real-time analytics or updates on traffic conditions (Tsakalidis et al., 2020). As in many other domains, digital technologies for bicycles and cyclists are also gaining ground (Boot et al., 2024): urban cycling is increasingly shaped by sensors, actuators, mobile applications, data portals, and digital platforms, generating a vast amount of data that can be used to help cyclists to plan their routes and urban planners to understand travel behaviour and identify areas with higher potential for improvement (Stamatiadis et al., 2017).

More specifically, Smart Cycling has spurred a major interest in new data models created specifically to represent particular dimensions of the cycling reality of a city or region. Currently, there is a wide diversity of compound indices to assess bikeability (Gholamialam and Matisziw, 2019; Grigore et al., 2019; de Matos et al., 2021) and bike-friendliness (Copenhagenize, 2019; Score, 2018; Vale, 2020); custom tools for estimating the economic value of cycling (Ferreira et al., 2020) or the potential for cycling (PCT.BIKE,

2017; Silva et al., 2021); commercial services based on data from mobile applications (Lee and Sener, 2021); and a broad range of smaller and specialised models, serving specific perspectives of analysis, such as slope (lab, 2020), pollution (Tran et al., 2020), route choice (Alattar et al., 2021), comfort (Berger and Dörrzapf, 2018) or accessibility (Vale, 2020). These models are often published as open-source projects, and, overall, they represent a wide spectrum of analyses of the many dimensions of urban cycling.

Similarly, Smart Cycling has driven the development of several digital platforms specifically designed for cycling. One of the most popular examples is Strava[1], both a digital platform and a social network. Strava has millions of active users who track, analyse, and share their activities using smartphones, fitness watches, or other devices. Other notable platforms include Bike Citizens Analytics[2] and Ride Report[3]. Due to their widespread adoption, these platforms have access to large amounts of data, which they process, aggregate, and monetise[4]. For example, access to Strava data can cost hundreds of thousands of dollars, depending on the geographic area and time of interest (Lee and Sener, 2021).

1.1.1 Cycling Analytics: Current Development Paradigm

Despite continuous efforts, the full potential of cycling analytics as a transformational force for urban cycling remains largely untapped. Many of its building blocks may already be there, but they are not connected in ways that allow them to deliver the type of shared, real-time collaborative application of data envisioned by smart cycling. There are many relevant models, but in the end, only a small community of users with strong technical knowledge on how to prepare, process and explore this type of data is able to work with them. Common citizens, and even the people responsible for managing cycling mobility, may not have the expertise, the time or the practices to get value from those models. This problem is particularly relevant for municipalities and urban planners who lack accurate, comprehensive and actionable data tools to comprehend the cycling reality.

The root cause of this problem seems to be the prevailing development paradigm, mostly based on self-contained and custom processes that do not offer the necessary context for distributed and open innovation. Currently, developing a new model typically involves a single entity gaining access to a dataset, applying a set of algorithms developed specifically for the study, generating new data structures, and creating tailored visualisations to present the results. This self-contained process is very flexible because there is no need to consider generalisation or integration issues. However, it leads to monolithic models that

[1] https://www.strava.com/

[2] https://cyclingdata.net/

[3] https://www.ridereport.com/

[4] https://www.strava.com/legal/privacy

become closed packages of data, expertise, methodologies, and conclusions that are hard to reproduce, generalise, recombine, or improve beyond their original context.

Similarly, current commercial cycling analytics platforms often adopt closed approaches, typically offering a limited set of features developed solely by their teams. Their users are often limited to using the services offered by platform creators, consulting a small and fixed set of information, and, at best, contributing to the platform by providing their personal data generated while cycling. This approach prevents interested parties, for example, from repurposing the platforms' processing models to analyse their datasets or exploring the platform data to develop and test new analytic models.

Due to the limitations of the current paradigm, no system has become widely accepted to support the collaborative development of analytics models for cycling mobility in urban environments. In reality, when someone wants to explore a dataset or develop a new model, it often feels like starting from scratch, forced to rebuild the entire process. In the absence of a comprehensive approach to assessing reality, tasks such as benchmarking cities' support for cycling, monitoring progress over time, or measuring the impact of policy initiatives become highly challenging, if not unfeasible Decision makers, who aim to understand travel behaviour and citizen needs, remain unaware of which data they should collect to achieve their objectives. Even when they have access to rich datasets, they miss the opportunity to analyse them effectively and improve citizens' quality of life. Consequently, investments are frequently misaligned with actual needs, exacerbating the feeling of dissatisfaction among the population and slowing the transition from conventional transportation modes to cycling.

1.1.2 A Solution Based on a Digital Platform Strategy

The concept of digital platform has become widely explored to address these types of challenges, and it could offer a viable and scalable approach to deliver the value of rich mobility analytics to urban cycling ecosystems. Digital platforms are a well-known strategy for optimising innovation efforts within an ecosystem through combinatorial and distributed innovation across organisational boundaries (Yoo et al., 2012), creating new sources of innovation based on the principles of convergence and generativity (Ciriello et al., 2018).

Convergence means that digital technologies can combine previously separate components or services. The new convergence spaces enabled by digital technology provide numerous innovation opportunities for products and services, potentially transforming a hugely fragmented space of data sources, mobility models, and Key Performance Indicators (**KPI**) into a shared and actionable set of data references for cycling ecosystems. Generativity means that digital technologies can produce unprompted change driven

by large, varied, and uncoordinated audiences (Zittrain, 2005). They are never complete, as they are inherently dynamic and malleable, continuously evolving and being extended by the actions of users or the contributions of a large community of product and service creators. Digital platforms offer communities the opportunity to innovate by customising parts of the digital platform to meet their specific needs.

A platform strategy may also offer the type of continuous service needed for real impact, such as readiness to be used and availability for multiple territories. Currently, most models are conceived either as conceptual tools or as a part of specific planning studies, lacking the capability to function as active services. The few systems that support the systematic publication of cycling metrics, such as the Propensity to Cycle Tool (Lovelace et al., 2017) or the Bike Friendly Index (Figueiredo and Vale, 2018), are focused only on a very specific range of metrics. A properly designed digital platform should be able to provide a systematic, collaborative, and incremental model development paradigm, supporting the convergence between multiple models and promoting their usage as shared references for cycling ecosystems.

1.1.3 Research Challenges

A digital platform constitutes a promising approach to fully unlock the potential of the analytical models for cycling in urban environments. However, in practice, the development of such a platform entails a range of challenges across multiple domains, including:

(1) Data heterogeneity: One of the primary challenges in cycling mobility analysis is data heterogeneity. In the case of cycling mobility, data arise from many fields and applications. The most common example is data about cycling demand, most commonly expressed as **GPS** traces or bicycle counts (Berger and Dörrzapf, 2018; Long and Zhao, 2020; Citizens, 2017). Because these basic data types may not provide a sufficiently rich understanding of reality, researchers have also produced other types of data to provide a clear understanding of reality, such as cyclist feedback, complementary sensors (e.g. accelerometers, biometric, video or images), or external data sources (e.g. accidents, weather, road network, altimetry, motorized volumes) (Resch et al., 2020; Tran et al., 2020; Kamel et al., 2020; Saghapour et al., 2017; Schmid-Querg et al., 2021). However, due to the lack of widely used standards for publishing these types of data, developers adopt those which best suit their needs and skills. This results in a myriad of data types, structures (Schabenberger and Gotway, 2005) and publication venues. In reality, there is no lack of data. The challenge is to obtain it and understand it. There are many cycling datasets available from various repositories across many cities and countries, but they tend to lack proper details about their properties; for example, they do not include information about their schemas, data collection processes, geographical scopes, and data quality. The platform should not impose rigid data structures,

as that could limit innovation and the evolution of better formats. Instead, the real challenge is to design a platform capable of integrating data from diverse sources and seamlessly converting between different formats, ensuring compatibility across various analytical models.

(2) Fragmented models: A second fundamental challenge lies in the development and publication of analytics models. Currently, there is no common workflow for sharing such models. As with datasets, researchers adopt the approach that best aligns with their skills and needs. While this approach speeds up the development process, it creates limitations when sharing with others. Even when the source code is available, it often lacks sufficient documentation, making it difficult to understand and adapt to different scenarios. In addition, mobility models inherently contain embedded knowledge acquired during their development; however, this knowledge is often omitted when the model is published, limiting its usability and reproducibility.

(3) Balance between generativity and control: The platform must achieve the optimal balance between flexibility and control. Excessive flexibility can lead to further fragmentation rather than convergence, while excessive control may constrain creativity and limit the development of new tools. The ideal solution must be flexible enough to foster innovation and generativity while maintaining sufficient control to promote the convergence of data representation and model development.

(4) Common reference systems: Mobility analysis relies on data acquired from diverse sources, each providing information in different formats and spatial reference systems. These heterogeneous data must be aggregated into a common reference system to ensure effective analysis. During aggregation, the geospatial data is reduced to fewer data units, including points, lines, polygons, or pixels (Blan and Butler, 1999). In mobility analysis, points may represent intersections or turning points, lines often correspond to roads, and polygons may include Local Administrative Units (**LAUs**), postal codes, or custom fishnets. Thus, the ideal reference system should be open, freely accessible, widely adopted, and globally comprehensive. These properties will enable models to be generalised across different contexts with minimal adjustments and effort. This would support data-driven improvements to cycling infrastructure and policies, and would improve benchmarking and comparisons across cities.

1.2 The Challenge of a Common Reference System

The set of challenges outlined exceeds the scope of a single doctoral thesis. For this reason, in this work, we will focus on the challenges related to the need for a common reference system. Such a system should serve as a reference point for many models and algorithms, allowing researchers to aggregate information

about cycling activities.

As a research perspective, we focus on open platforms to have more flexibility and control and avoid usage limits or fees. The system should also be widely accepted and used by different models, services, and applications across the globe, meaning that the system is representative of the problem that we are targeting, and it already has an impact on cycling mobility. Based on these criteria, the OpenStreetMap (**OSM**) project emerges as a promising candidate.

1.2.1 The OpenStreetMap Project

As part of the increasing demand for tools to support cycling mobility, OSM has become significantly relevant as an enabling technology for many cycling applications and services. OSM is a open and global geographic database (Barrington-Leigh and Millard-Ball, 2017) curated by a large and diverse community of volunteers, which includes individual mappers, informal OSM communities, for-profit organisations, and paid editors (Anderson et al., 2019). They rely on ground surveys, imagery, local knowledge, and government data to maintain OSM with up-to-date information, among others, about the road network, buildings, and administrative limits. These data are used in many applications, such as navigation services, urban planning, disaster management, and public health (Herfort et al., 2023). Its characteristics make OSM particularly valuable for reproducibility and accessibility in research and practical applications (Reggiani et al., 2022).

OSM uses three elements to represent geographical features, namely nodes (points in space), ways (linear features and boundaries) and relations (how elements work together)[5]. Each element has one or more tags describing its properties. A tag consists of a key-value pair, in which the values of each field are created freely and later conventionalised by the community[6]. In the context of mobility, OSM enables mappers to create a comprehensive representation of the existing road network infrastructure. Mapping a road typically begins with someone defining its geometry and typology (secondary, residential, pedestrian, ...). Over time, the community maps more details and keeps the information up-to-date. For a given road, mappers may describe the pavement conditions, movement restrictions (e.g., speed limits, traffic direction, permitted modes of transport), road names, and other pertinent information. As example, for mapping cycling infrastructure, mappers can use the "highway=cycleway" tag to express separate cycleways, and in cases where the cycling infrastructure is part of a road, mappers can include the "cycleway=*" tag to an existing way ("highway=*") (Contributors, 2025a,b).

[5] https://wiki.openstreetmap.org/wiki/Elements

[6] https://wiki.openstreetmap.org/wiki/Tags

The OSM road network is the building block of many cycling-related services and studies (Bergman and Oksanen, 2016), including navigation services that suggest the most suitable cycling routes (e.g., Mapbox Directions API[7], Ride with GPS route planner[8], OSRM[9], or OpenRouteService[10]), mobility analysis tools that help decision-makers plan cycling infrastructures (e.g., Bike Citizens Analytics[11] or Eco Counter "Eco-Display World Map"[12]) or research papers aiming to evaluate several dimensions of cycling mobility, such as bicycle network connectivity (Miah et al., 2023), bikeability (Wysling and Purves, 2022), road cycling safety (Bock and Verstockt, 2023), or route choice (Yeboah and Alvanides, 2015).

1.2.2 Bicycle Route Ambiguities

Due to its characteristics, OSM stands out as a natural candidate for the role of a universal reference system for aggregating cycling-related data. However, bicycle transportation planning requires high-quality and consistent cycling infrastructure data (Wysling and Purves, 2022). Therefore, OSM's ability to play this pivotal role in cycling information systems is strongly dependent on its ability to offer an accurate representation of cyclists' traces. In this case, an accurate representation should mean that the completeness and quality of its data (Barrington-Leigh and Millard-Ball, 2017) should guarantee a suitable match between its cycling road network model and the paths effectively used by cyclists.

While several authors support that OSM has good quality (Barrington-Leigh and Millard-Ball, 2017; Zhang and Malczewski, 2018), the reality of cycling paths is much fuzzier than the common assumption involving only dedicated bike lanes or car roads. Bicycle route choice is strongly personal, and cyclists combine different criteria when selecting their routes (Zimmermann et al., 2017). A bicycle route ends up being the result of a multi-objective optimisation process (Nunes et al., 2021) that comprises the selection of cycling tracks, but also many other types of roads (Reggiani et al., 2022). A realistic cycling journey may involve frequent switching between very heterogeneous roads with different profiles and purposes, such as footpaths, parks, and other unconventional paths that often are not represented on a road network model or at least not represented as cyclable routes (Schweizer et al., 2016). The fundamental issue is the existence of very diverse interpretations about the road network. Mekuria et al. (2012) describe two opposing views of the meaning of the cycling road: municipalities view the bike network as the set of links designated for cyclist use, whereas cyclists define it as the set of paths that do not exceed their tolerance

[7] https://docs.mapbox.com/api/navigation/directions/

[8] https://ridewithgps.com/routes/new

[9] https://project-osrm.org/

[10] https://openrouteservice.org/

[11] https://cyclingdata.net/

[12] https://eco-display-map.eco-counter.com/?siteId=100155323

for traffic stress.

Moreover, there are also ambiguities associated with OSM representations. Despite continuous efforts by the community, Hochmair et al. (2015) describe two types of errors: omission errors occur when physical elements are missing in the OSM database, while commission errors occur when features present in the dataset do not exist in the real world. Since OSM relies on crowd-sourced contributions, maintaining consistent tagging is challenging. Mappers may lack sufficient training in geography or surveying, apply diverse mapping practices, or have different interpretations of the tags, making their contributions inaccurate, incomplete, inconsistent, or vague (Jackson and Zhang, 2016; Nelson et al., 2021; Ferster et al., 2020). Despite established community conventions, OSM tagging methods vary across regions, particularly for paths and cycle tracks (Ferster et al., 2020). For example, the tag "highway = cycleway", intended to represent a dedicated cycling infrastructure, is used to model both cycle tracks and off-street paths. This approach blurs the distinction between different cycling facilities.

These ambiguities may strongly impact OSM's ability to offer a complete representation of cycling traces. Consequently, the lack of a complete representation may undermine OSM's role as an enabling technology for cycling information systems, compromising the quality of the tools and services that rely on OSM for road network data. For example, routing engines may overlook popular but misrepresented cycling paths, resulting in suggested routes that may be longer, steeper, or less safe. Moreover, mobility analysis tools will generate inaccurate interpretations about the cycling activity, leading to flawed recommendations for urban planners and decisions that fail to reflect the actual cycling behaviour.

1.3 Objectives

Despite its obvious potential as a shared reference system for cycling analytics, there are clear signals that challenge the ability of OSM to take that role. Based on the previous analysis, we formulate the following research question:

Can OpenStreetMap serve as a reliable reference system for cycling analysis and an enabling technology for cycling information systems?

This broad research question can be addressed through five specific research objectives:

0.1 Assess the existence of generic functionalities across cycling models, characterise the opportunity for horizontal services, and analyse how a digital platform strategy could bring many of those models into a common system with added value for different cycling.

0.2 Characterise the concrete causes and the main properties associated with the discrepancies be-

tween real cycling routes and how they are mapped into the OSM road network model.

0.3 Develop a supervised methodology to detect relevant discrepancies between cyclists' GPS traces and the routes produced by a map-matching algorithm and classify them according to their root cause.

0.4 Develop an unsupervised computational method to process sets of cycling traces and identify city areas prone to map-matching anomalies, and analyse how the knowledge about the map-matching anomalies can help cities uncover novel and actionable insights about real cycling behaviour.

0.5 Characterise the perspectives of key stakeholder groups about the gaps between common cycling paths and the cycling road network model in OSM, and explore how knowledge about these gaps can become a tool for congregating the perspectives of these stakeholders and reducing the representation gaps.

Regarding Objective 01, identifying the existence of generic functionalities would confirm the need for a new paradigm in mobility model development, one that fosters generativity and convergence. Defining the general properties of such a platform will help illustrate how a platform-based strategy can unlock the full potential of these models, while also clarifying the extent to which a common reference system is needed to aggregate and integrate data effectively. Regarding Objective 02, characterising the concrete causes and main properties of the discrepancies between cycling traces and map-matched routes will provide insights into whether OSM can serve as a reliable reference system for aggregating mobility data and its role as an enabling technology for innovation. Regarding Objective 03, the automatic identification and classification of these discrepancies will enable their detection at a larger and more realistic scale, providing an overview of the scope and magnitude of the problem. Regarding Objective 04, identifying systematic discrepancies will help reveal the urban areas most affected by this problem and indicate where interventions may yield the greatest improvements. Regarding Objective 05, characterising the perspectives of key stakeholders on these discrepancies will help to further shed light on their root causes, reveal how different stakeholders perceive the problem, and explore how potentially conflicting viewpoints can be reconciled into a common solution that benefits most citizens. Improving the alignment between cyclists' traces and OSM will ultimately enhance the quality of OSM as a reliable reference system for cycling analytics and as an enabling technology for innovation.

1.4 Thesis Structure

This first chapter defines and presents the context, motivation, and objectives of this PhD thesis, along with the research outputs made during the development of the reported research.

Chapter 2 resumes the state of the art relevant to this thesis. It starts by analysing several concepts related to cycling analytics, namely, bikeability, bike friendliness indexes, and analytics platforms. Then it explores related work in bicycle route preferences, OSM completeness, route representation politics and map-matching algorithms.

Chapter 3 addresses the first research objective by describing how current analytics models can be fragmented into smaller and generic units and discussing how a digital platform strategy could provide a more promising approach to their development, improving their value and enhancing their impact on the research community and society.

Chapter 4 addresses the second research objective by identifying the main source of discrepancies between cyclists' routes and the OSM road network model.

Chapter 5 addresses the third research objective by presenting a supervised methodology to automatically identify GPS traces that cannot be map-matched successfully into the OSM road network.

Chapter 6 addresses the fourth research objective by proposing an unsupervised methodology for extracting the segments of a GPS trace with no correspondence in the OSM road network and identifying the locations in a city where that problem is systematic.

Chapter 7 addresses the fifth research objective by assessing how cyclists, OSM mappers, and local authorities perceive this problem and discussing how the knowledge about these discrepancies can provide a convergence mechanism and a common ground to promote discussion between stakeholders for improving the cycling road network and cycling experience.

Chapter 8 concludes this thesis work by discussing the main contributions to each research objective, formulating an answer to the research question, and pointing out limitations and possible future research directions.

1.5 Research Publications

This research resulted in the following scientific publications.

Published Publications:

- Carvalho, C., Pessoa, R., & José, R. (2022). Cycling analytics for urban environments: From vertical models to horizontal innovation. In A. L. Martins, J. C. Ferreira, & A. Kocian (Eds.), Intelligent transport systems (INTSYS 2021) (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 426, pp. 135–148). Cham: Springer. https://doi.org/10.1007/978-3-030-97603-3_10 *(indexed in Scopus in* https://www.scopus.com/pages/publications/85126979287)

- Carvalho, C., & José, R. (2022, September). The gap between cycling practices and mapping services for smart cycling. In 2022 IEEE International Smart Cities Conference (ISC2) (pp. 1–7). IEEE. https://doi.org/10.1109/ISC255366.2022.9922089 *(indexed in Scopus in* https://www.scopus.com/pages/publications/85142033718)

- Carvalho, C., Ramires, M., & José, R. (2024). Automatic identification and classification of map-matching anomalies in cycling routes. In Proceedings of the 13th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS) (pp. 17–28). SciTePress. https://doi.org/10.5220/0012627700003714 *(indexed in Scopus in* https://www.scopus.com/pages/publications/85194201150)

Submitted for Review

- Carvalho, C., Ribeiro, P., José, R., "Systematic Identification and Analysis of Map-Matching Anomalies in Urban Cycling", submitted to Environment and Planning B: Urban Analytics and City Science

- Carvalho, C., Ribeiro, P., José, R., "Stakeholders' perspectives on the gaps between urban cycling routes and their representation on OSM", submitted to Transportation Research Interdisciplinary Perspectives.

1.5.1 Contributions per Chapter

The research contributions are closely aligned with its structure: Chapters 1 and 3 build upon the conference paper "Cycling Analytics for Urban Environments: From Vertical Models to Horizontal Innovation" (Carvalho et al., 2022). Chapter 4 is based on the conference paper entitled "The gap between cycling practices and mapping services for smart cycling" (Carvalho and Jose, 2022). Chapter 5 draws from the conference paper "Automatic Identification and Classification of Map-Matching Anomalies in Cycling

Routes" (Carvalho et al., 2024). Chapter 6 is based on the article entitled *"Systematic Identification and Analysis of Map-Matching Anomalies in Urban Cycling"*, submitted to *Environment and Planning B: Urban Analytics and City Science*. Finally, Chapter 7 draws from the article entitled *"Stakeholders' Perspectives on the Gaps Between Urban Cycling Routes and Their Representation on OSM"*, submitted to *Transportation Research Interdisciplinary Perspectives*.

Chapter 2

State of the Art

This chapter summarises the state of the art in key areas directly related to the work presented in this thesis.

The first part explores recent trends in cycling analytics by examining bikeability and bike-friendliness indexes. Although closely related, Lowry et al. (2012) distinguish between the two: Bikeability measures how conducive a city is to cycling by evaluating the connectivity of the bikeway network to key destinations, as well as the perceived comfort and convenience of cycling. Bike-friendliness builds upon bikeability by incorporating evaluations of municipal policies, investments in cycling infrastructure, educational initiatives, and the community acceptance of cycling. These indices represent an important category of models used to systematically and quantitatively measure how well the built environment supports cycling. Reviewing these indices provides an overview of the different dimensions that must be considered to evaluate cycling, and thus the diversity of data sources, transformations, and visualisations expected to be included in the platform.

The second part reviews a range of analytics platforms, focusing on solutions specifically designed for cycling mobility while incorporating relevant scientific studies and platforms that propose innovative approaches. This analysis identifies key features and best practices that should inform the development of a new cycling analytics platform.

The final part reviews literature addressing the role of OSM as an enabling technology for cycling services from four complementary perspectives: bicycle route choice preferences, OSM completeness, OSM representation politics, and map-matching and its limitations. First, route choice analysis gives an overview of the preferred and deterring route attributes for cyclists. This highlights what could influence cyclists to choose alternative paths instead of the expected path and provides a context for structuring the work in the following sections. Secondly, the literature on OSM provides an overview of how researchers have assessed OSM, its quality, and its limitations. Thirdly, the politics of OSM mapping provides insights about the mapping process, the conflicts between mappers to achieve an accurate representation of reality,

and what these conflicts can reveal. Finally, the literature on map-matching provides an overview of how researchers handled the limitations of current map-matching algorithms, either when map-matching is applied to cycling or when using incomplete road network data.

2.1 Bikeability

Bikeability is a relatively recent term derived from the concept of *walkability*. It aims to describe how the urban environment contributes to cycling adoption (Porter et al., 2020) and to identify areas conducive to cycling (Winters et al., 2013). The concept is closely related to other terms such as bike-friendliness (de Matos et al., 2021), bike-attractiveness (Kamel et al., 2020), and bike accessibility (Saghapour et al., 2017).

A bikeability index is typically generated through the weighted aggregation of multiple factors, allowing for comparing different paths (Krenn et al., 2015). However, there is no consensus in the literature regarding its exact definition or the attributes that should be used to assess the cycling environment (Muhs and Clifton, 2016). Lowry et al. (2012) defined bikeability as the comfort and convenience of a bikeway network for accessing important destinations. Grigore et al. (2019) expanded this definition by incorporating the perceived safety and attractiveness of streets and intersections along routes, as well as travel distances. Hamidi et al. (2019) offered a different perspective and defined bikeability as the ability to use bicycles to access public transport stations, such as trains and buses. Other authors, such as Long and Zhao (2020), measured bikeability using usage frequency and the built environment.

Despite their differences, these indices are a powerful tool for promoting cycling. They can help identify areas with poor conditions for biking, acting as a guide for urban planners, designers, and policymakers to target their efforts and investments to develop a more bicycle-friendly transport environment (Krenn et al., 2015). In addition, these indices can drive bicycle adoption, indicating to less experienced cyclists which places are more pleasant to ride.

Indicators

In literature, the most commonly used attributes fall into the following categories:

City topology: Typically measuring slope or elevation data. Steep terrain poses significant challenges for cyclists. Uphill routes require greater physical effort, while downhill sections can be hazardous due to increased speeds (Grigore et al., 2019).

Bikeway infrastructure: Assessing several characteristics of the bike network, including the pres-

ence of bikeways (Gu et al., 2018), their width, typology, whether they are dedicated to cyclists or shared with pedestrians or with motorised traffic (Schmid-Querg et al., 2021), and network coverage. Less common factors include the availability of bicycle parking facilities (Lin and Wei, 2018) and pavement quality (Manum et al., 2017).

Connectivity: Typically assessing the density of the bikeway network, its spatial coverage, and the density of intersections (Kamel et al., 2020; Lin and Wei, 2018).

Road network: Accounting for factors that may affect cyclist safety and comfort, such as the number of traffic lanes (Gholamialam and Matisziw, 2019), the presence of car parking spots (Manum et al., 2017), or shoulder width (Lin and Wei, 2018).

Traffic: This includes indicators regarding motorised traffic volume, heavy traffic areas, cyclist flow, and speed limits (Arellana et al., 2020).

Environment: Indicators in this category include the aesthetic quality of the landscape (e.g., greenery, trees, or water features), air quality, and the availability of lighting (Grigore et al., 2019; Porter et al., 2020; Tran et al., 2020).

Land use: Indicators describing activities carried out in an area (e.g., office, retail, industrial, educational), its points of interest (**POIs**), and demographic information (Long and Zhao, 2020).

Data Collecting Techniques

The data collection techniques used in these studies can be grouped into five categories: surveys, systematic observations, proprietary data, online data, and sensing.

Surveys: Some indices primarily rely on surveys, questionnaires, and on-site interviews with random or selected experienced cyclists. In these activities, participants are typically asked to share their opinions on their cycling experience and to evaluate the cycling environment. For example, Schmid-Querg et al. (2021) asked local commuters to rank different attributes of road segments and intersections, including the quality of bicycle infrastructure, motorised speed limits, bicycle parking, and intersection quality. The final index was derived from a weighted aggregation of these scores. However, the results from such methods tend to be very subjective, influenced by the participants (e.g. incorrect answers in questionnaires) or by the researchers (e.g., interpretation of qualitative statements) (Berger and Dörrzapf, 2018). Despite this, the approach remains valuable for comparing the perceptions of novice and veteran cyclists about the same environment and for exposing the factors that may be discouraging people from adopting cycling as their primary mode of transportation. In reality, every cyclist has their preference and perspective on the cycling ecosystem, and what is considered a good biking environment for some may not be the

same for other cyclists or different cycling purposes (Ma and Dill, 2015). Although conducting surveys and questionnaires can be time-consuming, recent technological advances in smartphones enable the automation of surveys through mobile applications, significantly reducing the time and effort required to collect these data.

Systematic observations: Other approaches are based on systematic observations of cycling infrastructure, either on-site or remotely using information systems. These approaches guarantee objective bikeability measurements. However, these assessments require well-trained professionals and do not scale efficiently in larger cities or multiple locations. For example, Arellana et al. (2020) used Google Street View imagery to evaluate several properties of each road in Barranquilla city, including the presence of bicycle infrastructure, the quality of bike path pavement, and the presence of obstacles. They reported that the assessment process, conducted by two people, took six months to complete.

Proprietary Data: Some authors use data provided by organisations to compute their indices. For example, Long and Zhao (2020) relied on bike usage data from a bike-sharing company and street network information from a navigation company. However, a limitation of this approach is that companies and public organisations may be reluctant to share their data, as it could be integral to their business models or be subject to legal constraints (e.g., user privacy). This reluctance may limit the applicability of such indices in regions where such data are not made available, despite their existence.

Online data: Other researchers explore freely available online data, including datasets that cover multiple cities and countries, such as OSM, census data, aerial imagery, and various official databases (Gu et al., 2018; Tran et al., 2020; Kamel et al., 2020; Saghapour et al., 2017). For example, Gu et al. (2018) evaluated the bikeability of four Chinese cities using exclusively open data, which included street network data from OSM, high-resolution street view images from Tencent, POIs from Gaode, and building footprint raster data from Baidu. Similarly, Tran et al. (2020) used open-source datasets, including land use data, road network information, Google Street View imagery, POIs, and a digital elevation model (**DEM**) from online databases. Kamel et al. (2020) focused on cyclist-car crashes (severity and location), bike and street networks, zonal boundaries, and annual average daily bike traffic, all obtained from public databases. Saghapour et al. (2017) evaluated accessibility levels in metropolitan Melbourne, Australia, and relied on mesh blocks from the 2011 Census, POIs from the Public Sector Mapping Agency, bicycle network data from the Victorian Government, and demographics and trip data provided by the Victorian Integrated Survey of Travel and Activity dataset.

Despite the high accessibility of online data, the challenges and concerns regarding its completeness and accuracy remain present. This is especially true for collaborative projects like OSM, where the data

is maintained by volunteers in their spare time. As a result, researchers must carefully verify and address any gaps in the data to ensure its reliability.

Sensing and mobile applications With recent technological advances, cyclists can now equip themselves with mobile applications, smart bands, smartwatches, and bicycle gadgets at a relatively low cost. This enables the mass recording of various types of data, which some authors use to compute indicators such as comfort, safety, or travel times (Berger and Dörrzapf, 2018; de Matos et al., 2021).

For example, Berger and Dörrzapf (2018) proposed a framework to identify stressful events during bicycle rides by combining different data sources and validated their findings with pre- and post-questionnaires. They used a GPS sensor, accelerometer, eye tracker, EEG, and EDA device to measure biophysiologic parameters and detect stressful events. In another study, Resch et al. (2020) used physiological measurements to identify areas of clustered stress and relaxation. The physiological data were complemented by an eDiary and cross-validated with first-person videos and georeferenced questionnaires. de Matos et al. (2021) combined data from a Global Navigation Satellite System (**GNSS**) data logger, a heart rate recorder device, and a video camera to quantify the travel time, caloric expenditure, effort distribution, and comfort for each segment of the road network. Additionally, they used official data sources to assess safety, infrastructure performance, and emission hotspots.

The main benefit of this approach is the possibility of computing bikeability in other and larger environments. Additionally, as researchers can gather real-time data, they can continuously monitor various aspects of cycling.

Indices representation

There are three main ways to represent a bikeability index:

City-level: A single overall value is assigned to the entire study area, such as in (Porter et al., 2020). This approach provides a high-level overview of how supportive a city is for cycling and enables comparing different cities.

Zonal level: The study area is divided into smaller zones (administrative or grid-based), with a bikeability score for each zone, such as proposed by Kamel et al. (2020). This approach provides a more detailed understanding of spatial variability within a city, allows for local benchmarking, and is valuable for identifying zones with lower levels of bikeability.

Network-level: Each street segment and intersection is individually assessed and assigned a value, as in (Schmid-Querg et al., 2021). This fine-grained approach offers the highest resolution and is especially useful for routing applications and infrastructure planning. For instance, de Matos et al. (2021) developed

a cycling navigation information system that supported seven different indicators, including travel time, safety, and physical effort. After collecting and processing the necessary data, they classified each road segment and applied Dijkstra's shortest path algorithm to compute optimal routes for each indicator across three case studies. This approach allows cyclists to customise their route according to their preferences. For example, a cyclist who prioritises comfort can select the route that optimises this indicator. Additionally, cyclists can compare a route across other indicators, checking, for instance, whether the most comfortable path is also safe or whether it involves a significant detour compared to the shortest route.

Discussion

During the analysis of bikeability indices, their diversity was particularly evident. The studies employ a wide range of methodologies, differing in the indicators considered, the data sources used, and how results were presented.

Firstly, the analysis of different methodologies supports a lack of consensus on the bikeability concept and on which indicators should be used to quantify bikeability. Each study adopts a unique set of indicators, with no single indicator common to all studies. The most common indicators were related to slope, the presence and typology of cycling infrastructure (e.g., existence, width, and separation from traffic), network connectivity, land use, and the cycling environment. The final index was typically computed through a weighted aggregation of the selected indicators.

Secondly, data collection methods also varied significantly. These include systematic observations, surveys, data provided by organisations, open-access repositories, mobile applications, and sensor-based devices. The type and availability of data directly influenced which indicators could be evaluated. In many approaches, the data collection and processing methods were time-consuming, especially those relying on manual observations.

Regarding results presentation, most studies presented the results graphically. Some divided the city or region into smaller zones, assigning a bikeability score to each, while others assessed bikeability at the street level. This flexibility allows researchers to creatively explore different dimensions of urban cycling conditions. However, these methods are often highly tailored to the specific datasets available and are usually implemented for a single city or study area.

Applying a bikeability index to a different context poses several challenges. Firstly, the same types of data are often not available. Even when similar data exist, it is usually in a different format and requires preprocessing. Furthermore, the lack of accessible or well-documented source code makes it difficult to understand and replicate the original algorithms. As a result, researchers are frequently forced to recreate

the methodology from the ground up. These challenges create a substantial barrier for those wishing to test or adapt a model to a new city. The need to repeat the entire process from data collection to algorithm implementation can be discouraging, leading to lost time, missed opportunities, and a slower pace of innovation. Ultimately, this jeopardises their widespread use and the development of more robust bikeability assessment models.

2.2 Bike-Friendliness Indexes

This section reviews a selection of widely recognised bike-friendliness indices, highlighting the components or indicators considered and the methodology to compute the final score. Bike-friendliness is a term very similar to bikeability, and even though some authors use both terms interchangeably, Lowry et al. (2012) distinguished between the two and defined bikeability as the comfort and convenience of a bikeway network for accessing important destinations and bike-friendliness as a combination of bikeability, laws, policies, and community education.

The Bike Score Index

One of the most popular bike-friendliness indexes is Bike Score® (Score, 2018), which builds upon the methodology originally developed for Walk Score®. The Bike Score methodology is based on thousands of votes from the Walk Score community and expert advice. The score is available for any address within the United States and Canada. However, the public website only provides results for cities with a population of 200,000 or more, covering a total of 130 cities.

The final score ranges from 0 to 100 and is calculated as the average of four equally weighted components: The **Bike Lanes Score** component assesses the quality of the cycling infrastructure based on data obtained from OSM. It accounts for the length and typology of nearby road segments, namely bike paths, bike lanes, and shared infrastructure. For a given location, the value for the component is obtained by summing up the length of all nearby bike lanes using a distance decay function, giving greater importance to closer segments. The **Hill Score** component evaluates the topographic steepness of the terrain by identifying the maximum gradient within a 200-meter radius of each location. This score is based on data from the National Elevation Dataset provided by the USGS. The **Destinations and Connectivity** component measures the network's distances to a diverse set of amenities and calculates connectivity metrics such as average block length and intersection density. Finally, the **Bicycle Mode Share** component quantifies how many people commute by bicycle, using data from the U.S. Census.

This index relies on open or accessible data (e.g., OSM and U.S. Census), making it relatively easy to apply in multiple geographic locations. However, its dependence on census data, conducted only once every ten years. This limits its ability to reflect recent shifts in commuting habits or demographic changes.

Copenhagenize Index of bicycle-friendly

The Copenhagenize Index (Copenhagenize, 2019) is a biennial ranking assessing the bicycle-friendliness of cities worldwide. The index evaluates the urban cycling environment on 14 parameters across three main categories, namely streetscape, culture, and ambition. Each parameter is rated on a 0–4 scale.

The **streetscape parameters** assess the quality and extent of the cycling network. These include the presence and typology of bicycle infrastructure (e.g., protected and segregated lanes, cyclist prioritisation, and network connectivity), the availability of bicycle facilities (e.g., bike racks), and the implementation of traffic calming measures (e.g., speed limits and awareness campaigns). The **culture parameters** measure the social acceptance and integration of cycling as a primary mode of transport. This includes gender distribution among cyclists, bicycle modal share and its growth, perceived safety, public image of cycling, and the use of cargo bikes. The **ambition parameters** evaluate the level of commitment from local authorities. Indicators include government advocacy for cycling, political support, the presence of bike-sharing systems, and the integration of cycling in urban planning.

In 2019, the index was applied to 115 cities worldwide. However, this index is not fully automated and requires expert assessments, which limits its scalability and reproducibility. On the other hand, relying on assessments by experts makes this process independent of specific datasets that may not be uniformly available across cities, allowing for more consistent cross-city comparisons despite data limitations.

Bicycle Cities Index 2019

The Bicycle Cities Index (Coya, 2019) was developed by Coya, an insurance provider acquired by Luko and subsequently by GetSafe. Although the original website is no longer available, it featured evaluations for 90 cities worldwide.

This index assessed 16 indicators grouped into six main categories: **Weather** indicators measured factors such as hours of sunshine, average rainfall, and frequency of extreme weather events. **Bicycle usage** indicators evaluated the percentage of people using bicycles for daily activities. **Crime and safety** indicators evaluated the number of cycling-related fatalities, accidents, and thefts. **Infrastructure** indicators assessed the availability of bicycle shops, the existence and quality of cycling infrastructure, and public investment. **Sharing** indicators evaluated the number of bike-sharing stations and the availability

of shared bicycles. Finally, indicators related to **events** evaluated initiatives that promote cycling, such as No Car Days or Critical Mass Score.

Each component was normalised on a scale from 1 to 100. The final index score was obtained by weighted aggregation of these normalised values.

They used a wide spectrum of data sources, including online repositories and APIs (e.g., OSM, World Weather Online API), local statistical departments, international organisations (e.g., Greenpeace, the World Bank, and the World Economic Forum), and social media platforms (e.g., Facebook).

Bike Friendly Index

The Bike Friendly Index (**BFI**) (Figueiredo and Vale, 2018) was developed as part of an initiative by the BEAM research group in collaboration with Pulsar!.

This index relies on open-source datasets to ensure replicability, including Census data, Copernicus, OSM, and the BASE Portal. It assesses the urban cycling ecosystem across five main components encompassing twelve indicators: slope, built environment, infrastructure, political commitment, and bicycle usage. The final score ranges from 0 to 10 and is calculated through a weighted aggregation of the component scores.

The researchers made the 2018 BFI results publicly available via an interactive website[1], providing values for all municipalities in mainland Portugal. The index revealed that cycling is still in its early stages of development in the country, with a national average score of only 2.3 out of 10. Murtosa (6.08) and Lisbon (5.84) were the only municipalities achieving scores above 5.

Most Bike-Friendly Cities in the U.S.

In 2022, Anytime Estimate (Estimate, 2022), a company specialising in real estate insights, released a ranking of the 50 Most Bike-Friendly Cities in the United States. The final score was calculated through a weighted aggregation of 10 factors, including a bikeability score, the percentage of workers commuting by bicycle, the number of bicycle shops, bike-sharing services, bike trails, status as a Vision Zero Community, interest in cycling (via Google Trends), historical precipitation, and air quality.

Anytime Estimate used data from different public repositories, including government agencies (e.g., U.S. Census Bureau, U.S. Department of Transportation, and U.S. National Centers for Environmental Information), non-profit organisations (e.g., Rails-to-Trails Conservancy and Vision Zero Network), Walk Score, Google Trends, and Yelp.

[1] https://www.bikefriendlyindex.com/mapa/

Best Biking Cities in the U.S

In 2023, LawnStarter (Maive, 2023), a lawn care company, ranked the 200 biggest U.S. cities for biking conditions. Their index was based on the weighted aggregation of 27 indicators, grouped into six major categories: Access, Ride Quality, Safety, Climate, Community, and Events.

Access evaluates how easily residents can reach bike trails, bike-sharing services, and bike shops. **Ride Quality** assesses the overall infrastructure, including bike score, ratings of bike trails, and presence of top-rated cycling routes. **Safety** measures include the presence of bike-friendly laws, the number of thefts, and cycling fatalities. **Climate** considers the number of extreme temperature days, monthly precipitation, air quality, and average sunshine. **Community** factors focus on engagement and adoption, including the number of bike commuters, local cycling clubs, meetups, and supporting businesses. Finally, indicators related to **events** track cycling-related activities, including races, tours, and community rides.

By comparing Bike Score, the Anytime Estimate and LawnStarter indexes, in every index, San Francisco and Portland appear in the top 3, highlighting their investment in cycling. However, some cities have different positions in the ranking. For instance, New York ranked 11th in the Anytime Estimate Index but climbed to 2nd place in the LawnStarter Index. Conversely, Minneapolis appeared in 1st place in the Bike Score index and dropped to 10th in the LawnStarter index. Similarly, San Jose dropped from 3rd in 2022 to 12th in 2023, illustrating how differences in methodology and weighting can influence a city's position.

PeopleForBikes' Best Places to Bike

PeopleForBikes City Ratings (PeopleForBikes, 2023) aim to identify, evaluate, and compare the best cities and towns for bicycling. In 2024, the index evaluated more than 2,500 cities worldwide.

The City Ratings are published annually each summer and assess six key factors, summarised by the acronym SPRINT: Safe Speeds, Protected Bike Lanes, Reallocated Space for Biking and Walking, Intersection Treatments, Network Connections, and Trusted Data. The tool used to calculate the City Ratings is the Bicycle Network Analysis (**BNA**) software, which is freely available and built upon open data sources such as OSM for road network data and Census demographic data.

The BNA scores cities in four steps: data collection, traffic stress analysis, destination access analysis, and score aggregation. In the second step, each road segment within the city boundaries is categorised as low or high in traffic stress based on predefined criteria. In the third step, the BCA calculates the accessibility to destinations from each Census block, considering only low-stress routes within a maximum cycling distance. Destinations are grouped into six categories: People, Opportunity, Core Services, Recreation, Retail, and Transit. Each block then receives a rating based on a weighted aggregation. Finally, in the

fourth step, the BNA calculates the category scores and the overall score for the entire city based on the population of each census block.

Propensity to Cycling Tool

The Propensity to Cycle Tool (**PCT**) (Lovelace et al., 2017) is a web-based platform designed to estimate cycling potential and the associated health and CO_2 benefits at the street level across England and Wales.

Using data from the 2011 Census on commuting and cycling to school, the PCT identifies desire lines—the most common routes between the centroids of origin and destination areas. For each desire line, it maps both the shortest and the quietest routes, offering a comparison of different path characteristics.

The PCT also simulates how travel patterns would change under different cycling uptake scenarios. Each scenario assigns a probability for converting car trips to bicycle trips, based on factors such as trip distance and hilliness. Example scenarios include the governmental target, gender equality in cycling uptake, and the "Go Dutch" model.

Users can explore data and results across multiple spatial levels, ranging from Lower Layer Super Output Areas (**LSOA**s) to Middle Layer Super Output Areas (**MSOA**), providing a detailed and flexible analysis of cycling potential.

City readiness for cycling

Zayed (2016) assessed the readiness of 42 Egyptian cities to become bike-friendly by focusing on socioeconomic and urban cycling characteristics.

The authors identified 12 urban attributes commonly associated with bicycle-friendly cities. From there, the authors selected 20 cities featured in the 2015 Copenhagenize Index and conducted a principal component analysis to determine which variables most correlate with cycling adoption. The final index considered five key variables: city population, road network length, city form, city area, and motorised transport modal split.

However, when applying this methodology, the authors reported difficulties computing some attributes due to limited data availability.

The Potential for Cycling

Silva et al. (2019) proposed a method to assess the cycling potential of starter cycling cities using a three-dimensional approach, considering: **Target-Population**, evaluating the size and spatial distribution of the population with a higher propensity to cycling (e.g., age, density, car ownership); **target areas**, assessing

the amount and spatial distribution of areas providing suitable conditions for cycling (e.g., accessibility to education facilities, city centres, railway stations); and **political commitment to cycling**, measuring the extent to which local policies support cycling (e.g., coverage of cycling infrastructures and availability of bicycle parking). Each dimension's score was computed through a weighted aggregation of attributes, with values normalised on a 0–5 scale. The attribute weights were defined based on a thorough literature review. After computing the scores, areas were categorised into classes A (score > 3.7) to C (score < 2.3) for each dimension. Each area was assigned a three-letter classification (e.g., AAA), representing its rating in each of the three dimensions. This methodology was applied to various Portuguese cities in (Silva et al., 2018) and (BooST, 2018).

Urban Mobility Readiness

The Urban Mobility Readiness Index (Forum, 2024) was developed through a collaboration between the Oliver Wyman Forum and researchers at the University of California.

The index incorporates 71 metrics across five key dimensions, with 15 new metrics introduced in the 2024 edition. These dimensions evaluate: **Social Impact**, measuring the social, environmental, economic and business benefits of mobility systems; **Infrastructure**, assessing aspects such as micro-mobility availability, public transit systems, connectivity, and overall infrastructure quality; **Market Attractiveness**, focusing on the availability of public transit options and smart mobility; **System Efficiency**, evaluating aspects such as the demand and transport planning or modal mix optimisation; **Innovation**, considering the adoption of new technologies and the quality of human capital;

In addition to the overall score, the index provides three sub-indices focusing on sustainable mobility, public transit, and technology adoption.

The index has expanded its coverage over time. In its 2024 edition, it evaluated 70 major cities worldwide, with San Francisco ranking first overall.

Discussion

Table 1 summarises the bike-friendliness indices examined in this review. It presents the date of the most recent update, the frequency of updates, the method of publication, and the geographical scope of each index.

Similar to bikeability, bike-friendliness indexes aim to quantify how conducive a city is to cycling. However, they extend bikeability: In addition to evaluating infrastructure and environmental conditions, they often incorporate indicators related to policy frameworks, public investment, cycling events, and

Table 1: Bike-friendliness indexes assessed in this review

Name	Year	Update Frequency	Representation	Spatial Coverage	Main Components
Bike Score Score (2018)	2021	-	Heatmap	Any address in the United States and Canada	4 equally weighted components: Bike lanes, Hills, Destinations and road Connectivity; Bike commuting mode share
Copenhagenize Index Copenhagenize (2019)	2019	Biennial	Zonal (Cities)	115 Cities Worldwide	14 indicators across 3 main components: Streetscape, Culture, and Ambition
Bicycle Cities Index Coya (2019)	2019	-	Zonal (Cities)	90 Cities Worldwide	16 city indicators across 6 main components: Weather; Bicycle Usage; Crime and Safety; Infrastructure; Bike-Sharing; Events;
Bike Friendly Index Figueiredo and Vale (2018)	2018	-	Zonal (Municipal)	Portugal	12 indicators across 5 main components: Slope; Land Use; Infrastructures; Policies; Bicycle Usage
Anytime Estimate's Most Bike-Friendly Cities in the U.S. Estimate (2022)	2022	-	Zonal (Cities)	United States	10 indicators
Lawnstarter's Best Biking Cities in the U.S Maive (2023)	2023	-	Zonal (Cities)	United States	27 indicators grouped into six components: Access; Ride Quality; Safety; Climate; Community; Events
PeopleForBikes' Best Places to Bike PeopleForBikes (2023)	2024	Annual	Zonal (Cities)	+2,500 Cities worldwide	6 indicators
City readiness for cycling Zayed (2016)	2017	-	Zonal (Cities)	42 Cities in Egypt	5 Variables: City Form; City Area; City Population; Road Network Length; Motorized Transport Modal Split
Propensity to Cycling Tool (PCT) Lovelace et al. (2017)	2017	-	Zonal (LSOA; MSOA)	United Kingdom and Wales	Commute trips
The Potential for Cycling Silva et al. (2019)	2018	-	Disaggregated spatial analysis; Zonal (Cities)	Portuguese Cities	10 indicators across 3 categories: Target-population; Target-Areas; Polical Commitment
Urban Mobility Readiness Forum (2024)	2024	Annual	Zonal (Cities)	70 Cities Worldwide	71 indicators across five dimension: Social Impact; Infrastructure; Innovation Market Attractiveness; System efficiency;

community perceptions.

Some indices, such as Bike Score® and the Copenhagenize Index, are applied across a wide range

of cities. This broad applicability allows for inter-city comparisons and can motivate public authorities to prioritise cycling initiatives to achieve higher rankings.

However, despite their utility, these indices face similar challenges to bikeability indices, primarily due to their closed or opaque methodologies. Even when the methodology is published, it is typically limited to high-level descriptions intended to justify the index's credibility rather than to enable replication.

As a result, individuals or researchers interested in evaluating their cities often lack the necessary guidance to reproduce the models, including the algorithms used or the required data formats. This limits the development of more comprehensive and generalisable indices that could facilitate cross-city comparisons and better support evidence-based decision-making by public managers.

2.3 Cycling Analytics Platform

This section describes analytics platforms whose focus or properties are related to the work described in this work. Namely, we prioritised platforms dedicated to cycling analytics and included some platforms dedicated to mobility in general, some scientific articles, and other platforms with relevant features. The goal of this review is to identify common features and limitations of current systems.

For each platform, we evaluated its main purpose; the existence of an analytics dashboard or map where users can perform their analysis or visualisations; the existence of a mobile app where cyclists can record their activities; the existence of an API where developers or other interested parties can submit or retrieve data; the existence of a web client where cyclists can plan routes, manage their accounts, and see personal statistics; how the owners monetise their ecosystem; the existence of community-engaging strategies; its main source of data; and the ability for users to develop and publish their mobility models. Finally, we also registered the level of access we had when evaluating a system or a component, namely, full demo, in cases where we had access to the full system features without paying; free demo, in cases where we had access to a demo version of the system with limited features; documentation, in cases where we had access to documentation describing premium features; and paper, in cases where the platform was described in a research paper and was not implemented or available. On some platforms, we had different access rights to different components. For example, when reviewing Cyclers Analytics, we reviewed the mobile application's freely available features, but we only had documentation to review the analytics component.

Cyclers

Cyclers (2019) is an application ecosystem aiming to leverage cycling, which includes a mobile application, a web route planner, and an analytics platform.

The "Cyclers Analytics" platform allows users to "visualise, analyse & optimise the micro-mobility transport system". Cyclers Analytics accepts data in different formats (e.g., GPX trajectories or GeoJSON data) and integrates it with Cyclers' built-in data, such as maps. According to the available documentation, the platform can generate reports using the trip and route analytics and system analytics. The trip and route analytics include: "Trip Patterns", calculating the number of riders moving between different zones at different times; "Traffic Intensities", showing the intensity and directionality of micro-mobility traffic on the street level; and "Traffic Speed and Fluency", showing how fast cyclists go and where they need to wait. The system analytics include: "Link Attractivity", showing which links attract riders and which links the riders prefer to avoid; "Ridership Potential", showing where cycling/micro-mobility has the highest potential to attract riders; and "Travel Times", showing how long it takes to reach a given area by bike or scooter.

The mobile application uses a freemium model in which users have free access to basic features (route suggestions from A to B). By paying a regular fee or a one-time payment, users have access to premium features, such as route customisation, battery-saving navigation, or turn-by-turn navigation. The system uses gamification to incentivise users to ride, ranks users by distance or rides recorded, and promotes challenges and competitions.

The web application allows users to plan their routes based on their preferences, such as minimum distance, bicycle type, traffic, surface, and hills. The planner gives users the best matching route for their preference, along with a set of alternative routes and their characteristics. For each route, the planner shows the bike-friendliness, the expected travel time, the percentage of the route with high traffic, the percentage of unpaved or badly paved route segments, the percentage of steep segments with a slope greater than 6%, and the elevation profile. It also includes the distance, the expected caloric expenditure, and CO2 emission savings. Figure 1 show an example of a route suggestion.

Finally, Cyclers also offers a routing and navigation API/SDK for cycling and micro-mobility, where developers can have access to circular routes or routes between two points.

Strava Metro

Strava Metro (Strava, 2019) is a web platform for mobility analytics provided by Strava, and it is possible to analyse cycling and walking activities. The demo included two main features: a dashboard and a map.

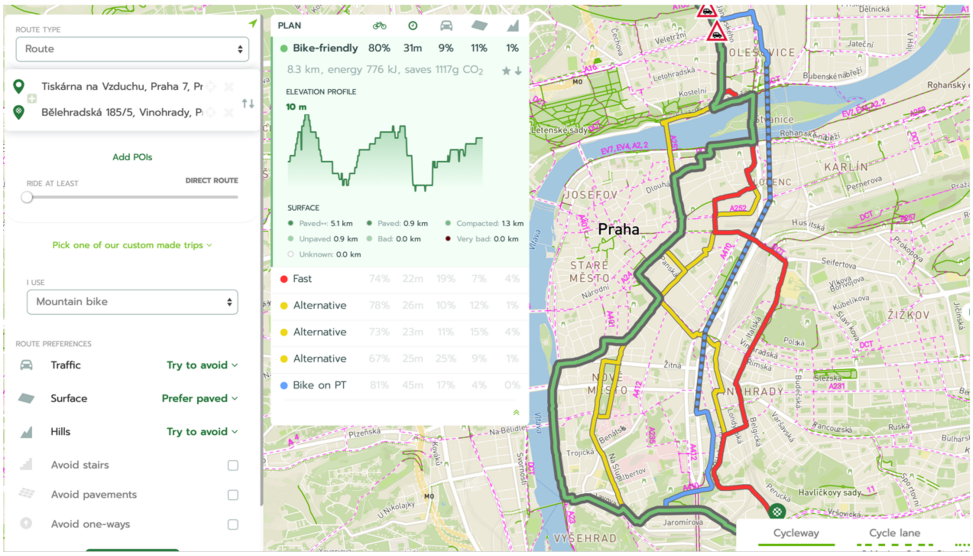


Figure 1: Cyclers Navigate Route Planner (Cyclers, 2019)

The dashboard included two sections: one numerical and one graphical. The numerical section presents information referring to the last week, month, or year, which includes the total and weekend trips, the number of cyclists, visitors, CO2 saved, and the cumulative distance and elevation gain. The graphical section presents the monthly and weekly evolution of the total number of trips, people, distance, and CO2 saved. It also presents the demographic distribution of riders and bike types. In some graphs, it offers the possibility to filter by trip purpose or by year.

The map visualisation offers five options for analysis, namely "Streets", "Routes", "Origins & Destinations", "Travel Times", and "Heatmap". The "Streets" option allows users to obtain aggregated information at a road level. This included trip information (total activities, the average speed, the distribution of trip purpose and bike type), demographic distribution (gender, age), and time distribution (trips during the morning and evening peaks, trips by hour, by day of the week, and by month). The "Routes" option allows users to obtain route information from A to B based on directness, popularity, or a weighted combination of both. The "Origins & Destinations" section allows users to visualise the places where most people start or end their rides. Users could select an area and obtain detailed information on where cyclists come from, where they go, and the average distance and duration of trips. Users could also filter by trip purpose, time of the day, day of the week, and time interval. The "Travel Times" section provides an overview of which points are accessible within a maximum travel time given a starting or destination point. Finally, the "Heatmap" section represents the most popular roads in a city.

Strava Metro uses the data collected by the Strava mobile application, one of the most popular mobile applications for sports. The mobile application employs several gamification strategies to incentivise users

to engage in physical activities, including allowing users to establish their groups and challenges.

Strava adopted a freemium model for their mobile application. Regarding Strava Metro, while Strava provides this tool for free, interested users must first gain authorisation. To become a Strava Metro partner, interested users must submit an online questionnaire describing their background and projects, depending on which they may or may not be accepted. Finally, Strava provides an API for users to retrieve their data and a web client for users to manage their accounts and see their statistics.

Bike Citizens Analytics

Bike Citizens Analytics (Citizens, 2017) is a GPS data analysis tool designed specifically to analyse bicycle traffic provided by Bike Citizens. Some of the available features include detailed information about specific road segments, such as cyclist volumes, direction, and speed. The platform also provides maps showing the average speeds per road, relative speeds, and the attractiveness of different routes and provides heatmaps displaying the number of cyclists per road, the shortest routes simulated based on actual cycling patterns, and locations where delays commonly occur. These analytics are based on GPS data collected from users using the Bike Citizens mobile application.

The mobile application helps users with city navigation, employs gamification through challenges, rewards, and badges to incentivise people to cycle, and allows users to contribute to the community by sharing their data. Bike Citizens employs a freemium approach, where users have access to premium features within a 7 km radius, which users can set only once. Bike Citizens also provides users with a web client where they can see their statistics, a web-client route planner, and complementary products and gadgets to attach to the bicycle.

Madrid Cycle track

Romanillos and Austwick (2016) developed the Madrid Cycle Track Initiative[2], a platform designed to collect cycling routes and other information from volunteers. Using this platform, users could register and download a GPS app, upload and draw their routes, visualise statistics about their activities, participate in groups, send messages, view the activities of other users, and track the evolution of trips made over time within the community. The website featured several dynamic maps to illustrate the real flows of casual cyclists and bike messengers in Madrid. This included an online map showing the routes of the volunteers and the bike messengers, a time-lapse animation showing the flow of cyclists, and another time-lapse animation showing the activity of the Madrid Bike-Share System.

[2] https://huellaciclistademadrid.es

Pedal Heat®

Pedal Heat® (KAPPO, 2016) was a platform that helps cities to visualise cycling traffic. Users could install a mobile application that incentivised them to ride a bicycle through virtual rewards. The recorded information was then sent to the platform to be visualised. The visualisation layers included the cycling infrastructure, a heat map of the most demanded street segments, and cyclists' flow in real-time. Currently, the platform and mobile application are not available online.

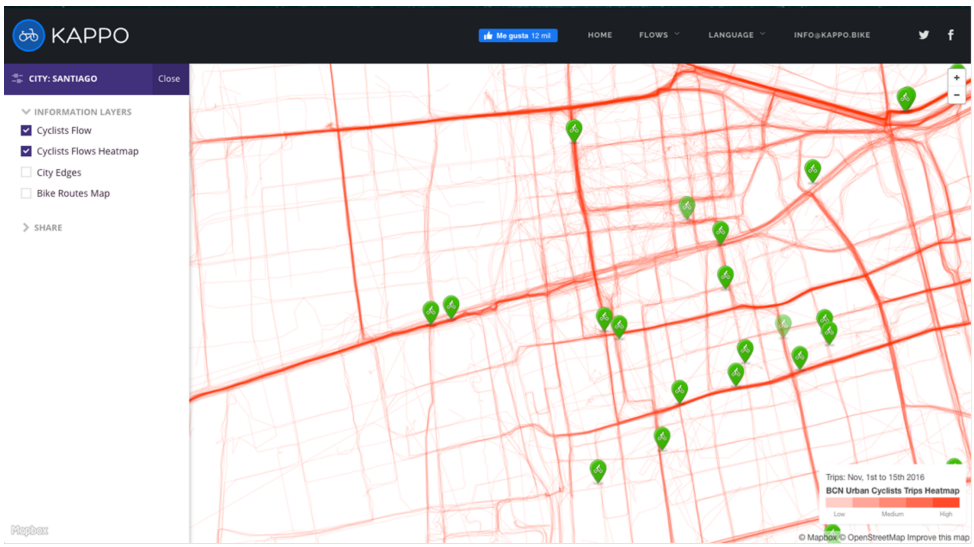


Figure 2: Pedalheat® - Urban Cycling Trips Heatmap (KAPPO, 2016)

Ride Report

Ride Report (Report, 2020) is a platform to manage micro-mobility, helping mobility providers to manage their system and components, such as shared e-bikes, e-scooters, and shared cars.

Mobility operators have access to a dashboard that enables performance analysis of their systems. Available metrics include total trips, average trips per day or time interval, trips originating from business districts, trip distance and duration, their distribution, origin-destination, and the most popular routes by trip mode and demographics. It is also possible to filter these metrics for a specific area. This platform is built based on the Mobility Data Specification (**MDS**), which is a set of Application Programming Interfaces (**API**s) for mobility operators and agencies to standardise their data to be shared, ingested, or analysed. It focuses on dockless e-scooters, bicycles, mopeds, and carsharing (OMF, 2021).

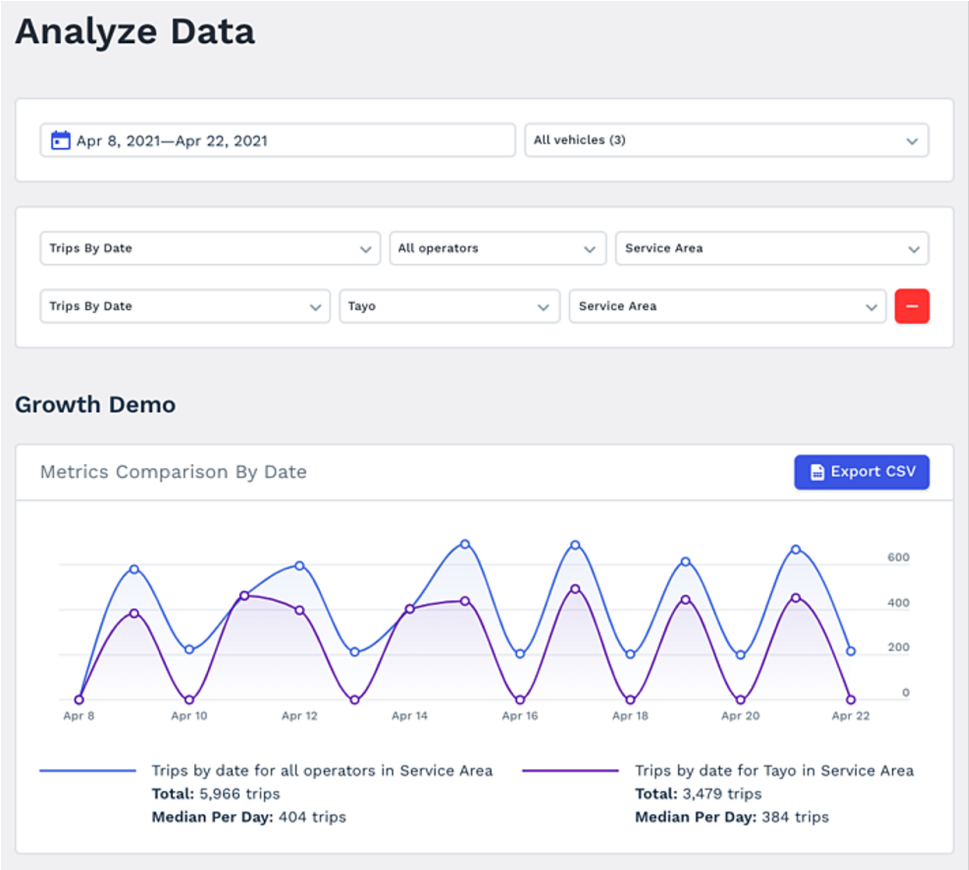


Figure 3: Ride Report Analyze Tab (Murray, 2021)

Currently, they offer an "Open Data Portal" featuring the "Global Micromobility Index", which provides insights into the total number of trips, trips per day, and average vehicles across all participating cities. The portal also ranks each city and enables users to filter results by mode of transportation (bikes, e-bikes, and scooters) and by time interval. Users can select a specific city to view detailed data, including a map highlighting the busiest streets, along with statistics on average vehicles, total trips, total distance, average trip speed, and temporal graphs displaying trips per day, trip distance per day, trips per vehicle per day, and active vehicles per day.

In the past, they offered a mobile application for users to record their cycling activities and evaluate the travelled route. The information recorded was sent to a centralised service, where it was processed and aggregated, resulting in a comfort map. This classified each street segment as comfortable or stressful. However, this application and interactive map were discontinued (Gomez, 2017).

Eco Counter and Eco Visio

Eco Counter is a company specialised in solutions to count pedestrians and cyclists. Their website provides two free resources: the "Cycling Data Dashboard" and the "Eco-Display World Map". The first is a dashboard that tracks the growth of cycling by country. The second is an interactive map showing the location of active counters and their historical activity. For each location, it is possible to obtain total counts, daily averages, and a histogram with daily, weekly, and monthly activity. They developed and commercialised the "Eco-Visio", an online platform to manage counting sites and analyse data. It also allows managers to measure the impact of weather and temperature on cycling and to create and share reports and dashboards with data visualisation modules. Finally, the "Eco-Visio API Evo" is a Representational State Transfer (**REST**) interface that provides access to count data.

StreetLight Data

StreetLight Data (StreetLight, 2019) is a mobility platform that aggregates data collected from multiple sources, including smartphones, navigation devices in connected cars and trucks, and data provided by Data Partners. The platform enables users to analyse various dimensions of mobility, such as the average annual daily traffic (**AADT**) for a study area, specific road, or entire region, and traffic volume at various intersections, along with trip speed, duration, and distance. It also provides insights into the origins, destinations, and routes between zones or roadways made by car, truck, bike, or walking; detailed statistics for each road segment; and insights into trip purposes and demographics.

xyzt.ai

The xyzt.ai is a no-code, self-service, collaborative platform for traffic and mobility analytics where users can visualise and analyse spatio-temporal data. By dragging and dropping their data or connecting sensor networks through an API, users can combine different sources to create personalised maps and dashboards and perform different types of analysis, such as trend analytics, wait time analysis, density analysis, origin-destination analysis, or dwell time analysis. The xyzt.ai is compatible with different data sources and types (e.g., floating car data, trips data, events data) and allows users to create custom algorithms and data formats. However, it is unknown if these custom algorithms rely only on built-in capabilities or if users can develop their own independently.

MoocRP: An Open-source Analytics Platform

Pardos and Kao (2015) developed an open-source analytics platform for massive open online courses. Despite not being related to cycling or mobility, we included this platform due to its unique features offered for researchers and developers. This platform follows a modular approach, allowing researchers to develop new analytics models, use their own datasets, or use those provided by others. The web application provides a user-friendly interface for listing approved analytics models and visualisations. Each component is associated with the datasets that can be applied to it. Additionally, an instance of the platform can be run on local machines.

To access a specific dataset, users must submit a data request, which is reviewed by an administrator who can either grant or deny access. Once granted, the dataset can only be downloaded once. Each dataset is linked to a folder where the data is stored. When a new folder is detected, the platform automatically adds it to the list of available datasets. The process of adding a new data model is manual: the administrator fills out a form, specifying both a user-friendly name and a system-friendly name for the dataset. The system then automatically creates the necessary directories. The platform incorporates several security mechanisms, and only approved affiliates from trusted institutions are allowed to use the platform.

Discussion

Table 2 presents the full list of the platforms analysed and summarises the results. In total, we analysed 10 different platforms. Six of those platforms were dedicated to cycling, three were broader platforms dedicated to mobility, and one was dedicated to massive open online courses. The latter platform was included because it allowed users to develop and reuse analytic models, an approach similar to our objective.

The platforms can be categorised into three groups. The first group consists of platforms dedicated to cycling analytics. Usually, a mobile application is available for cyclists to track their cycling activities. The data corresponding to cycling activities is then aggregated and presented to decision-makers through dashboards and maps, offering insights into cyclist activity and demand. These platforms often provide web clients for cyclists to manage their accounts, plan routes, and access detailed information about cycling conditions. Cyclers, Strava, and Bike Citizens exemplify this approach. The second group focuses on mobility analytics in general (StreetLight Data; xyzt.ai). These platforms do not specifically target cycling and lack dedicated mobile applications for cyclists or route planners. Instead, they provide urban planners with comprehensive analytics capabilities for a broader understanding of urban mobility. To obtain data, these platforms either establish partnerships with data owners or rely on the data from entities conducting

Table 2: Analytics Platforms Reviewed

Name	Purpose	Mobile App	Analytics Dashboard	API	Web Client	Price	Community Engagement	Data Sources	Own Models	Assessment Method
Cyclers Analytics	Cycling	Yes	Yes	Yes	Yes	Mobile App: Freemium; Analytics: Paid	Gamification	Mobile App;	No	Analytics (Docs) Mobile App (Free Demo) Web Client (Free Demo)
Strava Metro	Cycling; Pedestrians	Yes	Yes	Yes	Yes	Mobile App: Freemium; Analytics: Free with approval;	Gamification	Mobile App;	No	Analytics (Free Demo; Docs) Mobile App (Free Demo) Web Cliente(Free Demo)
Bike Citizens Analytics	Cycling	Yes	Yes	No	Yes	Mobile App: Freemium; Analytics: Paid;	Gamification; Digital Campaigns; Ride in Teams;	Mobile App; External Data	No	Analytics (Docs; Free Demo) Mobile App (Free Demo)
Madrid Cycle Track	Cycling	Yes	Yes	No	No	Free	No	Mobile App; Crowdsourced GPS traces;	No	Analytics (Docs; Full Demo)
Pedal Heat®*	Cycling	Yes	Yes	No	No	Free	Gamification	Mobile App	No	Analytics (Docs; Full Demo) Mobile App (Docs)
Eco Counter	Cycling; Pedestrians	No	Yes	Yes	No	World Map: Free; Eco-Visio: Paid;	No	Bike and Pedestrian Counters	No	Analytics (Docs; Free Demo) Evo-Visio (Docs)
Ride Report	Management of Shared Mobility	No	Yes	No	No	Free	No	Crowdsourced	No	Open Data Portal (Full Demo) Mobility Management Platform (Docs)
StreetLight Data	Mobility	No	Yes	Yes	No	Paid	No	Data Partners	No	Analytics (Docs)
xyzt.ai	Mobility; IoT; Maritime	No	Yes	Yes	No	Analytics: Paid	No	Own Datasets	No	Analytics (Docs)
MoocRP	Massive Open Online Courses	No	Yes	No	No	Free	No	Own and Public Datasets;	Yes	Analytics (Paper)

the analysis. The third group relates to platforms focused on managing micromobility and urban mobility. Although these platforms were not our primary focus, they offer valuable analytics and KPIs to help city or business managers optimise mobility. As part of their value proposition, they provide managers with interesting analytics and KPIs to help them manage their city or business. In the case of Ride Report, they also maintain an open data portal that compares bike and scooter-sharing demand across different cities.

A common pattern observed across these platforms is the lack of a platform strategy that enables the development of new mobility models. Typically, these platforms limit users or managers to a fixed set of analytic capabilities, such as analysing cyclist volumes, origin-destination patterns, or user behaviour. Despite having access to vast amounts of data, their restrictive design does not allow users to create new mobility models, thus preventing them from leveraging the available data.

One solution that targeted this issue was MoocRP, an open-source analytics platform designed for MOOCs. Unlike other platforms, this platform allows researchers to access multiple datasets, develop and share new models and visualisations, or use the models developed by others. While developing a workflow, the output of one model can also serve as the input for another. A similar approach could be adopted in the context of cycling analytics. By enabling the development of new models on top of existing ones and leveraging available datasets, the platform could facilitate a continuous cycle of innovation, ultimately

leading to more advanced and comprehensive mobility solutions.

2.4 Cyclists' Route Preferences

Research on cyclist route preferences focuses on capturing cyclists' perceptions about the road network characteristics, quantifying the most valued attributes, and assessing how much cyclists are willing to detour for preferred conditions. Research has shown that bicycle route choice involves a joint consideration of convenience, safety and leisure (Chen et al., 2018).

Positive Attributes

Generally, cyclists are found to prefer shorter and flatter routes, with fewer intersections per kilometre, and ride on buffered or separated bicycle facilities with good street lighting and slower road traffic (Prato et al., 2018; Chen et al., 2018; Ton et al., 2017; Menghini et al., 2010; Caulfield et al., 2012; Li et al., 2019).

Regardless of their confidence, cyclists strongly prefer bike lanes to other bicycle facilities (Caulfield et al., 2012; Hood et al., 2011). Those dedicated infrastructures positively affect the feeling of a reduced distance. Prato et al. (2018) found that a route is perceived as 24.9% shorter when cyclists ride on bike lanes and bike paths during peak hours. Broach et al. (2012) found that travelling on bike paths is equivalent to reducing the distance by 26.0% for non-commute trips (16.0% for commute trips), while travel on bike boulevards is equivalent to decreasing distance by almost 17.9% (10.8% for commute trips).

Cyclists are also found to place a relatively high value on enhanced neighbourhood bikeways with traffic-calming features (bike boulevards) and bridge facilities (Broach et al., 2012), and show some interest in routes through residential streets (Caulfield et al., 2012).

Regarding the environment, cyclists prefer scenic areas (e.g., parks, forests, and coastal locations) over high-rise urban environments, especially on long trips (Łukawska et al., 2023). They perceive these routes as shorter, but only depending on the air temperature (Prato et al., 2018).

Negative Attributes

The attributes that negatively impact route choice include trip length, excessive deviations from the shortest paths, steeper slopes (especially above 6%), longer trips without bike paths, unsignalled left turns, intersection control, interactions with traffic or pedestrians, and high adjacent traffic volumes (Hood et al., 2011; Broach et al., 2012; Li et al., 2019).

Broach et al. (2012) found that a 1% increase in route distance reduces the likelihood of choosing a route by about 5% for non-commute trips (9% for commutes). Additionally, cyclists are more concerned with relative rather than absolute route deviation. This means that, for example, a 200-meter increase in distance will have a bigger impact on a 1 km route than on a 10 km route. Trip time negatively impacts the probability of car and public transportation users changing to a bicycle to commute de Magalhães and Rigatto (2025).

Cyclists also prioritise avoiding steeper inclines, focusing more on the maximum gradient than the average gradient (Menghini et al., 2010). As the slope increases, so does the willingness of cyclists to take longer detours (Broach et al., 2012).

Regarding signage, cyclists normally avoid stop signs and traffic signals unless these signs are needed for crossing or to turn left at high-traffic streets (Broach et al., 2012). To avoid unsignalled left turns, cyclists may travel about 16.2% farther (9% for commutes) (Broach et al., 2012).

High crowding levels are also found to negatively affect the route preference and the perceived safety (Uijtdewilligen et al., 2024). Between 85% and 91% of the cyclists feel safer on separated bicycle tracks with low-to-moderate crowding. However, this feeling drops to 44%-62% at high crowding levels, particularly affecting older cyclists and women. Shared paths and dedicated footways are perceived to be 16% and 51% longer, respectively (Łukawska et al., 2023). Regarding on-street car parking characteristics, all cyclists prefer no parking to any form of parking on their route (Sener et al., 2009), especially for long commuting trips. All bicyclists except young adults prefer angled parking to parallel parking, as the latter configuration can lead to dooring problems.

Cyclists also avoid unpaved roads (Prato et al., 2018) and stairs (Łukawska et al., 2023). They seem to prefer roads built using asphalt rather than cobblestone (Hardinghaus and Weschke, 2022). Concerns about bicycle theft increase the emphasis on travel time and reduce the overall appeal of cycling (Márquez and Soto, 2021).

Cyclists typically will not go the wrong way down a one-way street without considerable savings in effort (Hood et al., 2011). The perceived distance is 2.4 times longer than the one in the correct direction. However, some cyclists have a positive preference for shortcuts involving being in the wrong direction (Prato et al., 2018). Other studies report that counterflow is perceived as 29% and 35% longer for short and long trips, respectively (Łukawska et al., 2023).

In cases where the perceived safety of the bike lanes is compromised, cyclists may reject them and prefer to use other infrastructures (Duc-Nghiem et al., 2020). This may be due to high vehicle volume, vehicles stopping on the bike lanes, narrow traffic lanes, narrow bike lanes, and wider sidewalks. The

latter may attract more cyclists and reduce bike lane usage.

Preferences heterogeneity

However, despite general trends, route choice preferences vary significantly (Prato et al., 2018) due to multiple factors such as gender, age, experience, riding with or without children, bike type, past accidents, the city cycling culture, trip purpose, and time of day (Hardinghaus and Weschke, 2022; Uijtdewilligen et al., 2024; Márquez and Soto, 2021). People have different perceptions of what it means for a path to be safe (Heinen et al., 2010). Even the same road may not appear safe or accessible for everyone. A route with 3% gradient in a small segment may look healthy for some commuters, but an obstacle for others (de Magalhães and Rigatto, 2025).

Infrequent cyclists, females, and people cycling with children have a stronger preference for bike paths and protected bike lanes than males (Poudel and Singleton, 2022; Hardinghaus and Weschke, 2022). For the people cycling with children group, preferences for living streets are nearly tripled compared to the others (Hardinghaus and Weschke, 2022). On the contrary, confident cyclists had significantly weaker preference for protected facilities (Poudel and Singleton, 2022). In Amsterdam, cyclists are even insensitive to separate cycle paths (Ton et al., 2017), presumably due to the guidelines for Dutch infrastructure or due to the location of the cycle paths in the network.

Infrequent cyclists feel less safe on large roads than experienced cyclists. However, this feeling can be mitigated with protected bicycle tracks (Łukawska et al., 2023). Cyclists have no stronger preference for painted lanes over roads with no infrastructure, as painted bicycle lanes alone do not provide sufficient protection on large roads.

Risk perception increases with age (Márquez and Soto, 2021). Young cyclists are less concerned about safety and are more likely to belong to a segment with a lower fear of traffic. Compared to younger cyclists, older groups have a lower perceived safety of high levels of crowding (Uijtdewilligen et al., 2024) and have a higher preference for lower speed limits for cars and side streets (Hardinghaus and Weschke, 2022).

People who have experienced collisions or near misses are less deterred by a 35 mph ($\approx$56.3 km/h) speed limit over a 25 mph ($\approx$40.2 km/h) speed limit and by medium or high traffic volumes over low traffic volumes (Poudel and Singleton, 2022).

Recreational cyclists prefer bicycle routes with attractions along the way, basic facilities, tourist information centres, and bike paths (Chen and Chen, 2013). Cyclists with high levels of recreational specialisation are more likely to choose longer and more challenging routes than cyclists with low recreational

specialisation (Chen and Chen, 2013).

Data Sources and Methodologies

These studies mainly rely on two main data sources: either surveys where cyclists state their preferences (Sener et al., 2009; Caulfield et al., 2012; Márquez and Soto, 2021), or GPS traces with cycling activity (Hood et al., 2011; Broach et al., 2012; Ton et al., 2017). While some authors employ face-to-face questionnaires to collect their data (Márquez and Soto, 2021), others rely on web-based surveys (van der Waerden et al., 2024; Sener et al., 2009). The latter is considered inexpensive to the researcher in terms of disseminating information about the survey, easier for respondents to answer, and environmentally friendly (Sener et al., 2009). Comparing GPS traces with survey data, Prato et al. (2018) argues that, while collecting stated preferences data has the advantage of controlling the experimental environment, it does not observe actual behaviour and that large-scale revealed preferences data (e.g., using GPS data) is preferred because it captures cyclists' actual preferences and surpasses possible technical problems related to tracing and map-matching. The GPS trace data is also usually enriched with other information, including weather, trip purpose, and demographic or personal information (Hood et al., 2011; Broach et al., 2012). Other complementary data sources include DEM data (Broach et al., 2012), road network data provided by regional planning organisations (Broach et al., 2012), and transport supply (Huber et al., 2024).

Regarding methods, the authors employ a plethora of techniques with different purposes, including map-matching algorithms to allocate the GPS points to the street network (Hood et al., 2011; Menghini et al., 2010; Ton et al., 2017), algorithms for mode detection (Menghini et al., 2010), such as algorithms based on fuzzy logic methods (Hood et al., 2011), multi-agent transport simulation to generate simulated origin–destination pairs (Menghini et al., 2010), Multinomial Mixed Logit model for investigating random taste variation (van der Waerden et al., 2024), the multinomial logistic regression model for route choice analysis. (Huber et al., 2024), or the k-means clustering method (Ton et al., 2017).

2.5 OSM Road Network Completeness

The popularity and widespread use of OSM have led many authors to assess the quality and completeness of OSM data. Typically, the methods used can be classified as intrinsic or extrinsic (Vierø et al., 2024b). Intrinsic methods focus on internal qualities of OSM, such as analysing the edit history, contributors, or tag completeness, while extrinsic methods compare OSM with other datasets.

Overall, most studies have shown that OSM data has good quality. In some cases, it is even more complete than municipal data (Ferster et al., 2020) as informal paths may be mapped in OSM, and new infrastructures are often mapped in OSM before they are released on datasets maintained by cities (Nelson et al., 2021).

Hochmair et al. (2015) compared the OSM road network with data for US and European cities and found that the OSM data has relatively good quality and particularly high quality for designated lanes. While Amsterdam and Berlin have a large network of bicycle trails, the US cities have a higher density of designated lanes.

Barrington-Leigh and Millard-Ball (2017) compared the OSM road network with aerial imagery and its historical growth and concluded that, globally, OSM is 83% complete, and in 40% of the countries it covers more than 95% of the roads. Therefore, the authors suggested that OSM already is or will soon be a reliable tool for researchers and policy-makers.

In another study, Ferster et al. (2020) compared OSM with reference data for six Canadian cities and found relatively good concordance in the overall length of bicycle infrastructure between OSM and city-provided data. More specifically, the authors found a very high concordance in two cities and reasonably high concordance in the remaining four. However, this concordance varied depending on the tags. Roads tagged as "on-street bicycle lanes" were the most concordant, while cycle tracks and local street bikeways were the least concordant.

With a different approach, Graser et al. (2015) assessed the quality of the OSM road network for performing vehicle routing. The authors compared the OSM road network with the Austrian reference graph and found a close alignment between the one-way street and turn restriction information. Namely, 95.4 % of the reference graph one-way street restrictions were found to be represented in the OSM road network. Additionally, with a median absolute length difference of 1.0%, the authors stated that OSM shows promising capabilities for vehicle routing applications.

Wasserman et al. (2019) assessed the performance of OSM to compute Level of Traffic Stress (**LTS**), a prominent metric to measure the facilities' attractiveness for cycling. The authors compared OSM-derived LTS predictions with ground-truth LTS scores and found high concordance between both LTS scores. The results showed that 89.9% of the length of the network was correctly identified as either high or low stress. However, some street typologies and urban contexts were more prone to errors. This included areas that might be under-represented in tag completeness, such as suburban or rural locations, and in denser areas, where street typologies might be more complex and potentially misrepresented in OSM (Wasserman et al., 2019). This contradicts Barrington-Leigh and Millard-Ball (2017), which states that completeness has a

U-shaped relationship with population density. Both sparsely populated areas and dense cities are the best mapped, as the densest cities are home to many potential mappers, while the least dense areas are easier to trace.

With opposing results, Vierø et al. (2024b) assessed the quality of OSM and GeoDanmark (an open bicycle network dataset) for network-based analysis of cycling conditions and stated that neither dataset is of sufficient quality and that dataset merging is necessary to obtain a complete dataset.

Hochmair et al. (2015) identified two types of errors, namely omission and commission. Omission errors occur when a cycle lane is not represented in the road network or does not include proper tags. Commission errors occur when non-existent roads are represented in the OSM database, the geometry is incorrectly defined, or it includes tags indicating permission for cycling. As the OSM database is dependent on crowd-sourced contributions, a key challenge is to achieve consistent OSM tagging for different bicycle infrastructure types: people may lack training in geography or surveying, employ diverse labelling practices, and have different interpretations of the tags. As a consequence, their contributions can be inaccurate, incomplete, inconsistent, or vague (Jackson and Zhang, 2016; Nelson et al., 2021; Ferster et al., 2020)

To improve the reliability of OSM data, common recommendations for practitioners include: the validation of OSM data against reference sources, such as using Google Maps as a potential source to improve the OSM database (Hochmair et al., 2015); combining multiple data sources to achieve comprehensive coverage (Vierø et al., 2024b); implementing regular quality monitoring techniques (Bres et al., 2023); and the use of specialised tools to conduct bicycle network analysis, such as tool similar to the BikeDNA (Vierø et al., 2024a). Despite its issues, Bres et al. (2023) argue that, with its large and dynamic community of contributors, OSM can become a regularly updated source of geographic data on cycling networks.

Quality Assessment Tools

Addressing the OSM data quality research topic, several researchers and other interested developers have proposed and published tools to detect errors in the OSM road network and help improve its quality.

An example of such a tool is OSMRunner (Tabet et al., 2021), designed to automatically detect connectivity errors. OSMRunner scans the entire road network and, for each road, it imposes condition checks that enforce the connectivity restrictions. For the problems found, it suggests to users possible fixes to achieve full connectivity.

Another example, the BikeDNA tool, developed by Vierø et al. (2024a), performs either a standalone analysis of one data set or a comparative analysis between OpenStreetMap and a reference data set. It considers data quality metrics, such as edge density, missing tags, or disconnected components, both

globally and locally on a grid cell level, to expose spatial variations in data quality. The results are summarised in PDF/HTML reports.

The "Is OSM up-to-date?" is an open source web application to assess OSM intrinsic quality based on the object history (Minghini and Frassinelli, 2019). The rationale is that nodes not modified for a long time have a higher probability of having outdated information. The authors considered five quality criteria, namely date of creation, date of last edit, number of versions, number of different contributors and update frequency.

MapRoulette (van Exel, 2012) is a web-based micro-tasking platform. It has a database with several mapping challenges, each divided into several smaller tasks. People interested in correcting OSM errors can select a challenge and start mapping. It provides a leaderboard ranking the monthly, yearly and all-time top contributors.

JOSM (Contributors, 2005a) is a desktop application for editing OpenStreetMap. It supports loading GPX tracks, background imagery, and OSM data from local and online sources to help users map the world by creating or editing existing nodes, ways and relations (Contributors, 2005b). JOSM Validator (Contributors, 2007) is a core feature of JOSM, which checks and fixes invalid data, such as duplicated nodes, incomplete ways, crossing ways, and other geometry and tagging errors. Errors detected by the Validator can be categorised as errors, warnings and other depending on their severity. While some errors must be manually fixed, others can be automatically fixed (Contributors, 2007).

2.6 OSM Representation Politics

Volunteered geographic information (**VGI**) is a collaborative approach based on digital tools to collect, analyse, and share geographic information (Engler et al., 2014; Ferster et al., 2018). This includes a variety of activities, from unintentional actions (e.g., georeferencing pictures shared on the internet) to involvement in scientific research (Ferster et al., 2018). VGI data collection and sharing, conceptualised as bottom-up processes, enable individuals to document their realities and express their needs to municipalities and can take a leading role in providing spatial information to planning efforts (Hachmann et al., 2018).

Crowdsourcing mapping, a specific type of VGI, can be an alternative to conventional mapping, challenging the hegemony of official and commercial cartographies (Perkins, 2014). Collaborative mapping platforms, such as OSM, allow local communities to become part of the power structures and influence what and who is on the map (Panek and Netek, 2019). They offer a forum for different voices, enabling new ways of economic and societal development and new ways of living (Perkins, 2014; Panek and Netek,

2019). These platforms offer the possibility of literally putting communities on the map by gathering information on uncharted or poorly represented areas, such as informal settlements or developing countries (Hachmann et al., 2018; Panek and Netek, 2019).

Cartography as political and heavily influenced by socio-political habits

However, concerns about data quality and incomplete representation are the two main barriers for VGI adoption (Ferster et al., 2018). Additionally, crowdsourced mapping may reify power relations (Perkins, 2014). Cartography is deeply political, and the cartographic gaze consists of multiple experiences from multiple mappers with numerous purposes (Wilmott, 2019). Their mapping practices are heavily influenced by personal preferences, knowledge, and habits (Glasze and Perkins, 2015), an entanglement between people, tools, landscapes, and cultures (Wilmott, 2019). As such, crowdsourced data products, such as OSM, are the sum of individual contributions with varied personal and institutional interests (Quinn and MacEachren, 2018).

These situated knowledges impact both the drawing and the reading of the map. In the case of OSM, while community conventions may provide general guidelines on how to conceptualise and use tags, their understanding during use and reading is affected by important cultural, social, economic, historical, and governmental distinctions (Wilmott, 2019).

Mapping Tensions, Edit Wars, and Conflict Resolution

The entanglement of these factors may give rise to mapping tensions or disputes. OSM Mappers engage in 'edit wars' when consensus is not achieved regarding the appropriate tags or the existence of certain features (Contributors, 2023). These are especially frequent in contentious areas, such as Cyprus, Jerusalem, or Crimea, where different place names and borderlines have been recorded and overwritten (Glasze and Perkins, 2015). Another source of conflict in OSM lies in how to render its information. For example, until 2008, the most important renderers translated all places of worship as a cross, which triggered several critical statements in the OSM discussion lists (Glasze and Perkins, 2015).

OSM communities attempt to resolve those disputes using various techniques and publish the results in the OSM Wiki (Glasze and Perkins, 2015). As an example, the engagement of commercial actors in OSM raised concerns by local mapping communities that OSM's diverse voices could be marginalised, that OSM data quality could be compromised, and that corporations would fail to disclose their mapping activities or introduce biases (Schröder-Bergen et al., 2022). Based on the discussion within the community, the

OSM Foundation proposed the "Organised Editing Guidelines"[3] (**OEG**), a document summarising the expectations, consensus, and established conventions to govern how corporate mapping should occur (Chauhan et al., 2024). Nowadays, these guidelines continue to be referenced in discussions and may be subject to modifications as new use cases and conflicts arise, new mappers join the community, and underlying values evolve (Chauhan et al., 2024). For example, when the OEG was proposed, local mappers were concerned that it could diminish their autonomy by placing excessive requirements on the existing OSM community.

While the 'on the ground' mapping rule is suggested as the default route in the OSM Wiki to solve disputes (Contributors, 2023), this rule tends to reproduce traditional cartographic patterns, favouring concrete and other physical structures over use and meaning (Glasze and Perkins, 2015). As a consequence, less prominent structures (e.g., "backyard mosques") tend to wait to be integrated into OSM. Additionally, this rule reaffirms the present spatio-social order and thus the power relations behind it (Bittner, 2017b). For the Palestinian case, "many aspects of the factual material space might appear not as neutral physical objects but as results of suppression, in which case any 'accurate' spatial representation, such as OSM, becomes objectionable" (Bittner, 2017b).

It is also found that VGI tools tend to reproduce specific community-based biases and do not unambiguously reproduce social inequalities (Bittner, 2017a). For example, while OSM data in Jerusalem tends to be richer in Jewish/Israeli areas than in Arab/Palestinian areas, the opposite is true for Wikimapia. This may be due to OSM being dominated by Israeli and international mappers, whereas Palestinians have hardly contributed to OSM (Bittner, 2017b).

Tensions as an enabler of new politics

The efforts to define clear categorisation systems do important political work in revealing tensions, different perspectives, and the difficulty of generalising from situated knowledge (Wilmott, 2019). Multiple realities overlapping or coexisting in contradiction, such as the differences between a top-down governmental understanding of the territory and a situated and local sense of lived experience, can reveal insights about how the land is used versus how it is owned (Wilmott, 2019).

As such, the wikification of mapping might facilitate a more mutable politics (Perkins, 2014). The tensions produced through processes of cartographic classification offer new pathways for engaging mapping as a political and world-building tool and may open up alternative ways of thinking about spaces, politics, landscapes, and environments (Wilmott, 2019).

[3] https://osmfoundation.org/wiki/Organised_Editing_Guidelines

2.7 Map-Matching

Map-matching is a key process in modelling mobility. Given a road network model and a GPS trajectory of a vehicle, it tries to find the most probable sequence of road network edges on which a vehicle has travelled (Yang and Gidófalvi, 2018).

However, applying map-matching for bicycle trips is particularly challenging, as cyclists frequently use paths that are not designed for cycling or that might not be included in the road network data, such as parks or shortcuts (Berjisian and Bigazzi, 2023; Schweizer et al., 2016). The built environment also strongly influences the performance of map-matching algorithms (Trogh et al., 2022). In dense urban areas, characterised by tall buildings and a complex road network with multiple parallel roads, the probability of having map-matching anomalies is higher compared with zones with a sparser road network.

Map-Matching applied to cycling

Several authors targeted these challenges and proposed new map-matching algorithms tuned for cycling mobility.

Bergman and Oksanen (2016) proposed a method based on the Hidden Markov Model, which favoured bikeways extracted from OSM. Schweizer et al. (2016) proposed a buffer-based map-matching algorithm that maximises the likelihood that a route is identical to the real route. Their method uses network attributes to estimate the route in case of incomplete GPS data and can identify if cyclists used a reserved bikeway, where available. Trogh et al. (2022) proposed a map-matching algorithm that supports trajectories on foot, by bike, and by motorised vehicles. Depending on the transportation mode, some roads are discarded based on their tags. For example, highway roads are discarded when map-matching a trace made by a cyclist. Gao et al. (2024) improved the pgMapMatch algorithm, tailoring it to cycling in urban environments. For this, they introduced a classification of road availability for cyclists and extended the multi-objective map-matching scoring system.

Map-Matching with incomplete road network data

Other authors acknowledge the incompleteness of road network data and propose approaches capable of performing off-road map-matching or detecting various categories of road network errors (Murphy et al., 2019; Behr et al., 2021; Yang et al., 2024).

Murphy et al. (2019) proposed a map-matching algorithm for on- and off-road tracking. Their method can switch between the two modes as much as necessary. It uses a standard Hidden Markov Model for

on-road tracking and a closed-form Kalman Filter for free-space tracking. Off-road trajectory portions are generated as a fallback when the road cannot accommodate the observed vehicle motion.

Considering that pedestrians and cyclists can go through omitted roads and open areas, Behr et al. (2021) proposed an approach that supports map-matching of trajectories that possibly contain on- and off-road sections. They extend the road network data model by triangulating all open areas and adding tessellation edges. They consider each GPS point as a possible matching candidate, and the unmatched candidates are added as nodes in the road network graph if no similar path is found in the road network.

The off-road sections can then be used to interpolate missing road segments and to update the road network (Shan et al., 2015; Sasaki et al., 2019). With a similar objective, Yang et al. (2024) combined partial map-matching (**PMM**) with a bidirectional recurrent neural network (**BiRNN**) model to detect various categories of road network errors from trajectory data. First, they apply PMM to build associations between the matched part and the road network. Then, they use the map-matching outputs, raw trajectories, and road network to extract context features. Finally, apply the BiRNN model to perform sequential classification and obtain point labels.

Frequently, the datasets used to test these algorithms consist of trajectories made by motorised vehicles, such as taxi trips. The authors assume that, since vehicles move on roads, vehicle trajectories are usually restricted to the road network and unmatched trajectories can be used to find new roads (Tang et al., 2019). Although this approach is reliable for cars, the paths made by cyclists may not be eligible to be represented in a road network, as we could be updating the road network with illegal roads or connections.

Map-Matching Error Detection

In a different approach, some authors have tried to identify map-matching errors.

Berjisian and Bigazzi (2023) proposed a method to flag map-matching results requiring visual inspection. The error indicator is obtained by combining the normalised values of several similarity measures between the GPS trace and the resulting map-matched trace. Cases where the indicator is above 0.5 are flagged as potentially erroneous.

In another study, Dey et al. (2022) proposed a method comprising two phases. In the first phase, an unsupervised learning model classifies each GPS point as good or bad based on its distance to the map-matched segment and estimated GNSS error. Then, the map-matched segments are voted as good or bad, based on the majority of points associated with them. In the second phase, they detect unrealistic travel behaviour using the "edit distance".

2.8 Summary

The review of the bikeability and bike-friendliness indexes highlighted an existing diversity in the methodologies: They are composed of varied indicators, rely on data from diverse data sources, and are presented in different ways. This flexibility offers researchers the possibility to expand their creativity and evaluate several characteristics of a city. However, most of these processes are self-contained, based on a specific dataset and usually applied to a single city. Whenever someone wants to test a model, most of the times they will not have access to the source code or even the data formats, and they are forced to repeat the whole process from the beginning. This makes these models difficult to generalise, reproduce, and repurpose to other scenarios or datasets, if not impossible. As for the analytics platforms, most approaches are focused on a specific data source, such as a mobile application, or compute a limited set of indicators, such as origin-destination or distance travelled. Other platforms offer more features: they characterise the network infrastructure and compute optimal routes between two points based on specific preferences. However, no cycling analytics platform allows users to upload their datasets and make full use of the platform's models to gain insights about their city.

This supports our view that the full potential of cycling analytics as a transformational force for urban cycling remains largely untapped. Regarding multiple initiatives, currently, no model is seen as a shared reference to drive decision-making. This problem may not be related to the merit of the models but rather to their development paradigm, leading to monolithic and closed processes. The current paradigm does not allow for incremental innovation of new, more complex models that better describe the cycling reality. Although a huge amount of data is available, the models cannot be adapted to leverage existing datasets. This is a major research gap that reduces the value of both existing datasets and mobility models.

The lack of actionable and comparable metrics presents a huge problem for urban planners who need data to guide their investments to meet users' needs. This consequently affects cyclists and people who want to adopt the bicycle as their usual means of transportation by not giving them the ideal conditions to feel comfortable and safe travelling on their bicycles. A new mobility model development paradigm is thus mandatory to allow the development of new models that build on top of the existing ones and aggregate different data sources.

The literature on OSM completeness revealed that OSM has a strong potential to become an enabling technology for cycling services. OSM's road network data is complete in many countries, and changes in the road network infrastructure are often reflected sooner in OSM than in official datasets. Additionally, several authors found that OSM data has sufficient quality to support many mobility-related applications,

such as vehicle routing or level of traffic stress metric computation.

However, the review of cyclists' route preferences shed light on the cycling route choice process and revealed that cyclists have different views of what they consider cycleable routes. Despite general trends, such as cyclists preferring shorter and direct routes and avoiding steep slopes or high crowding levels, there is also heterogeneity in route choice preferences. Different cyclists consider factors differently, and even the same cyclist may consider the same factors differently when facing distinct scenarios.

These diverse behaviours pose an enormous challenge for information systems designed to model cycling behaviour or provide optimal routes. In this context, for OSM to become an enabling technology for cycling services and support cycling services for every cyclist profile, it must accommodate this heterogeneity. Moreover, the challenge of representing real cyclist routes can be exacerbated by conflicting mapping practices, as mappers may have different interpretations of tags or how to represent cycling-related elements. These conflicting perspectives may even culminate in mapping tensions, which can reveal the differences between how the land is owned and how it is used. In the case of cycling, they can reveal how public authorities and cyclists perceive the road network infrastructure by revealing the differences between the designed and the actual routes.

Map-matching algorithms can be a valuable tool for revealing the gaps between mapping practices and cycling activity. Map-matching snaps GPS traces of cycling activity into a road network model, allowing us to create an aggregate perspective of cycling in a city. By identifying the gaps between the GPS traces and the map-matched routes, we can reveal the paths used by cyclists that have low or no representation in the OSM road network.

Chapter 3

Cycling Analytics Models Decomposition

3.1 Motivation

In this chapter, we aim to assess the existence of generic functionality across some of those models and characterise the corresponding opportunity for horizontal services that could facilitate their development and convergence. We also aim to make an initial analysis of how a digital platform strategy could provide a more promising approach to address the problem and bring many of those models into a common system with added value for different cycling stakeholders.

3.2 Methodology

The methodology of the work described in this chapter includes two distinct phases. In the first phase, to validate the existence of generic functionality across multiple models and identify concrete opportunities for convergence around shared services, we analysed 3 commonly used cycling mobility models. The approach is to decompose those models into a set of atomic code blocks and analyse the existence of shared generic operations. Each code block should correspond to a sequence of related instructions aiming at a common action that would not make sense to execute partially, in this or other contexts. We then analysed each of these atomic blocks to assess their potential for generalisation or integration. A high level of shared generic operations should indicate a high potential for the creation of horizontal services that could significantly lower the effort needed to develop those model implementations.

In the second phase, we conducted an initial analysis of how digital technology platforms may provide a more promising path for addressing the mobility models' current problem. Our approach is to analyse how the general principles of digital platforms, combined with the results from the study of this concrete problem domain, can provide a foundation for a platform-based approach to cycling analytics. In the process, we also expect to reach an initial perspective on what might be some of the expected properties

of such a platform.

3.2.1 Model Selection

To accomplish our goals, we started by selecting 3 cycling mobility models to be analysed and decomposed into smaller atomic blocks. We selected models developed as part of research initiatives, for which we have access to the source code. Two of those models are published on GitHub and have several recommendations stars. The third model was developed by our research team and includes procedures for describing cycling traffic at each road. These 3 models are relatively simple in their algorithms, but address common and yet very diverse data processing goals in the cycling domain. We selected a small number because the aim was not to consider the complete range of possibilities, but only to have a diverse sample to assess some of the common operations that compose the algorithms that create those models. Table 3 presents the 3 models used as input for our study.

Table 3: Mobility models analysed in this study.

Model	Description
Hilliness (lab, 2020)	This model takes a Digital Elevation Model and a set of Street Network Data and classifies network segments according to the respective cycling effort associated with street slope.
Accessibility (Vale, 2020)	This model takes as input an origin-destination matrix with travel costs and a list of opportunities at destinations and classifies city areas according to their accessibility. Travel costs can be very diverse, e.g. travel time, distance, slope or risk.
Road Characterisation [Own model]	This model takes a set of cycling traces and a scope of the OSM road network and classifies network segments according to aggregate properties, e.g. average speeds, acceleration markers, stops, turns, and volume.

3.3 Results

Table 4 presents the results from the analysis of the 3 models. Namely, Table 4 is a selection of the atomic code blocks which we identified as having the potential to be offered to many different models as common horizontal features in a digital platform. These algorithms are often created for a specific purpose and,

even if they often include many library functions, they also have strong dependencies associated with the concrete assumptions of a particular implementation.

Table 4: Possible horizontal features identified in the analysis and their dependencies.

Operation	**Analysis of possible dependencies**
Obtain road network data from OSM	This requires domain knowledge about the usage of the OSM API. There is an implicit spatial scope that makes this operation specific to the implementation. The output data is generic.
Obtain shapefiles for region boundaries	This requires domain knowledge about the availability and integration of region boundaries. There is an implicit spatial scope embedded in some operations. Output data is generic.
Annotate the road network with slope information	This requires domain knowledge on how to obtain and apply elevation data for a concrete region. There is an implicit spatial scope embedded in some operations. The output data is generic.
Classify a road network according to a cycling effort model	This requires domain knowledge on classification models for cycling effort. There is an implicit spatial scope embedded in some operations. The output data is generic.
Estimate accessibility from the Origin-Destination (**OD**) matrix and opportunities	This requires domain knowledge on classification models for cycling accessibility. There is an implicit spatial scope embedded in some operations.
Map-matching cycling traces into OSM	This requires domain knowledge on map-matching methods.
Compute average speed for road network segments	Code can be reused with any compatible trace data format, but the ability to aggregate multiple traces is also essential.

The first observation is the high number of procedures that are not specific to each of these models and could potentially be offered as horizontal features on a digital platform to serve similar needs across multiple cycling models. From the analysis of merely 3 models, we were able to identify 7 atomic blocks corresponding to operations that other models may also need to execute in very similar ways. These results are strongly aligned with our initial hypothesis about the existence of generic functionality across many of these mobility models.

Table 4 also represents the diversity of dependencies that we found in the analysis of these 3 models, and which may prevent their re-usage outside their initial implementation context. A very common de-

pendency is the specific domain knowledge embedded across these implementations. This knowledge is often present in the usage of specific enabling services or data sources needed by those models, e.g. map-matching, application of elevation data, mapping locations to administrative regions and their respective boundaries or working with OSM. None of these is necessarily very complex, per se. However, having to learn about the correct usage of multiple services, all of which are just enablers for the concrete purpose of the model being developed, may represent a huge barrier to the development of new models or the usage of existing ones, even for users with general knowledge on mobility systems. Another common form of dependency is related to the implicit scope assumptions, and particularly those referring to the spatial context of the work. For example, by focusing the model on a city, the selection of spatial data and the scope of the visualisations created can all be easily aligned. While some of these dependencies may be obvious from the implementation and potentially easy to change, many others are more deeply embedded in the code and severely limit the applicability of the same implementation to other spatial contexts.

There are also implications that can be linked directly to the self-contained nature of these models. An important group of operations involved the association with various types of geographic data, e.g. obtaining OSM road networks, obtaining elevation data to determine road segment slope, mapping traces into the OSM road network or postal codes. These are not only very generic operations, but they are also operations that could be done only once to serve the needs of many models. However, since each model is self-contained, it needs to independently obtain all of its data and repeat the process for each execution. In a platform approach, this type of global reference data could be managed as a horizontal feature to be shared across the entire system and made available to multiple models. Likewise, a shared service would also provide a much-needed aggregation context where data from multiple models could be treated in an integrated way and used to generate significant new value from the existing data. In particular, geospatial references are the ultimate convergence mechanism, allowing spatial relationships encoded in source data to support the aggregation of multiple independent data under a common spatial reference, and creating an immediate association between data that refers to the same region, postal code, road or road point.

There are also commonly used functions that operate on a specific dataset. Typical examples may include converting between data representation formats, calculating metadata for a given trace, or calculating the distance between two points in a road network. The existence of these functions saves model creators from additional programming effort and promotes more robust solutions. While we can expect many of these operations to become part of a digital platform, this form of generalisation is already supported by a multitude of code libraries available from multiple sources. Since their execution is a self-contained process, similar to a microservice paradigm, they do not have external dependencies. Still, a coordinated

library could play an important role in promoting convergence towards common approaches that improve the transparency of the models that use them and potentiate convergence towards common system approaches and data formats. It should also reduce the complexity involved in selecting and learning how to apply the functions from new libraries. A final observation refers to the huge diversity of data types and formats that are involved in all of these models. Regardless of the obvious benefits that may emerge from more standard approaches to data publication, a cycling analytics platform should be designed to accommodate this extreme diversity, offering multiple convergence mechanisms to absorb that heterogeneity through various types of data transformations, rather than trying to impose its own data formats.

3.4 Key Properties for a Platform Strategy for Cycling Analytics

In this section, we make an initial analysis of how a digital platform strategy could provide a more promising approach to address the challenges of urban cycling analytics. Considering that this is still largely an emerging domain, it may be too early yet to fully comprehend the nature of this new opportunity space and the concrete details of what these smart cycling ecosystems will look like. It will take some time and many cumulative experiments before we can reach a stage where a more grounded analysis of those properties will become possible.

At this stage, we can build on the results of our study and also on the general principles of platform innovation (Yoo et al., 2012; Zittrain, 2005; Ciriello et al., 2018) to make an initial analysis of the key properties that may characterise a platform strategy to uncover the potential of digital innovation in urban cycling analytics. These properties include a platform with a well-defined scope, multi-sided, and with global coverage, based on the principles of distributed and combinatorial innovation, enabling both convergence and generativity.

Cycling specific: A common approach of many platforms is to target a relatively narrow market in order to gain traction (Edelman, 2015). A platform must clearly define its boundaries and target audience to attract the right users. In our context, the platform should focus specifically on micro-mobility needs. While sharing many principles with more mature forms of mobility analysis, e.g. public transportation or automotive, cycling is substantially different regarding the nature of the data sources, the nature of road classifications, the key performance indicators or the role of real-time data. Focusing on the specifics of micro-mobility means placing those core needs at the forefront of the analytics systems. Still, cycling

should not be seen as an isolated reality in urban mobility, and the platform should always consider its own integration with broader mobility systems, particularly from a Mobility as a Service (**MaaS**) perspective (Jittrapirom et al., 2017).

A multi-sided platform for the cycling ecosystem: A basic premise for any Digital Platform is the ability to congregate many users and very different types of stakeholders. A multi-sided platform generates value by facilitating direct interactions between two or more sides (Leong et al., 2024). The network effects are central in their value proposition and can trigger a self-reinforcing virtuous cycle (Acs et al., 2021): as the number of users of one group increases, so does the value offered by the platform to that group (direct network effect) or to the opposite side (indirect network effects). As the platform's value grows, so does its ability to attract new users. In our context, cycling communities should find in this platform the ideal environment for a broad range of collective activities that will significantly expand the cycling experience. However, this form of convergence should also bring together many other types of stakeholders, such as bike sharing operators, bike shops, city authorities, mobility authorities, bike-friendly businesses, citizens in general, public parking facilities, bike manufacturers, and digital cycling technology companies. By significantly lowering the barriers for interaction, a platform should help to blur the boundaries between these many separate domains and unlock new forms of value exchange that will benefit all the stakeholders involved.

A global service for many local contexts: When users are used to a service, they expect it to be available anywhere they need it with a similar user experience. Moreover, the need for cycling services is a global need. Therefore, the proposed platform should offer a broad set of tools, services, data and algorithms to address the cycling challenges of any city in the world, regardless of its characteristics or level of cycling readiness. Also, the scale associated with a global service is essential for reaching the critical mass needed for the platform to be perceived by others as a catalyst for convergence. However, this global perspective should be counterbalanced by the need to connect to local cycling ecosystems, which constitute a major source for locativeness and generativity-based innovation. This may involve the emergence of local cycling groups, access to detailed information about the local infrastructures and regulations, the opportunity to participate in the co-creation of local mobility policies or specific MaaS integrations.

Collective Intelligence: As a result of pervasive information technologies, innovation processes became democratised and distributed, with new products or services becoming the result of innovative pro-

cesses involving many people and organisations (Yoo et al., 2012; Ciriello et al., 2018). The term collective intelligence in a platform ecosystem refers to the process in which diverse groups of people cooperate and combine their knowledge to generate ideas and solve problems, which otherwise would not be solved (Elia et al., 2020). In this context, the platform should have the capability to observe and learn about cycling mobility in a way that supports optimisations and informs decision-making at multiple levels of the system. Under a smart cycling paradigm, the bike should no longer be a single isolated entity. Instead, it should become a vehicle that can sense and share data, interact with other road vehicles, and learn from its environment and how it is used. This capability to observe reality, learn from the initiatives undertaken, and continually adjust its services to incrementally and measurably improve the overall performance of the system should result in safer, more efficient or more enjoyable rides. Moreover, the significant volumes of digital traces generated from those rides should become a huge driver for new innovations, many of which will not even have been anticipated by the platform creators. Rather than a centralised data collector, the system should explore forms of situated intelligence that begin directly on the bike, e.g. to characterise the riding context, detect near miss situations or even send accident notifications, and go all the way up to the broader understanding of urban mobility.

Horizontal Innovation and Combinatorial Innovation: A normal consequence of generative platforms is that innovation activities increasingly become horizontal as efficiencies are gained by applying the same innovation activities and knowledge across multiple products or platforms (Yoo et al., 2012). Moreover, the combinatorial innovation principle means that the creation of new products or services may result from combining existing modules, which are often designed without fully knowing the whole design or how they will be used (Ciriello et al., 2018; Yoo et al., 2012). A platform should thus provide a horizontal technical foundation upon which many complementary products and services can easily be developed and combined, each offering its own set of knowledge-intensive features. This should significantly lower the barrier for new innovations, in the form of new models or new multiple recombination possibilities that can emerge from re-purposing existing models across many usage contexts. The horizontal innovation involved would also enable the expertise associated with data models to become readily available across domains and usage contexts. This would help to bring together what are now separate experiences and domains, such as urban planning, urban KPIs, mobility models development, cycling activism and cycling itself. Finally, generativity would mean that many more and more advanced models could quickly emerge from the concrete learning resulting from the wide availability and usage of those models.

Progressive convergence of mobility models and data: A digital platform can harness convergence by bringing together separated products, user experiences, and industries (Yoo et al., 2012; Ciriello et al., 2018). From the results of the study with mobility models, we can identify the need to promote convergence between many independently developed analytics models. This should go beyond the use of a shared set of horizontal services or libraries and lead to a mesh of multiple interconnected models that may depend on each other as different layers of data processing. Given the inherent diversity of mobility models and data sources, any convergence strategy should be very minimalist, avoiding constraints on how those models are executed, on what programming language they are developed or on the types of input data they might use.

Regarding data, a platform may promote some convergence by facilitating the use of canonical data formats for representing the most common types of data. While totally optional, a few well-known data formats for common data types, promoted by an extensive collection of conversion and aggregation functions, may offer a shared resource that makes data convergence simpler, whenever suitable.

Regarding the algorithms of the models, the convergence process should promote their decomposition into basic steps. Each model needs to be described in a way that formally defines the model generation process as dataflows, with specific data sources, data transformations, concrete parameterisations of the algorithms used and concrete and publishable outputs. This could then be used to promote the progressive mapping of some of those parts into horizontal features offered directly by the platform. The explicit description of the transformation processes would thus constitute a key enabler for convergence. Even a simple description process, when made coherently for multiple dataflows, can help to blur differences between various processes and emphasise any existing similarities. This should be a fundamental first step to decouple some of those elements from specific processes and make them available as horizontal features readily available for composing new models.

This should have a major impact on the development of cycling mobility models, reducing the development effort and potentially improving their overall quality, as most of their code would now be based on mature services with greater development quality, better performance and simplified programming models. It should also make the overlap between different models more explicit. Very similar models can be described in different ways and using diverse terminology. While it may not be possible to define absolute boundaries between model concepts, this analytics platform may help to promote convergence towards recognisable model categories with similar goals and offering similar properties. This should facilitate comparative analysis that places side-by-side alternative ways to produce similar outputs and assess their relative performance. This progressive abstraction and decomposition should ultimately lead to a

many-to-many relationship between data sources and models, with common data types being explored in multiple ways by many complementary models and some models supporting the ability to be generated from alternative sources, and in this way be able to adjust themselves to the data reality of different cycling ecosystems.

3.5 Summary

Achieving the full potential of Intelligent Transportation System (**ITS**) requires the convergence of the efforts and the knowledge of multiple entities. This is also true for cycling mobility and for the many models that aim to provide particular perspectives on the cycling reality of cities. Currently, their prevailing development paradigm is based on dispersed, vertical, self-contained and custom processes, which fail to offer the necessary context for distributed and open innovation. As a consequence, each new effort results in a new package of data, expertise and methodologies that are hard to reproduce, generalise, recombine or improve outside their original context.

This chapter has shown that, despite their current dispersion, commonly used cycling models are composed of basic processes with a high potential to be used as horizontal features in the context of a digital platform approach. This represents an initial validation of the problem and highlights an opportunity for a new type of solution based on a platform strategy. This chapter has also identified some of the dependencies that are often embedded in the implementation of those models and prevent them from being widely repurposed by people outside their initial development context. These are fundamental hints for understanding how they can be redesigned around horizontal features shared among multiple model implementations.

Building on those results and on the more general principles of digital innovation associated with digital platforms, this chapter has also outlined a set of key design principles for the development of a digital platform strategy for cycling analytics. This should constitute a valuable contribution to inform the development of a new generation of cycling platforms for urban environments.

One key observation is that many of the generic atomic blocks (identified in this chapter and during the related work review) consist of models associating geographical information with a common reference system. Common examples included associating with a road segment, its slope, traffic volume, or average speeds. These associations are essential for providing actionable information to both cyclists and decision makers, directly supporting decisions and activities in urban mobility.

As discussed in section 3.3, the reference systems serve as the ultimate convergence mechanism for

harmonising models and datasets, enabling immediate associations between different data sources that describe the same region. The demand for cycling services is global, yet without an adequate reference framework, models become difficult to replicate across various contexts, and comparisons between models targeting similar problems remain problematic. As a result, many generic code blocks risk remaining context-bound, limiting their potential and reducing the value of development efforts. Therefore, before dealing with common challenges related to the development of a digital platform, such as selecting adequate technologies, designing the co-creation procedures, designing effective governance mechanisms, or establishing network effects, it is essential to resolve this foundational requirement.

As introduced in Chapter 1, OSM stands out as a natural candidate given its open-source model, global coverage, recognised good quality, and widespread adoption in research and practical applications. However, a key requirement for OSM to become an enabler for new technologies is that it needs to represent real cyclists' traces. The uncertainties about OSM's capability to represent real cyclists' traces stem from the fuzziness of what constitutes the cycling road network and the ambiguities regarding road representation and mapping practices.

The following chapters will address the main research question of this thesis and explore whether OpenStreetMap can serve as a reliable reference system for cycling analysis and an enabling technology for cycling information systems.

Chapter 4

The Gaps Between Cycling Traces and OSM Road Network

4.1 Motivation

The perception about the limitations of the OSM road network data model to represent real cyclist routes emerged from initial experiments involving the collection and processing of sensor data to characterise urban cycling behaviour at both individual and aggregate levels, as reported in (Cabral et al., 2021). The list of sensors included accelerometers, gyroscopes, GPS, sound, and distance sensors. Among the collected data, we obtained multiple GPS traces of cycling activity, which we subsequently attempted to map-match into the OSM road network. While working with the data, we observed several instances where the map-matching procedure failed to find an accurate representation of the GPS trace, resulting in noticeable discrepancies between GPS traces and the corresponding map-matched routes. We refer to those discrepancies as map-matching gaps, as defined Definition 1.

Definition 1. *Map-matching gaps correspond to any situation in which the attempt to map-match a cyclist's GPS trace into the OSM road network results in a map-matched route that does not accurately represent the original trace.*

A closer examination revealed that these gaps were frequent and stemmed from a variety of factors. Based on our preliminary analysis of the cycling data, we identified an initial set of causes behind those gaps, as presented in Table 5.

In this chapter, to gain a deeper understanding of map-matching gaps, we aim to characterise the concrete causes and the main properties associated with the discrepancies between real cycling traces and how they are mapped into the OSM road network model. We also aim to assess the potential impacts of those mismatches and explore alternatives that may help to mitigate or even eliminate the negative consequences.

Table 5: Common situations for the occurrences of map-matching gaps.

Case	Case Description
Recent roads	The cyclist rides on a street that was recently subjected to traffic changes, which may not yet be represented in the OSM road network model.
Shortcut	The cyclist takes a shortcut, i.e., a non-official but convenient path that may not be represented in OSM.
Wrong direction on a one-way street	The cyclist rides against traffic on a one-way street.
Pedestrian	The cyclist rides on a sidewalk or pedestrian path.
Dirt track	The cyclist takes a dirty road, possibly created through the passage of pedestrians or other cyclists over time.
Zebra crossing	The cyclist uses a zebra crossing to go from one side of the road to another (either dismounting or not).
Pedestrian Bridge	The cyclist takes a pedestrian bridge.
Semi-public roads	The cyclist takes a road located in a semi-public area, such as a university.
Adjacent roads	Car and cycle roads are adjacent, and even minor positioning errors may result in an incorrect match.

4.2 Methodology

To capture a broad range of map-matching gaps, we planned an experimental route in Braga encompassing 19 situations likely to cause discrepancies between the raw GPS trace and the corresponding map-matched route. Braga is a mid-sized city with limited cycling infrastructure, where cyclists often choose unconventional paths to avoid sharing the roads with cars. Each situation listed in Table 5 was incorporated into the route design at least once. Since the route was specifically planned for this purpose, this experimental route did not aim to be representative of the number of potential problems that may occur on real cycling traces, or representative of the frequency and the relative weight of the various types of map-matching gaps. Nonetheless, all the situations included in this route corresponded to realistic cycling trajectories commonly used by cyclists in their daily journeys. Therefore, the potential map-matching gaps are thus genuine, even though their concentration within this route was artificially created for this study. It is also important to note that we do not expect to capture the full spectrum of potential categories.

We collected data by riding the planned route with a bicycle equipped with a smartphone and an

action video camera, both of which were attached to the bicycle's handlebars. Figure 4 represents a map view of the GPS trace recorded in Braga. On the mobile phone, we used the GPS logger application[1] to register the GPS trace. To avoid unintended effects on the map-matching algorithm, we configured the application to avoid using battery optimisation or features to filter noisy GPS signals. The video feed was used as a complementary data source, especially for clarification in cases where additional knowledge about the trajectories could become necessary for the correct interpretation of the results. The video also allowed us to generate images to illustrate the various seed cases embedded in our experimental route.

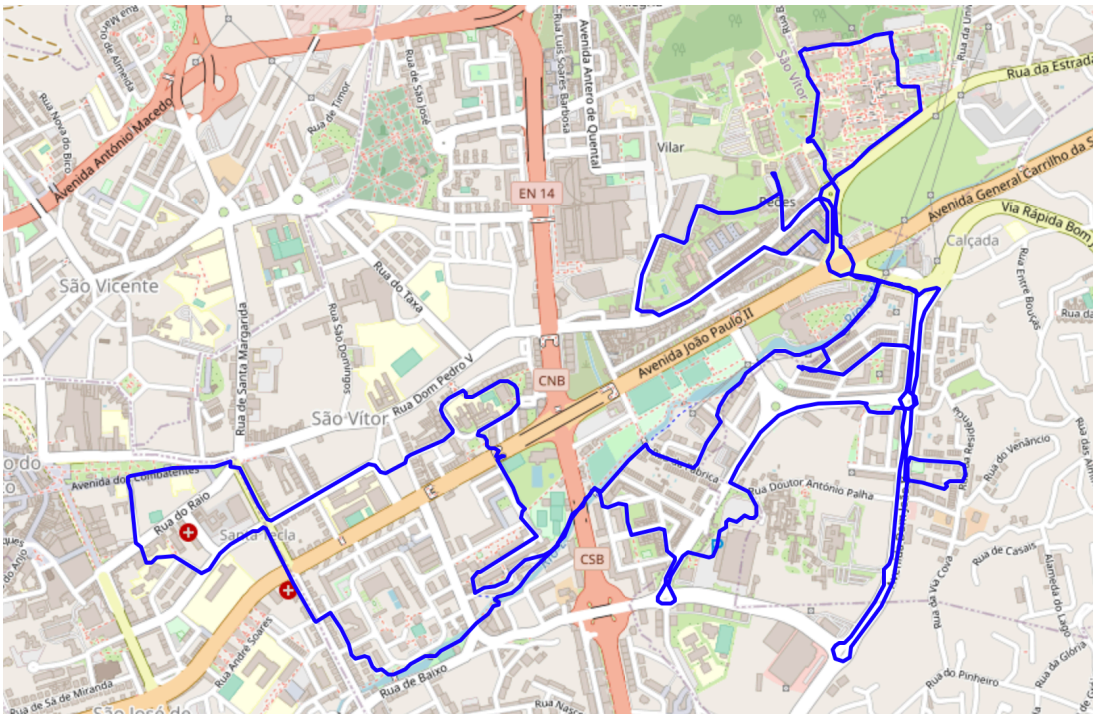


Figure 4: Map view of the original bike trace recorded in Braga.

4.2.1 Route Processing - Map-Matching

To analyse the trace data, we started by map-matching the GPS trace onto the OSM road network. For this purpose, we used the map-matching engine developed by Mapbox (Mapbox, 2020). This engine is OSM-based and offers 100,000 free requests per month. A single request to the Mapbox Map-Matching API[2] can have between 2 and 100 coordinates and should be framed by a specific mobility profile: driving, cycling, or walking. We configured the map-matching service to consider the bicycle profile, which is designed to avoid highways and give preference to streets with bike lanes. The response of a map-matching request will fall into one of the following 3 categories:

1 https://play.google.com/store/apps/details?id=com.peterhohsy.gpsloggerlite

2 https://docs.mapbox.com/api/navigation/map-matching/

(i) **Map-matching Error:** The GPS trace could not be map-matched into the OSM road network. The result is an execution error, and no route is returned.

(ii) **One map-matched polyline**: The GPS trace was map-matched into a single polyline.

(iii) **Multiple map-matched polylines:** The GPS trace was map-matched into several polylines. This indicates that some portions of the request could not be map-matched into the OSM road network, resulting in discontinuities.

We should emphasise that our objective is not to evaluate the performance of any map-matching algorithm. Rather, Mapbox was selected as a representative example of a widely adopted, state-of-the-art map-matching service.

The execution of the map-matching process was made for the entire trace and was totally independent from any initial considerations about possible map-matching gaps or seed situations. Following Mapbox guidelines, the points were filtered so that every point in the request would be at least 10 meters apart from the previous one and registered at least 5 seconds after the previous one.

4.2.2 Identification of Map-Matching Gaps

To analyse the map-matching results, we created an OSM map representation in QGIS[3], overlaying the original GPS trace and the route proposed by the map-matching process. We then analysed the whole route to identify any cases that could correspond to the definition of a map-matching gap. All the map-matching gaps reported were identified through manual inspection of the map visualisations developed in QGIS. In cases where QGIS did not provide sufficient information to infer the causes of the gaps, we further investigated the OSM road network database to identify the underlying causes. This involved reviewing the specific tags assigned to each road segment and analyzing the topological connections at intersections.

More specifically, we signalled for further analysis any situations where the trace and the route proposed by the map-matching seemed to either diverge or follow parallel directions. We also used our own field knowledge about each concrete situation and the video feed whenever needed to assess in more detail the characteristics of each situation and to decide whether or not they qualified as a gap. The selection criteria were that a gap should correspond to a failed mapping or to a mapping made to a road segment that was not representative of the trace section that it was meant to model. For example, a segment made on a sidewalk that was mapped to the main road would qualify as a map-matching gap.

[3] `https://qgis.org/`

It is important to note that the analysis of the map-matching results was independent of the 19 seed situations considered when planning the experimental route. The usage of manual inspection was needed because we had to start by gaining a deeper understanding of the specific ways in which these mismatches occur, before we could be ready for more automated and quantitative methods of analysis.

4.3 Results

In total, we identified 16 situations that corresponded to our definition of map-matching gaps, as provided in Definition 1. To support subsequent analysis, we registered, for each gap, a unique ID (1-16), its start and endpoint and a problem description. Figure 5 shows an example of one of those 16 situations. The GPS trace is represented by the blue continuous line with arrows, while the map-matched route is represented by the dashed red line. In this example, the cyclist used the sidewalk next to a one-way road and in the opposite direction to traffic. The map-matching algorithm returned a route that clearly does not represent the GPS trace.

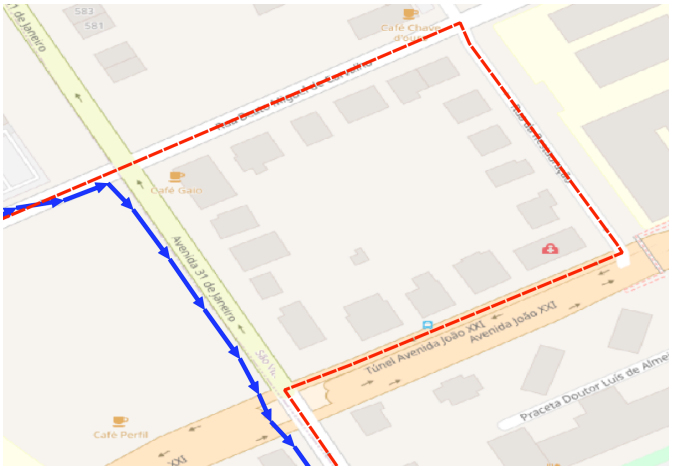


Figure 5: Map-matching of a GPS trace performed on a sidewalk along a one-way road.

4.3.1 Gap Categories

After the identification and description of map-matching gaps, we then aggregated those gaps into categories according to the assessment of the respective cause. This resulted in the categorisation presented in table 6.

Missing Segment: We found 6 occurrences related to missing road segments. Figure 6 shows the occurrence ID3, representative of this category. In this case, the cyclist used a pedestrian pathway behind a building which was missing in the OSM database.

Table 6: Identified gap categories

Gap Categories	Gap IDs	#	Category Description
Missing Segment	3; 7; 8; 9; 15; 16	5	Occurrences where cyclists rode on segments not defined in the OSM road network.
Missing intersections	15	1	Occurrences where the intersections between two or more roads are not defined in the OSM road network.
One-Way	12; 13; 14	3	Occurrences where the cyclist rode on a one-way road but contrary to the traffic direction.
Competing Roads	1; 2; 6; 11	4	Occurrences where the map-matching was unable to correctly match the real trace segment due to multiple viable alternatives that the cyclist can ride on.
Semi-public Roads	4; 5	2	Occurrences where the cyclist used semi-public roads, tagged as private on the OSM road network.
Complex Crossing	10	1	Occurrences where the cyclist arrived at a crossing and took a non-conventional path to avoid dangerous situations.



Figure 6: Missing Segment (ID3)

The omission of road segments has a direct impact on map-matching and on routing. Map-matching algorithms cannot consider those roads eligible for aggregating GPS traces and are induced into total failure, such as in cases similar to Figure 6, or to produce a detour to connect eligible traces. Whereas routing engines do not consider those roads when computing route suggestions, and will produce a route that does not match real cyclists' desire lines. The omission of road segments also impacts mobility

models through the loss of relevant information or misattributing data to incorrect roads.

Missing Intersection: We found 1 situation where the map-matching has produced a gap due to the omission of an intersection between two roads, as shown by Figure 7. In this case, due to the omission of that intersection, the map-matching algorithm was forced to find a route that could connect both ways. The alternative route found did not represent the GPS trace. These types of omission errors also impact route suggestions and mobility models.

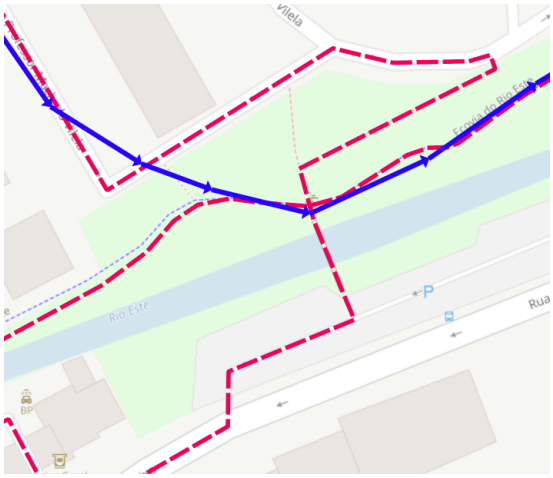


Figure 7: Missing Intersection (ID15)

This example was also assigned to the "missing segment" category because, at that time, a small shortcut through the grass was not represented in the OSM road network. However, that segment is already visible in Figure 7.

One-way: We found three situations where the map-matched routes were not representative of the cyclist's trace due to the cyclist travelling against the traffic on a one-way road. Figure 8 shows an example of this gap category, where the cyclists used the sidewalk to avoid riding against traffic on the main road.

Analysing the three similar cases, we could observe that the map-matching results differ. In case ID12, shown by figure 5, the map-matched route clearly does not represent the cyclist's trace. In cases ID13 and ID14, map-matched returned a route close to the original trace. However, despite its proximity, they were inaccurate, as the cyclist used the sidewalk to travel. These different outcomes were related to the map-matching internal configurations and how the algorithm uses OSM road network data during the matching process. In case ID12, the road was tagged as "secondary", while in cases ID13 and ID14, the roads were tagged as "tertiary" and "residential". It is also possible that, in case ID12, the alternative route was faster because the cyclist could ride mounted, while in cases ID13 and ID14, the matched routes were the fastest even if dismounted. It is also important to note that, in all three cases, OSM had no representation of the sidewalk.



Figure 8: One Way (ID13)

These examples highlight the importance of defining the desired behaviour of map-matching accordingly to the objectives or assumptions of the study, such as considering or not one-way roads eligible for matching GPS traces in both directions. It is expected that other map-matching algorithms and even the same algorithm configured differently will produce different results. Nonetheless, the problem still relies on the fact that some paths made by cyclists have no support in the official road network.

Competing Road: We found 4 occurrences related to competing roads. Figure 9 shows an example of a gap of this category. In this case, the cyclist used the dedicated cycleway; however, the map-matched snapped the GPS trace into the pathway at first and then into the main road.

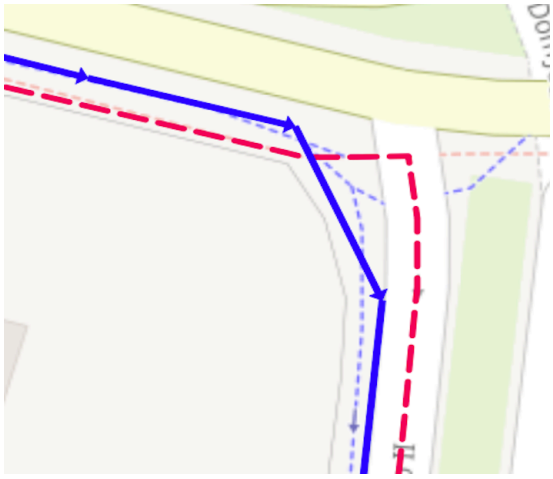


Figure 9: Competing Road (ID1)

Competing roads might correspond to different lanes of the same road or roads close to each other, such as a cycleway parallel to the main road. Unlike cars, cyclists can use both roads and bikeways, which significantly increases the complexity of the map-matching algorithm when both roads are next to each

other. In these cases, even minor GPS errors might be enough to cause the mismatch.

Complex Crossing: We found one occurrence where the cyclist used the zebra crossing to avoid using the main road, as shown by Figure 10. In this particular case, the map-matching algorithm snapped the GPS trace into main roads instead of mapping into the sidewalk and zebra crossing.



Figure 10: Complex Crossing (ID10)

Applying map-matching to similar scenarios is additionally challenging because, in those situations, there are multiple alternative roads, and even minor GPS interferences may induce the map-matching algorithm into error.

Semi-public Roads: We found 2 occurrences related to cycling riding on semi-public roads, as exemplified by Figure 11. In this case, the cyclists rode on roads and pathways located inside the University of Minho campus.



Figure 11: Semi-public Road (ID4)

These types of errors are closely related to the configuration of the map-matching algorithm. In the

case of the Mapbox map-matching service, it does not consider roads tagged as "private=yes" eligible to be used during the map-matching process. It is expected that other algorithms with different configurations will produce different outcomes.

4.4 The Limitations of the OSM Road Network Model

The results confirm various mismatch scenarios between actual cycling trajectories and their map-matched representations onto the OSM road network. More specifically, they suggest that the main overall cause of this problem is directly linked to limitations of the road network model. From all six categories of map-matching gaps identified in this study phase, only one, competing roads, cannot be linked to those limitations. Two roads adjacent to one another are a worst-case scenario for map-matching of cycling traces. As cyclists may potentially use any of those roads, both possibilities need to be considered, and even small positioning errors may be enough to produce a wrong match. For all the other cases, however, the problem is not associated with any complexity of the map-matching process. It is always a direct consequence of a mismatch between the real GPS traces and the roads represented in OSM. Since there is no suitable match, the map-matching process could never provide a proper one.

These findings indicate that current road network models, such as those predominantly used in OSM, fail to adequately capture the specificities of cycling and micro-mobility. Even if this is not necessarily a limitation of OSM itself, the prevailing road network models seem to be based on the assumption that cyclists will always follow either car roads or cycle paths. Consequently, it fails to consider all sorts of shortcuts, alternative paths, or specific crossing strategies at complex intersections that cyclists resort to daily in an attempt to make their journeys more efficient and safer.

Given this discrepancy between actual cycling behavior and the underlying road network models, frequent map-matching failures are an inevitable consequence. Crucially, each unmatched segment constitutes a significant loss of data, directly compromising the ability of cycling information systems to accurately reflect real-world cycling activity. This has a twofold impact. The first is the impact on the accuracy of the models that can be created based on the road network information. Those models can be very valuable for urban planning and mobility management, but if they are based on an inaccurate road network model, then they will also be inaccurate or even misleading. Also, if the road network does not represent all the possibilities, and particularly those that cyclists seem to prefer, navigation services will necessarily suggest routes that are not ideal. If a cyclist systematically receives route suggestions that seem to ignore what the cyclist perceives as some obviously better alternatives, then the confidence in that navigation service

is broken. This can be particularly negative for beginner cyclists or tourists, who are less knowledgeable about the shortcuts or alternative roads that could offer them more comfort, safety, or efficiency.

From a technical standpoint, these findings highlight the potential for automating the large-scale identification and characterisation of map-matching gaps. This would create awareness about their existence and could serve to trigger revision processes that included an official assessment of concrete situations. Unmatched trace segments would no longer be wasted information. They would become the core input for the progressive alignment between OSM and the reality it intends to represent. Some cases could be addressed by simply updating OSM representations. Others could serve as a sign for the need to perform small interventions in the physical infrastructure to prevent people from taking specific paths that might be considered dangerous, or make any necessary adjustments to officially recognise them as cycling routes. Yet another alternative might involve the creation of new OSM tags to signal de facto routes, possibly making them an option that users may or may not consider when using mapping services.

4.5 Summary

All over the world, in high-profile cycling cities or in cities just trying to make their first steps towards sustainable urban mobility, local cyclists apply their local knowledge to identify the best possible routes across a complex and heterogeneous network of roads. The main contribution of this chapter was the identification and characterisation of an important gap between those cycling traces and their representation in the road networks of geographic systems. More specifically, there are parts of the traces that represent the cyclist's trip that cannot be map-matched into the road network of OSM because they have no representation. We then discussed the impact of those map-matching gaps on the quality of the services provided to urban cyclists and on the value of the data available for micro-mobility planning and management.

Chapter 5

A Supervised Method to Automatically Detect and Classify Map-Matching Gaps in Cycling GPS Traces

5.1 Motivation

In the previous chapter, we characterised a broad range of situations corresponding to gaps between cycling traces and their representation in the OSM road network model, deepening our knowledge about the specific ways in which these mismatches occur and their characteristics. During that study, we manually inspected the map-matched GPS trace to identify these gaps. However, this approach was feasible because we were only considering a single GPS trace. To scale the identification of map-matching gaps into a larger and real context, this would be a time-consuming task.

In this chapter, based on the deeper understanding of the map-matching gaps resulting from the previous chapter, we propose a methodology based on supervised machine learning to detect relevant discrepancies between cyclists' GPS traces and the routes produced by a map-matching algorithm and classify those discrepancies accordingly to their root cause.

5.2 Methodology

The methodology for this study involved training and evaluating machine learning approaches to automatically identify and classify map-matching gaps. The process consisted of six main steps. First, we collected a set of GPS traces representing cycling activity in four European cities. Second, we gathered the corresponding OSM road network data for those cities. Thirdly, we map-matched the GPS traces into the OSM road model. Fourth, we manually inspected the map-matching results to establish ground-truth data. Fifth, we computed several similarity measures to compare each GPS trace with its corresponding map-matched route. Finally, we trained and evaluated a set of binary and multi-class classification models

using the similarity measures as features and the ground-truth data as labels. During this step, we applied hyperparameterization with cross-validation to improve the confidence in the results.

5.2.1 Mobility Data Acquisition and Preparation

To guarantee diversity and authenticity, we collected GPS traces from four very distinct cities, regarding their size, cycling culture, and mobility policies. More specifically, we considered the cities of Braga (Portugal), Seville (Spain), Paris (France), and Amsterdam (Netherlands). Using the *Wikiloc*[1] website, we searched for cycling traces in the selected areas and downloaded a set of at least 20 GPS traces corresponding to routes that had been effectively made by cyclists in each of those cities. In total, we downloaded 98 GPS traces. To improve the granularity of the analysis and fully understand the behaviour of the map-matching process, we divided these traces into slices of approximately 1,000 meters, depending on the distance between consecutive GPS points. The last slice of each GPS trace was composed of the remaining GPS points and would thus be smaller than 1,000 meters. This resulted in 935 trace slices.

5.2.2 Road Network Data Acquisition

For each of the selected cities, we created a database with the respective OSM road network model. We started by obtaining the road network data, using the *geofabrik*[2] website. Secondly, we used the *osmium*[3] tool to select the data corresponding to the main urban area. Then, we applied the *osm2po*[4] tool to convert the selected road network model into a routable model. For this conversion, we considered car, pedestrian, and cyclist roads. Finally, we created an instance of a PostgreSQL[5] database with the PostGIS extension and imported the topology data using the *psql*[6] tool.

5.2.3 Map-matching

In this study, our focus is not on the quality or any particular properties of map-matching algorithms. We are only concerned about the discrepancies between the road network model and the cyclists' traces. To reduce any effects of the algorithm selection on the results of our study, we opted for the pgMapMatch (Millard-Ball et al., 2019) algorithm. This is a widely used algorithm, which has been classified as the

[1] https://www.wikiloc.com/

[2] https://www.geofabrik.de

[3] https://osmcode.org/osmium-tool/

[4] https://osm2po.de

[5] https://www.postgresql.org

[6] https://www.postgresql.org/docs/current/app-psql.html

best-performing algorithm by Berjisian and Bigazzi (2023) and seemed acceptable as a representative example of the current state of the art in map-matching algorithms. The advantages of the pgMapMatch algorithm are that it grants more control over the map-matching process, its code is publicly available, and it is possible to configure the road network model data that the algorithm will use to match the GPS traces.

We thus used pgMapMatch to perform a map-matching operation in each of the slices obtained in the previous step. For each city, we started by configuring the *pgMapMatch*[7] tool to use the respective database instance as the source data for map-matching processes. Finally, we map-matched each GPS trace slice into the OSM road network model and built the resulting geometry.

5.2.4 Ground-Truth Data Acquisition

To obtain ground-truth data, we manually inspected the map-matching results to identify and categorise every map-matching gap. For each slice, we generated a map visualisation representing the road network, the original GPS trace slice, and the map-matched route. We then analysed each of these 935 visualisations to identify any anomalous situations. In this context, a map-matching gap was a case where it was obvious from visual inspection that the map-matched route was not the best option for representing the trace taken by the cyclist, except for some parallel roads, where human reasoning can be very effective at disambiguating by considering background knowledge about cycling and land usage in the area represented by the map. The categorisation of each map-matching gap was determined according to the most plausible cause for its occurrence. In some cases, assessing its cause required the inspection of the OSM road network data to get details about road topology and tags.

5.2.5 Generation of Similarity Measures

The key approach for automating the map-matching gap identification is to compare the similarity between the GPS trace and the road network to which the trace was matched. At this stage, we computed two base length features and six well-established similarity measures to compare each GPS trace slice with its map-matched counterpart. These similarity metrics are all based on literature. Five of them were also used in the work previously conducted by Berjisian and Bigazzi (2023). The complete set of extracted features is detailed below:

GPS Trace Slice Length (TL): The total length of the GPS trace slice.

Map-matched route length (ML): The total length of the map-matched route.

[7] `https://github.com/amillb/pgMapMatch`

Length index (LI): The ratio between the length of a GPS trace slice and the respective map-matched route (Schweizer et al., 2016). In optimal scenarios, this value would be close to 1.

Average distance (AD): The average of the distances between each GPS point of a GPS trace slice and the respective map-matched route (Schweizer et al., 2016). This metric uses the distance between each point in the GPS trace slice and the nearest point in the map-matched route. The order of the points is not considered.

Average distance error per record (ADE): The average value of the distances between each GPS point of a GPS trace slice and the map-matched route, as proposed by Berjisian and Bigazzi (2023). For each GPS point of a GPS trace slice, this approach considers the distance to two possible map-matched route segments: the segment assigned to the previous GPS point and the following segment. The smaller distance is used. However, if the cumulative distance between the first GPS point assigned to the first segment and the current GPS point is longer than the length of the first map-matched segment, the distance to the second is always used, and this process recommences.

Discrete Fréchet Distance (FD): The Fréchet distance assesses the similarities between two geometries and can be explained as: A man walks a dog with a leash. They walk on two curves independently with varying speeds. The Fréchet distance is the minimum leash length required to traverse both curves (Eiter and Mannila, 1994). The higher the Fréchet distance is, the less similar the curves are. There are multiple variants of this metric. In the weak Frechet variant, one or both "entities" can walk backwards. We use the strong Fréchet variant, where only movement forward is allowed. The discrete variant is an approximation for polygonal curves.

Dynamic time warping (DTW): DTW is used in many areas to measure the similarity or the distance between two sequences (Toohey and Duckham, 2015). In the context of our study, the sequences are composed of WGS84 latitude and longitude coordinate pairs. One sequence represents the GPS trace slice and the other the map-matched route. The distance between each pair of points is computed with the haversine formula.

Alignment (A): The alignment metric, as proposed by Berjisian and Bigazzi (2023), describes the average difference between the bearings of the GPS trace and the corresponding map-matched route over each 5 consecutive GPS points. To obtain the start point of the corresponding map-matched route segment, the first GPS point of the interval is projected into the map-matched route. The last point of the map-matched route segment is obtained by walking on the map-matched route a distance equal to the cumulative distance between the GPS points considered.

These eight features were computed for all 935 trace segments comprising our ground-truth dataset.

5.2.6 Classification models

Supervised machine learning classification aims to categorise data or predict outcomes based on prior labelled information (Singh et al., 2016). It is used in many data science problems and comprises two phases. First, the classifier is trained using a training dataset. Then, the performance of the resulting model is evaluated against labelled test data.

Initially, we trained a set of binary classifiers using the similarity measures as features and the ground-truth data as labels to detect gaps between OSM and actual cycling traces. Subsequently, we trained multi-class classifiers on the same dataset to categorize the identified gaps according to their root causes. In each stage, we tested different feature engineering approaches and conducted hyperparameterization optimization with cross-validation to ensure model robustness.

Performance Metrics

Model performance was evaluated using standard classification metrics. For the binary models, we computed accuracy, precision, recall, and F1 score (Iwendi et al., 2020; Gyawali and Qian, 2019). For the multi-class models, we considered accuracy alongside macro-averaged precision, recall, and F1 score (Takahashi et al., 2022). These metrics are well-known and used frequently in machine learning-related studies.

We now present the equations for these performance metrics. The following terms are used in the equations: true positives (**TP**), true negatives (**TN**), false positives (**FP**), and false negatives (**FN**).

Accuracy: Accuracy is the ratio between the sum of correctly predicted cases and the total number of cases. This metric is calculated in equation 5.1, as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

Precision: Precision measures the ratio between true positives and the sum of true positives and false positives. This metric is calculated in equation 5.2, as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

The macro-average precision score is calculated as the arithmetic mean of individual classes' precision scores.

Recall: Recall, also known as sensitivity, measures the ratio between true positives and the sum of true positives and false negatives. This metric is calculated in equation 5.3, as follows:

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

The macro-average recall score is calculated as the arithmetic mean of individual classes' recall scores.

F1 Score: F1 score represents the harmonic mean of the recall and precision values. This metric is calculated in equation 5.4, as follows:

$$F1Score = 2 * \frac{Precision + Recall}{Precision * Recall} \tag{5.4}$$

The macro-average F1-score is calculated as the arithmetic mean of individual classes' F1-scores.

5.3 Results

This section describes the results of applying the methodology described in the section 5.2. We start by describing and analysing the results obtained by applying the map-matching algorithm to input GPS traces. Afterwards, we proceed to describe the training and application of the binary classifiers. Then, we present the results of performing multi-class classification. Finally, we assess the method proposed by Berjisian and Bigazzi (2023) to detect the gaps and compare its performance against our binary classification.

5.3.1 Analysis of the Map-Matching Process

The first part of this study involved collecting GPS traces representing bicycle activity, slicing them into smaller portions, map-matching the resulting slices into OSM, and performing manual inspection of the results to produce ground truth data. We used the GPS trace slice as our unit of analysis.

The manual inspection of the routes produced by the map-matching algorithm allowed us to identify and classify every map-matching gap. Table 7 describes the overall results, and Figure 12 shows 6 examples of common map-matching gaps.

Out of the 935 slices analysed, 417 cases (44.6%) were entirely map-matched with success. The remaining 518 slices (55.4%) had at least a small portion with a faulty map-matched situation. In some cases, the map-matching process failed at several portions across the slice. In 45 of those slices (4.8%), the map-matching even failed due to different reasons. As a result, these cases were assigned to multiple categories.

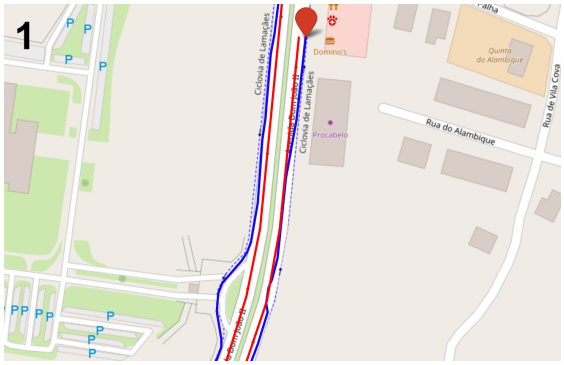
Comparing the map-matching results for each city, we observe that the success rate varied slightly. Amsterdam had the lowest success rate (39.8%), while Seville had the highest one with (47.1%). This can

Table 7: General stats: Map-matching gaps

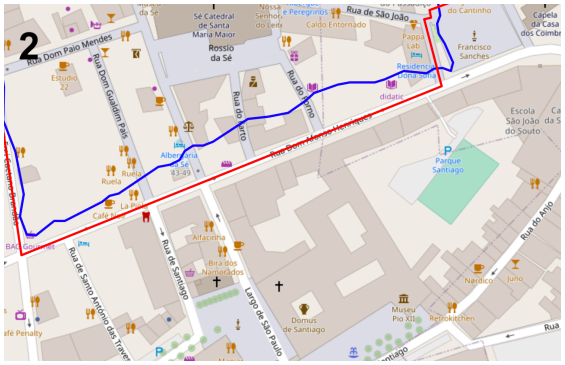
	Braga	Amsterdam	Paris	Seville	Total
GPS Traces	24	22	20	32	98
Total Length (Km)	213	179	96	406	894
GPS Trace Slices	226	191	98	420	935
% of Slices Without a Gap	44.7	39.8	42.9	47.1	44.6
% of Slices With At Least 1 Gap	55.3	60.2	57.1	52.9	55.4
% of Slices Assigned to One Gap Category	50.9	59.2	45.9	47.6	50.6
% of Slices Assigned to Multiple Gap Categories	4.4	1.0	11.2	5.2	4.8

be somewhat explained by the source of the map-matching gaps. In Amsterdam, the majority of these gaps were caused by "one-way" travel against traffic direction, which the pgMapMatching algorithm does not consider, even in cases where the cyclist used cycleways.

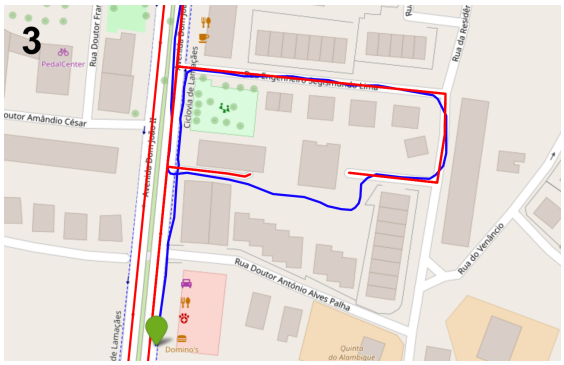
The manual classification of the map-matching gaps is represented in Table 8, which summarises the occurrences per category in each city. Since the total number of slices varies substantially across cities, we computed two ratios. The first ratio, called *"Rt"*, shows the ratio of occurrences of each category per



(a) Competing Roads



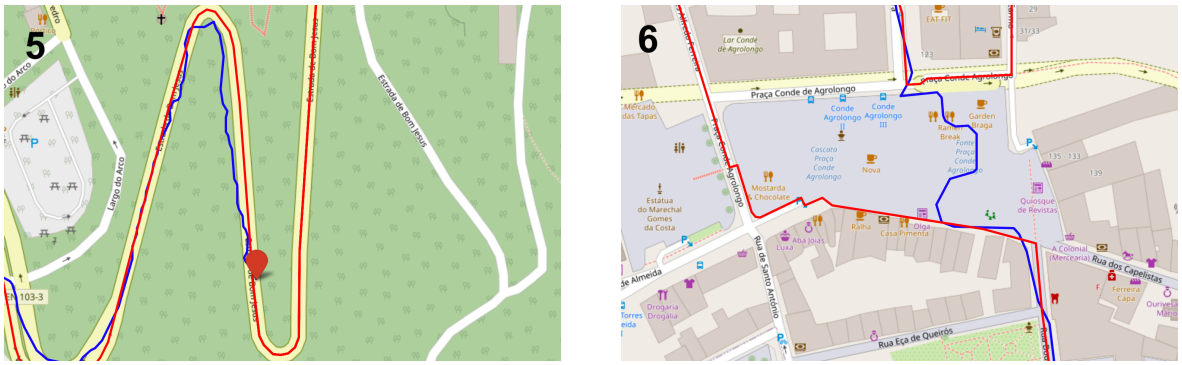
(b) GPS Error



(c) Missing Road



(d) One Way



(e) Map-Matching (f) Open Area

Figure 12: Examples of common map-matching gaps.

GPS trace slice. The second ratio, called *"Re"*, shows the ratio of occurrences of each category per gap. Since one gap can be assigned to one or more categories, the sum of Re can be higher than 100%. As we can see, despite similar map-matching success rates, the main source of error varies substantially across cities.

Table 8: Category occurrences per city.

Dataset	**Braga**			**Amsterdam**			**Paris**			**Seville**			**Total**		
Category	Cnt	Rt	Re	Cnt	Rt	Re	Cnt	Rt	Re	Cnt	Rt	Re	Cnt	Rt	Re
One-Way	53	23.5	42.4	71	37.2	61.7	30	30.6	53.6	52	12.4	23.4	206	22.0	39.8
Competing Road	9	4.0	7.2	21	11.0	18.3	20	20.4	35.7	72	17.1	32.4	122	13.0	23.6
Missing Segment	39	17.3	31.2	2	1.0	1.7	3	3.1	5.4	54	12.9	24.3	98	10.5	18.9
Map-Matching	24	10.6	19.2	2	1.0	1.7	3	3.1	5.4	49	11.7	22.1	78	8.3	15.1
GPS	2	0.9	1.6	18	9.4	15.7	11	11.2	19.6	13	3.1	5.9	44	4.7	8.5
Open-Area	5	2.2	4	0	0	0	0	0	0	2	0.5	0.9	7	0.7	1.4
Unknown	2	0.9	1.6	2	1.0	1.7	0	0	0	1	0.2	0.5	5	0.5	1.0
Circular Street	0	0	0	0	0	0	0	0	0	1	0.2	0.5	1	0.1	0.2
Complex Crossing	1	0.4	0.8	0	0	0	0	0	0	0	0	0	1	0.1	0.2
No Bicycle	0	0	0	0	0	0	1	1.0	1.8	0	0	0	1	0.1	0.2

The three main sources of map-matching gaps were directly related to the road network, namely one-way, competing-roads, and missing segments. This is in line with the results from Berjisian and Bigazzi (2023). They also point out that the most common sources of error were cyclists travelling in the wrong direction on a one-way street, travelling on missing links, and traces being map-matched into a parallel street.

The map-matching algorithm itself led to 78 map-matching gaps. In those cases, the algorithm wrongly returned the last portion of the map-matched segment duplicated. This would mean that the cyclist inverted

his direction of travel. However, by manually inspecting the results, we could observe that it was not the case. These errors were clearly due to map-matching algorithm limitations.

Bad GPS trace quality led to 44 map-matching gaps. In some cases, strong interferences on the GPS signals led the measurements to be recorded as being above buildings or rivers. In other cases, the GPS traces had a very low sampling rate or had long intervals without recordings. These situations, where the quality of the GPS trace is compromised, jeopardise the map-matching process.

Finally, we tagged 5 cases as "unknown" as we could not properly identify their main cause, even after inspecting the OSM road network data.

5.3.2 Binary Classification for Map-Matching Gap Detection

After performing the ground-truth analysis, we computed the similarity measures between each GPS trace slice and the corresponding map-matched route. We then developed a Python script to train and test binary classification models, using the sklearn[8] library. We considered 8 different classifiers: Logistic Regression (**LR**), Support Vector Classification (**SVC**), K-Neighbours (**KNN**), Decision Tree (**DT**), Random Forest (**RF**), Naive Bayes (**NB**), eXtreme Gradient Boosting (**XGBoost**), and Adaboost.

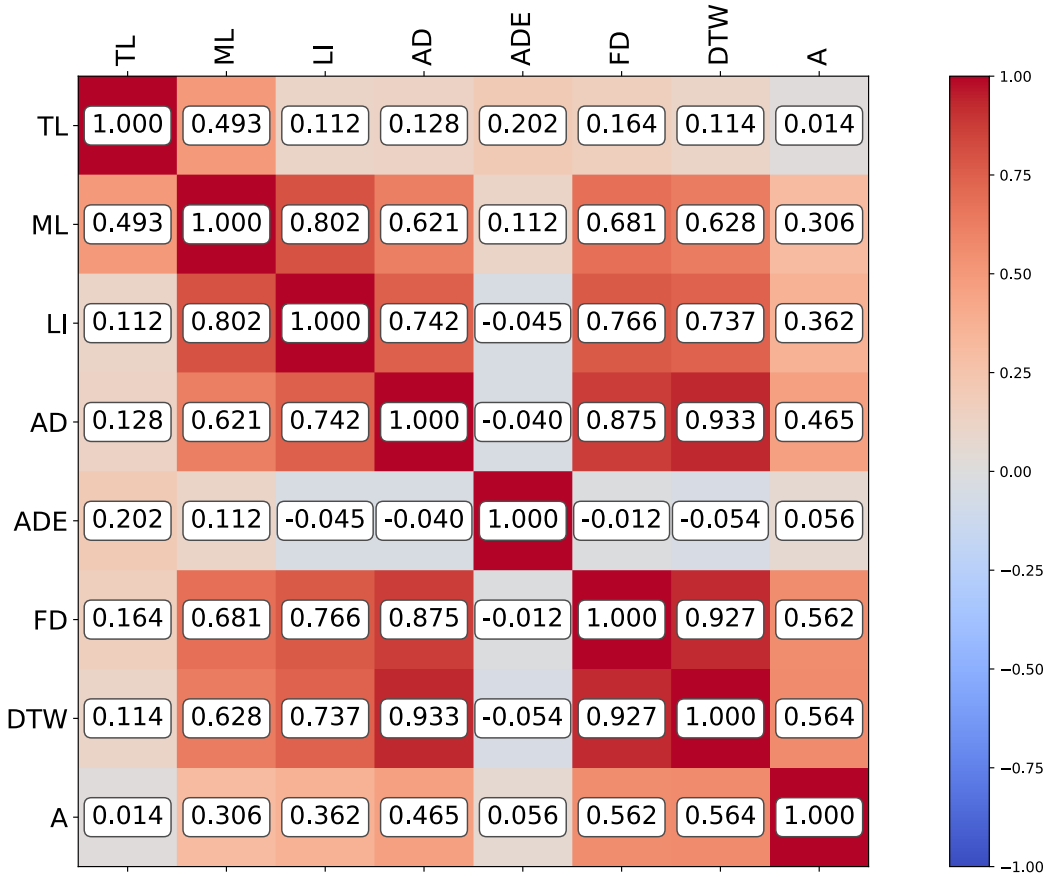
We employed a 75/25 hold-out split, randomly shuffling the data to allocate 75% to the training set and 25% to an unseen test set for final evaluation. To optimize the models without inducing data leakage, we performed hyperparameter tuning via random search with 4-fold cross-validation strictly within the training set.

We used four different feature engineering approaches, namely Analysis of Variance (**ANOVA**), principal component analysis (**PCA**), mutual information, and the Spearman correlation between features (Khalid et al., 2014). For each approach, we also tested with standardised values (**STD**). Depending on the approach, we used different sets of features:

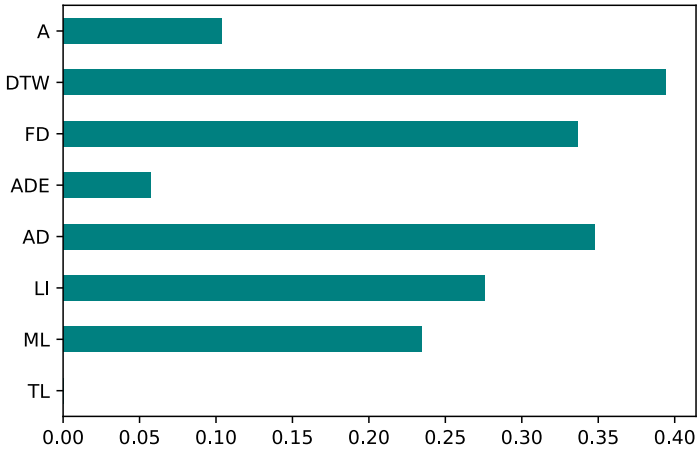
- "None": We used every feature available.
- "Corr": Based on the Spearman correlation between features shown in Figure 13a, we removed highly correlated features, using only the TL, ADE, DTW, and A.
- "Info-Gain": We considered the features with the most dependency on the result, namely DTW, FD, AD, and LI. The results of mutual information analysis are shown in Figure 13b.
- "PCA": We retained the top four principal components.

[8] https://scikit-learn.org/stable/

- "ANOVA": We selected the six most significant features based on their F-values.



(a) Spearman Correlation.



(b) Information Gain.

Figure 13: Correlation between features

Regarding the dataset, out of the 935 map-matching requests, 518 (55%) had anomalous results, showing that our dataset is balanced for the error identification phase. Table 9 shows a summary of the

results for the binary models, ordered by accuracy score. The "Version" column corresponds to the feature selection algorithm. Table 19 in Appendix A contains the results for every model trained.

Table 9: Binary Classification Results - Short Version.

ID	Model	Version	Prec.	Recall	F1	Acc.
1	**XGB**	ANOVA	0.883	0.931	0.906	0.893
2	**XGB**	ANOVA-STD	0.883	0.931	0.906	0.893
3	**DT**	Info-Gain	0.888	0.915	0.902	0.889
4	**KNN**	Corr	0.876	0.923	0.899	0.885
5	**KNN**	None	0.876	0.923	0.899	0.885
6	**KNN**	ANOVA	0.887	0.908	0.897	0.885
7	**XGB**	Corr	0.887	0.908	0.897	0.885
...	**...**	...	...	...	...	...
56	**SVC**	None	0.573	1.000	0.728	0.585

The results presented in Table 9 show that the performance of the models varied considerably, depending on the classifier algorithm and the feature engineering method used. Some models had a very low performance, with accuracy below 0.6. However, other models had a very good performance, with accuracy close to 0.89. The best-performing model was a trained XGB with ANOVA feature selection. The same model was also the model with the best F1 score, slightly above 0.9. Additionally, several other models had very similar performance, including decision tree, Adaboost, and logistic regression. Therefore, these results indicate that binary classification models can identify gaps between GPS traces and map-matched results with very good confidence.

5.3.3 Multi-class Classification for Map-Matching Gap Classification

In the second phase of our work, we trained a diverse set of multi-class classifiers, specifically NB, RF, LR, DT, KNN, and SVC. We utilized the same ground-truth dataset and similarity measures; however, to ensure robust model evaluation, we excluded instances tagged with multiple categories, as well as categories containing fewer than 20 occurrences due to their low statistical representativeness.

Consistent with our previous methodology, we employed a 75/25 hold-out split. To optimize the models while strictly avoiding data leakage, hyperparameter tuning was conducted via random search with 4-fold cross-validation exclusively within the training dataset. We evaluated these models across different feature engineering approaches, namely PCA, mutual information, Spearman correlation, and ANOVA.

Furthermore, in certain iterations, we applied the Synthetic Minority Oversampling Technique (**SMOTE**) preprocessing algorithm (Fernandez et al., 2018) to mitigate class imbalance among the map-matching gap categories. Specifically, in the iterations designated as "-SMT-U", we assessed model performance when applying SMOTE to oversample minority categories up to 100 entries, while simultaneously downsampling the "OK" category to 200 occurrences. Table 10 details the training dataset distributions before and after this specific feature engineering iteration.

Table 10: Train Dataset for the iteration: "-SMT-U".

	Train Dataset		Test Dataset
Category	Before	After (SMT-U)	Count
OK	309	200	104
Missing Segment	58	100	20
Competing Road	72	100	25
Map-Matching	59	100	17
One-way	128	128	44
GPS	31	100	9
Total	657	728	219

Table 11 presents the performance metrics of the top-performing models, sorted by accuracy. The comprehensive results for all trained models are available in Table 20 in Appendix B.

Table 11: Multi-class classification Results - Short Version.

ID	**Model**	**Version**	**Prec.**	**Recall**	**F1**	**Acc.**
1	**RF**	None	0.63	0.57	0.59	0.71
2	**RF**	Info-Gain	0.57	0.54	0.53	0.68
3	**RF**	ANOVA	0.59	0.53	0.54	0.68
4	**RF**	Corr	0.47	0.41	0.41	0.68
5	**DT**	ANOVA	0.54	0.51	0.50	0.68
6	**LR**	PCA	0.45	0.40	0.39	0.68
...	**...**	...	...	...	...	...
90	**LR**	Info-Gain-SMOTE	0.04	0.26	0.05	0.06

Figure 14 illustrates the confusion matrix for the best-performing model. Overall, the results indicate

that the predictive performance of the multi-class classification models remains limited. While the majority of these models successfully detect the occurrence of a map-matching gap, they frequently struggle to accurately classify its root cause. For instance, out of 104 true "OK" cases, 92 were predicted correctly. Regarding false positives, 12 occurrences were incorrectly classified as "OK", with the majority of these (10) actually belonging to the "competing-road" category. The confusion matrix further reveals significant misclassifications across various gap categories. Notably, most "missing-segment" gaps were erroneously predicted as "one-way" anomalies, and 12 out of 45 true "one-way" gaps were assigned incorrect labels.

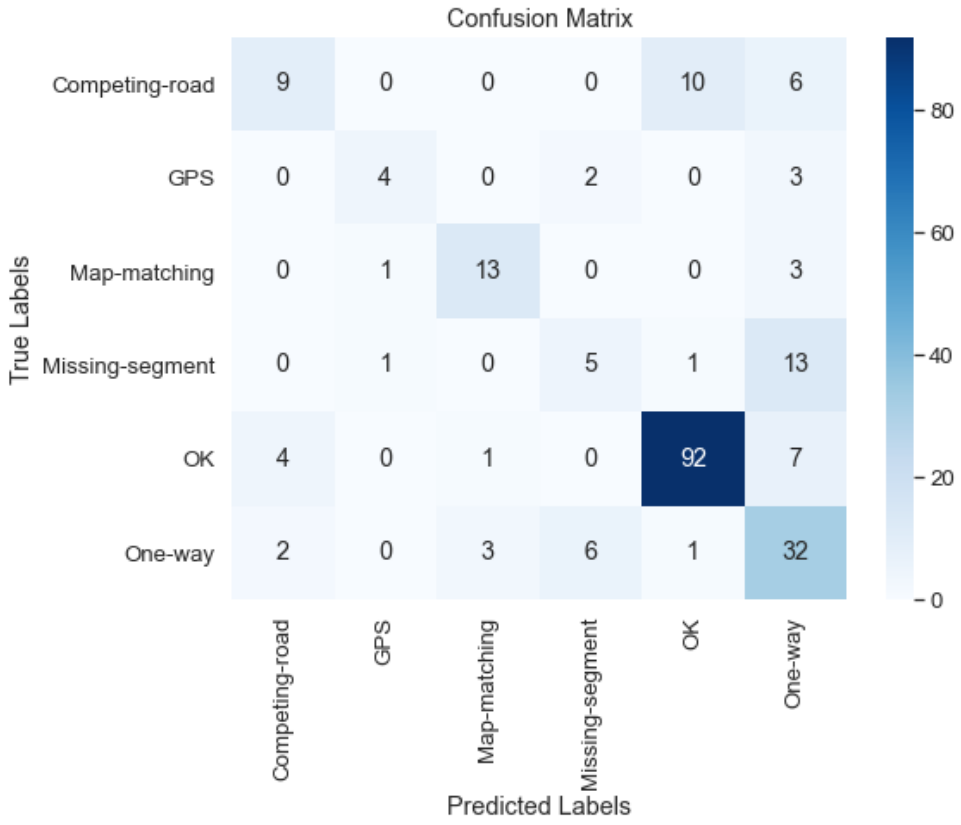


Figure 14: Confusion matrix for the RF model with no feature engineering.

In a real-world application context, these multi-class models could serve as a complementary tool to the initial binary detection. Despite their suboptimal classification accuracy, the outputs of these models can provide valuable heuristics for GIS analysts and mapping communities, streamlining the manual inspection process and facilitating targeted corrections to the road network model.

5.3.4 Error Indicator Method Assessment

The final phase of this study involved assessing the quality of the Error Indicator (**EI**) for detecting map-matching gaps and comparing its performance against the binary classification models. The EI approach is

among the few unsupervised methods capable of detecting map-matching gaps without relying on ground-truth data. It was originally proposed by Berjisian and Bigazzi (2023) and consists of equation 5.5, as follows:

$$EI = 0.39LI' + 0.94ADE' + 0.67A' + 0.96DTW' + 0.74FD' \tag{5.5}$$

This expression utilizes normalised values for each component, corresponding to the similarity measures defined in Section 5.2.5. The LI component was obtained by computing the absolute value of the Length Index minus 1, as map-matched routes shorter than the GPS trace (LI <1) would otherwise contribute negatively to the EI result. It is important to note that the normalisation process was conducted based on the range of values for each dataset independently. In their original work, entries with EI values greater than 0.5 were flagged for manual inspection as potentially unreliable results. Table 12 presents the outcomes of applying this method to our datasets.

Table 12: Error Indicator Performance.

	Braga	**Amsterdam**	**Paris**	**Seville**	**Total**
Accuracy	0.77	0.63	0.77	0.55	0.64
Precision	0.81	0.79	0.72	0.84	0.79
Recall	0.76	0.52	0.96	0.19	0.48
F1 Score	0.78	0.63	0.82	0.31	0.6

Of the 518 total map-matching gaps, the EI correctly flagged 251 and missed 267, yielding an overall recall of 0.48. Additionally, 68 cases without any gap were wrongly flagged for manual inspection. This makes an overall precision value of 0.79, meaning that, on average, 1 in every 5 cases was flagged incorrectly for manual assessment.

Comparing the outcomes across datasets reveals substantial variation. For smaller datasets, the EI demonstrated good sensitivity in detecting gaps but generated more false positives. Conversely, with larger datasets, precision increased, but the sensitivity to detect gaps experienced a significant decline. This phenomenon is primarily attributable to the presence of outliers exhibiting larger discrepancies between the GPS trace and the map-matching result, as their extreme similarity measures tend to exert a strong negative influence during the normalisation phase. This distortion ultimately reduces the sensitivity of the EI to detect subtler gaps.

Furthermore, the performance of this method varied considerably across different gap categories. Table 13 displays the recall values for the five most common categories. These values can be interpreted

as the proportion of occurrences within a given category that were flagged for manual inspection, given that the EI solely performs detection and does not classify the nature of the occurrences.

Table 13: Error Indicator recall for the five most common categories.

	Braga			Amsterdam			Paris			Seville			Total		
Category	Cnt	TP	R	Cnt	TP	R	Cnt	TP	R	Cnt	TP	R	Cnt	TP	R
One-way	45	36	0.8	69	41	0.59	19	19	1	42	7	0.17	175	103	0.59
Competing-road	8	1	0.13	20	2	0.1	16	14	0.88	57	0	0	101	17	0.17
Missing-segment	30	28	0.93	2	0	0	1	1	1	45	17	0.38	78	46	0.59
Map-matching	23	16	0.7	1	0	0	2	2	1	41	10	0.24	67	28	0.42
GPS Error	2	0	0	18	14	0.78	7	7	1	11	5	0.45	38	26	0.68

Notably, the EI struggled significantly to detect "Competing Road" occurrences. For this specific gap type, the EI yielded a recall value equal to or below 0.13 across three datasets, with the "Paris" dataset being the sole exception. This underperformance is directly tied to the presence of outliers in larger datasets. In "competing road" gaps, the deviations between the GPS trace and the map-matched route are often subtle, causing their similarity measures to closely resemble those of correctly map-matched cases. Due to the normalisation process, extreme outliers mask these subtle differences, allowing them to pass undetected by the EI method. Consequently, even for gaps typically characterized by stark discrepancies between the GPS trace and the map-matching result — such as one-way violations or missing segments — the EI only achieved recall values of 0.17 and 0.38, respectively, for the Seville dataset.

Comparison between EI and ML approach

Comparing the performance of both methodologies reveals that the binary classification models significantly outperform the EI method. The top-performing binary classifier achieved an accuracy of 0.893 and an F1-score of 0.906, whereas the EI method yielded an average accuracy of 0.64 and an F1-score of 0.60. Furthermore, the EI method solely flags anomalies without providing any insights into their root causes. In contrast, despite the limited accuracy of our multi-class models, the categorical indications they provide remain highly valuable for GIS analysts and mapping communities responsible for updating OSM data. Another substantial advantage of our ML approach is its suitability for on-the-fly inference. Unlike the EI method, which requires dataset-wide normalisation of similarity measures prior to computation, our trained models can evaluate new individual instances independently. This characteristic makes the ML approach highly scalable and far better adapted for real-world production environments handling continuous streams of GPS data.

5.4 Summary

Detecting the gaps between real cycling traces and the existing road network models is an essential first step to improve these models and, consequently, offer better insights to decision-makers and more reliable services to cyclists. In this chapter, we assessed the feasibility of using machine learning classifiers to automatically detect and classify the cases where map-matching fails to properly map-match cycling traces into the OSM road network model.

Results show that binary classification models were able to identify map-matching gaps with good performance. The best classifier, XGBoost, obtained an accuracy of 0.893 and an F1 value of 0.906. The top-performing binary models significantly outperformed existing unsupervised baselines, specifically the EI method (Berjisian and Bigazzi, 2023). We observed that EI performance depends largely on the dataset. The sensitivity of this method decreases as the size increases due to a higher probability of existing extreme outliers.

We also trained several multi-class classification models. However, their performance was not very good. The best classifier, Random Forest without feature selection, achieved 71% accuracy. Despite being able to distinguish between cases with and without gaps, in most cases, it failed to classify those gaps according to their root causes.

Despite these limitations, the predictions estimated by these models can help stakeholders to find and correct the errors in the road network data model and to create an overview of the anomalous segments per city. In particular, practitioners and researchers can integrate the proposed approach into geographic information systems or spatial analysis pipelines to process cycling trajectories more accurately. This method can detect when the information is being assigned to the wrong segment, thus improving insights and suggestions given to decision-makers and to cyclists. Moreover, detecting when the GPS trace could not be map-matched onto the road can lead to the discovery of omission errors, thus contributing to its progressive improvement. Furthermore, this methodology can assist urban planners and municipalities in pinpointing unmapped segments and informal cycling routes, thereby informing future infrastructure investments.

Chapter 6

An Unsupervised Method to Identify Anomalies in Road Network Models for Urban Cycling

6.1 Motivation

In Chapter 4, we defined map-matching gaps as situations where a map-matched route fails to accurately represent the cyclist's actual path. Subsequently, in Chapter 5, we proposed a supervised method to automatically detect these gaps and classify them based on their probable root causes. However, analyzing trajectories in isolation yields a limited perspective on broader urban cycling mobility. When examining a single GPS trace, map-matching gaps may not be highly significant due to the presence of external confounding factors, such as GPS inaccuracies. Furthermore, an isolated map-matching gap may merely reflect an idiosyncratic detour—where a cyclist traverses an unconventional route rarely used by others—rather than a systemic infrastructural issue.

To move beyond the scope of individual trajectories and understand the meaning of map-matching gaps within large datasets of GPS traces, we introduce the novel concept of a map-matching anomaly in Definition 2:

Definition 2. *A map-matching anomaly is a systematic discrepancy between the road network data model and a recurrent pattern of GPS traces crossing a specific area.*

Such a discrepancy inevitably forces the map-matching algorithm to systematically generate gaps when processing traces in that specific region. When analyzing massive volumes of cycling data at a city-wide scale, these anomalies become substantially more relevant than isolated trace errors. While individual GPS errors introduce noise, they tend to be randomly distributed and ultimately cancel each other out when aggregating large datasets. Conversely, anomalies are spatially bound and represent systematic errors stemming directly from localized structural limitations within the road network model. Therefore, anomalies are neither a consequence of GPS noise nor a limitation of the map-matching algorithms themselves.

Even an optimal map-matching algorithm processing a flawless GPS trace would fail to produce a suitable representation, simply because the required physical route does not exist within the digital model. The inevitable output is a map-matching gap.

Figure 15 illustrates three scenarios corresponding to our anomaly definition. The figure demonstrates how a single continuous GPS trace (red dotted line) can fragment into multiple map-matched segments (blue arrows). The underlying base road network is depicted by thinner grey lines, while the dashed circles highlight the three distinct map-matching gaps. Although we represent only one gap per anomaly for visual clarity, real-world scenarios would exhibit a dense cluster of gaps concentrated around each anomalous zone.

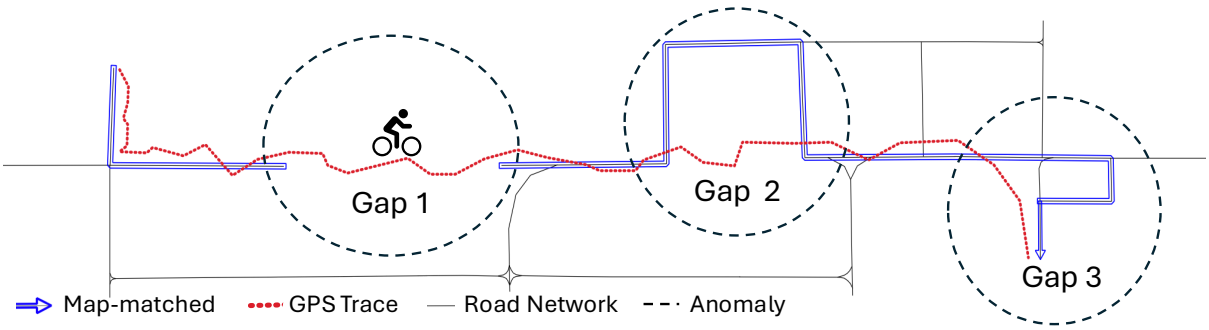


Figure 15: Examples of three map-matching anomalies.

In Figure 15, Gap 1 represents instances where the map-matching algorithm fails to find any match, resulting in an error. Gap 2 depicts scenarios where no feasible match exists, prompting the algorithm to generate a detour completely unrepresentative of the actual trajectory. Finally, Gap 3 illustrates cases where a geometrically appropriate candidate road exists, but the algorithm considers it ineligible for cycling (e.g., riding against traffic on a one-way street or cycling on a pedestrian-only walkway). In all these cases, an anomaly exists because in the road network model, there are no viable roads to represent the cyclist's path. This lack of representation will inevitably lead the map-matching process to produce systematic map-matching gaps.

If left unaddressed, these anomalies can significantly distort the interpretation of aggregate cycling activity. Such systematic misrepresentations inevitably lead to flawed assumptions and, consequently, misguided urban planning decisions. Conversely, by collectively analyzing these patterns, anomalies can serve as powerful indicators, revealing true cyclist behaviours and exposing structural flaws in current road network mappings.

This chapter proposes an unsupervised computational method to process large sets of cycling traces

to pinpoint urban areas prone to map-matching anomalies. Subsequently, it discusses how insights derived from these anomalies can empower municipalities to uncover novel, actionable perspectives on real-world cycling dynamics.

6.2 Methodology

The methodology proposed to identify map-matching anomalies is based on a large dataset of cycling trips recorded in Braga, Portugal. Figure 16 shows a heatmap with the spatial distribution of the GPS traces across Braga, where each red line represents a single GPS trace.



Figure 16: Heatmap with the spatial distribution of the GPS traces in the Bicification dataset.

These data were collected as part of a collaboration between Braga Municipality and the Bicification project[1]. In this project, registered users could receive monetary rewards for logging their bike trips in the city using a mobile application. This was a cycling incentive and a way to collect valuable and reliable trajectory data. During the observation period (May 31 to September 26, 2022), 262 registered users produced 16,292 GPS traces of their cycling trips, corresponding to 94,467 km of cycling trajectories in the streets of Braga. Braga is a mid-sized city with approximately 150,000 inhabitants. This large concentration of cycling trips in a relatively small region provided us with enough data points to assess

[1] https://bicification-project.eu/

cycling patterns and move beyond the circumstantial analysis of particular trips. This large concentration of cycling traces was very valuable for the goals of our study.

Figure 17 illustrates the proposed methodology for identifying map-matching anomalies. This process consists of four steps, namely, map-matching, gap extraction, anomaly identification, and anomaly analysis.

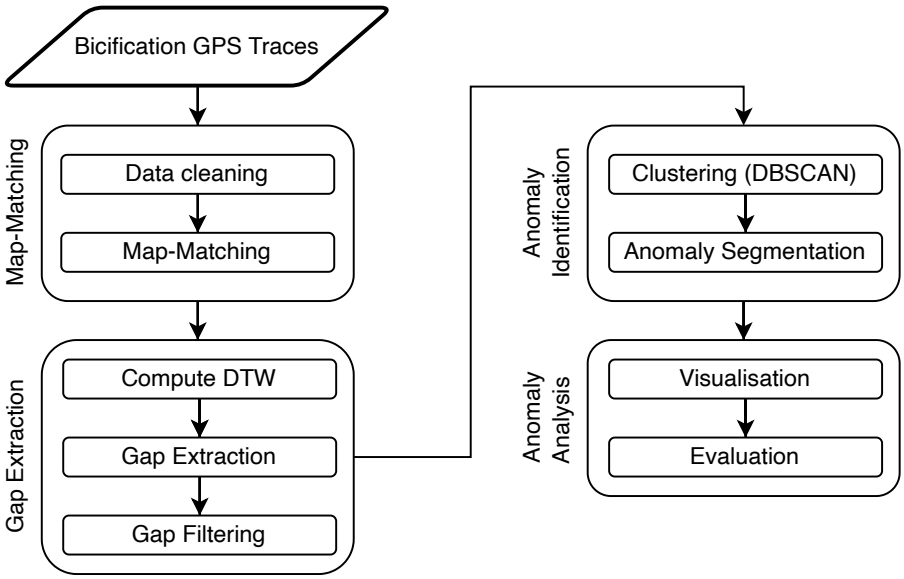


Figure 17: A four-step methodology for anomaly identification and analysis.

We started by cleaning the Bicification dataset and map-matched the eligible GPS traces to the OSM road network. Then, we analysed the map-matching results to identify the map-matching gaps. For this purpose, we used reference similarity measures to compare the GPS trace and its map-matched polylines and identify the portions where both geometries appear less similar. Finally, we applied a spatial clustering algorithm to identify the map-matching anomalies, the areas of the city where these gaps occurred systematically.

The next section presents a detailed description of each step. After each description, it presents the results of applying each step to the Bicification dataset. This approach provides essential context for understanding the subsequent steps. This structure facilitates comprehension of the logical flow of the proposed method.

6.3 Process and Results

To ensure the reliability of our data, we started by conducting a data-cleaning procedure to remove abnormal situations from the original dataset. We first excluded the GPS traces recorded outside our area of interest, specifically within the boundaries of Braga City. Then, to reduce the impact of low-quality GPS traces, we split any GPS trace where two consecutive points were at least 200 meters apart into separate traces. Such gaps in a trajectory often occur due to GPS signal loss or temporary interruptions in recording. Removing these gaps helps prevent false positives.

Results: Table 14 summarises the results of applying this data-cleaning procedure. In total, from the original 16,292 GPS traces, we excluded 1,595 abnormal GPS traces, and after the splitting step to the remaining 14,698 GPS traces, we obtained 16,746 GPS traces eligible to be map-matched.

Table 14: Results of applying the data-cleaning procedure.

Description	Count	%
Bicification GPS Traces	16,292	-
GPS Traces Excluded	1,594	9.8
Eligible GPS Traces Before Splitting	14,698	90.2
Eligible GPS Traces After splitting	16,746	-

6.3.1 Map-Matching

In the map-matching step, we aligned the eligible GPS traces onto the OSM road network utilizing the Mapbox algorithm, previously detailed in section 4.2.1. The decision to employ Mapbox's service over pgMapMatching was primarily driven by computational scalability. The pgMapMatching algorithm exhibited significant latency, often requiring several minutes to process a single trajectory — a limitation severely exacerbated by longer traces. Given the overall dataset volume, processing the entire dataset with pgMapMatching would be computationally prohibitive. Moreover, consistent with the preceding chapters, our focus is not on the quality or any particular properties of map-matching algorithms. We are specifically looking for map-matching anomalies. While the resulting set of anomalies identified could be slightly different using other map-matching tools or by different map-matching parameterisation, our map-matching anomaly concept is inherent to the road network data models and therefore should be largely independent from the specifics of any map-matching algorithm. In fact, such flexibility is expected. Regions may have

different rules for similar scenarios. For example, in some regions cyclists may be allowed to ride against traffic, whereas in others such behaviour is prohibited. Thus, what may constitute an anomaly may vary across regions. Therefore, the anomaly identification process should be adaptable to these perspectives, for example, by enabling the usage of different map-matching algorithms and parameterisation. In this case, we configured the map-matching service to consider the bicycle profile, which is described to avoid highways and give preference to streets with bike lanes. This mode also considers one-way roads and sidewalks eligible for matching cyclists' GPS traces.

To ensure optimal results in the map-matching process, we preprocessed the 16,746 eligible GPS traces to ensure that consecutive GPS coordinates were at least 10 meters apart. Additionally, due to the GPS coordinate limit per request, eligible GPS traces exceeding this limit were divided into segments of 100 points each, with the last segment potentially containing fewer points.

Results: This optimisation produced 43,761 segments, each corresponding to a map-matching request. Of those requests, 320 resulted in map-matching errors (e.g., an execution error where no route was returned), 37,998 resulted in one map-matched polyline, and 5,443 resulted in multiple map-matched polylines. Table 15 summarises the results of applying the map-matching procedure.

Table 15: Results of applying the map-matching procedure.

Description	Count	%
Eligible GPS Traces After splitting	16,746	-
Total Map-Matching Requests	43,761	-
Map-match: Error	320	0.7
Map-match: One Polyline	37,998	86.8
Map-match: Multiple Polylines	5,443	12.4

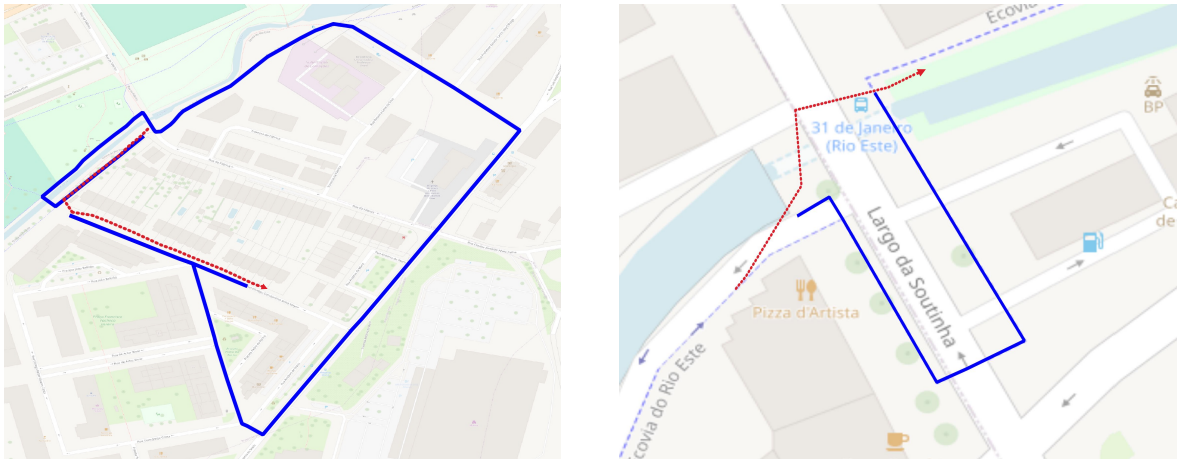
6.3.2 Gap Extraction

The gap extraction step aims at the unsupervised identification and extraction of the map-matching gaps. In cases where map-matching failed to produce any match, we straightforwardly considered them as a map-matching gap. The main challenge is to define rigorous criteria to identify the gaps corresponding to routes that misrepresent the source trace. To address this challenge, in the cases where the map-matching produced one or multiple map-matched polylines, we used the DTW algorithm (Müller, 2007) and the Optimal Warping Path (**OWP**) to compare the GPS trace with its corresponding map-matching

polyline and identify the segments of both geometries that are below a certain level of similarity. DTW is a technique to find the optimal alignment between two sequences (time series) (Müller, 2007) and is widely used in different fields, such as speech recognition, trajectory similarity (Tao et al., 2021), dynamic signature verification (Jiang et al., 2022), and the stock market (Han et al., 2020). For two sequences, the DTW value is well-defined. However, they can have different warping paths. The OWP is a path that optimises the sum of all local distances. A local distance refers to the distance between two elements of each sequence based on a cost function. The cost function can be defined according to the purpose of the study. In our study, the sequences are composed of WGS84 coordinates. Therefore, we used the Haversine distance formula as the cost function. It is important to note that before computing the DTW, we interpolated points every 2 meters in the GPS and map-matched geometries. This solution reduced the impact of long intervals between consecutive points, which would otherwise lead to high local distances during the OWP computation, even when both geometries appear similar.

After computing the OWP, we classified map-matching gaps as corresponding to cases where the local distance was higher than 10 meters for at least 5 consecutive points. Where a map-matching gap was detected, we registered the "gap segment" and its "gap trace". The former corresponds to segments in the map-matching response, and the latter corresponds to the part of the source GPS trace that produced the gap.

Results: In total, we extracted 153,170 map-matching gaps. Figure 18 presents 2 examples of those gaps. The red dotted line represents the gap trace, while the blue represents the gap segment.



(a) Missing Road (b) Unconventional Road Crossing

Figure 18: Example of two different map-matching anomalies.

Figure 18a illustrates a scenario where the cyclist used a ramp not mapped in OSM to transition from a cycleway to a residential road. The map-matching algorithm attempted to find a path connecting the

cycleway and the residential road. Figure 18b shows a situation where the cyclist crossed the road outside the designated area. The map-matching algorithm snapped the GPS trace to the nearest zebra crossing.

Gap Filtering

A large number of the map-matching gaps produced in the previous step correspond to very short route segments. To avoid many false positives and because we aim to focus on an aggregate perspective, we decided to remove gap where (1) the gap traces were shorter than 40 meters or (2) where the ratio between the length of the gap trace and the gap segment was between 0.7 and 1.3. We consider false positives the cases in which the map-matching algorithm successfully snapped the GPS into the correct road, but GPS interferences created a deviation in the GPS trace sufficient to induce our algorithm to consider it as a gap. We also removed gaps where the gap trace consisted of the beginning or end sections of the GPS trace. These sections can easily lead to map-matching errors due to the calibration of the GPS sensor or the map-matching algorithm itself. While GPS problems could be addressed in the data cleaning step, the interferences from map-matching calibration can only be detected in this phase.

Results: After the gap filtering procedure, we ended up with 29,151 eligible map-matching gap. Table 16 summarises the results of applying the gap extraction and filtering.

Table 16: Results of applying the gap extraction procedure.

Description	Count	%
Total Map-Matching Requests	43,761	-
Total Map-matching Gaps	153,170	-
Eligible Map-Matching Gaps	29,151	19.0

During our analysis, we observed that most gaps with a ratio greater than 1.3 were removed, as they involved the beginning or end sections of the GPS traces.

6.3.3 Anomaly Identification

In the map-matching anomaly identification step, we aimed to identify the regions with a high density of gaps, situations matching our concept of map-matching anomalies. The goal is to provide an aggregate view of these gaps, focusing on their quantitative relevance, especially on truly systematic cases, indicating an underlying problem. Addressing these issues could have a significant impact on a large number of cyclists. A large cluster may not only be large due to having many gaps but also because it aggregates

several smaller clusters that, despite having different characteristics such as movement patterns, share the same underlying cause. For this purpose, we applied the DBSCAN[2] clustering algorithm to identify the map-matching anomalies. To apply this procedure, we selected the midpoint in each gap trace, since the midpoints of gaps passing through the same area tend to converge to the same location.

The rationale behind this procedure is that in zones where the road network does not represent the cyclists' paths, the map-matching is likely to produce recurring gaps, leading to areas with a high concentration of gaps. In contrast, sporadic events, such as strong GPS interferences or map-matching malfunctions, mainly result in spatially dispersed gaps. As DBSCAN groups data points based on spatial proximity, it allows us to mitigate the impact of sporadic events and highlight the most common gaps impacting most cyclists.

DBSCAN requires two parameters: the minimum number of points composing a cluster and the distance threshold. After some experimental iterations, we selected a fixed distance threshold of 15 meters and a minimum number of 20 points. However, the optimal value for these parameters may vary, depending on the number of gaps extracted, especially the minimum number of points. As the number of gaps increases, the minimum point threshold should be adjusted accordingly to prevent the formation of overly large and dispersed clusters by the DBSCAN algorithm.

Results: Table 17 shows the results of applying DBSCAN. In total, using the 29,151 eligible map-matching gaps, DBSCAN formed 157 clusters. In other words, it identified 157 anomalies. The clustering step excluded a total of 14,121 gaps.

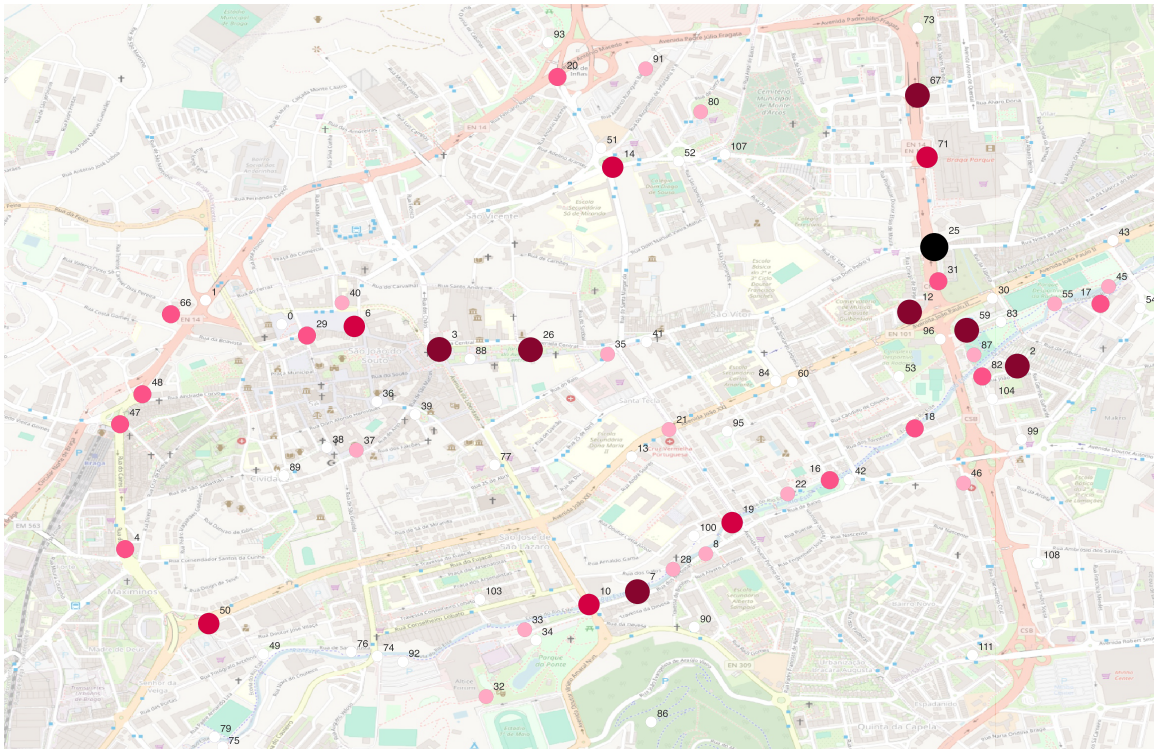
Table 17: Results of applying DBSCAN to the eligible map-matching gaps.

Description	Count	%
Eligible Map-Matching Gaps	29,151	-
Gaps Ignored During DBSCAN	14,121	48.4
Anomalies Identified	157	-

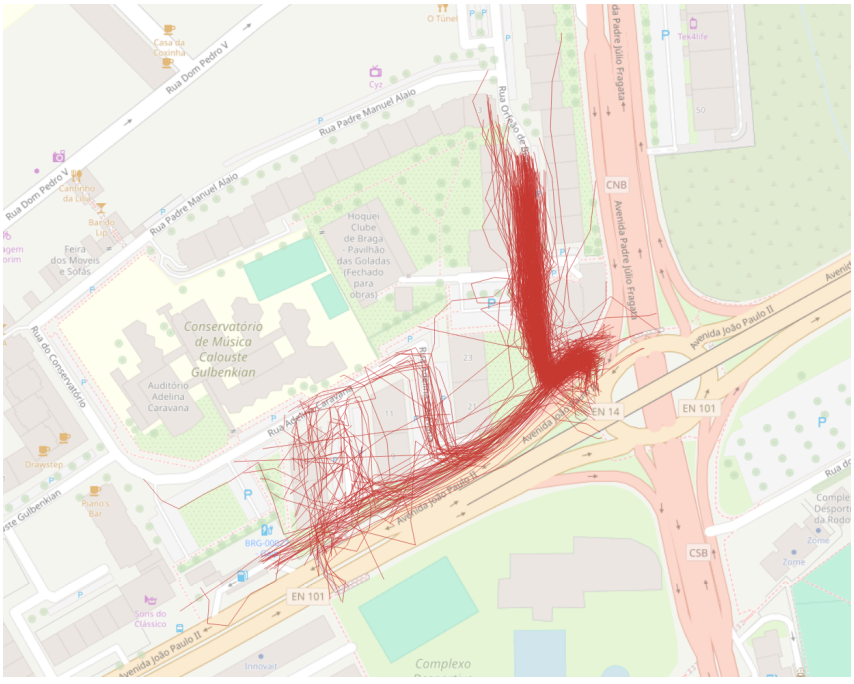
Figure 19a shows the spatial distribution of anomalies across the city of Braga. Each anomaly is represented by a circle, whose size and colour represent the number of gaps in each anomaly. As the size of the circle increases and the colour darkens, so does the number of map-matching gaps forming the anomaly. The number near each circle represents a unique ID to identify each cluster. This ID was generated automatically during the clustering process. Figure 19b shows the map-matching gaps that were assigned to the anomaly with ID 12. In this example, we can observe a clear and predominant pattern

[2] https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html

near the roundabout, where the density of gap traces is the highest. We can also observe that there are other movement patterns with less representation.



(a) The spatial distribution of anomalies.



(b) An example of an anomaly ID12.

Figure 19: Anomaly Identification results.

Figure 20 shows the distribution of the number of gaps per anomaly. Analysing this distribution, we

could verify that of the 157 clusters obtained, 125 clusters had fewer than 100 gaps. On the opposite side, 9 clusters contained more than 420 gaps. In particular, the two most significant clusters contain more than 1,000 gaps. Although anomalies are recorded in various areas of the city, this relative distribution shows that certain areas require greater prioritisation. Targeted interventions in these high-density clusters have the potential to benefit a substantial number of users.

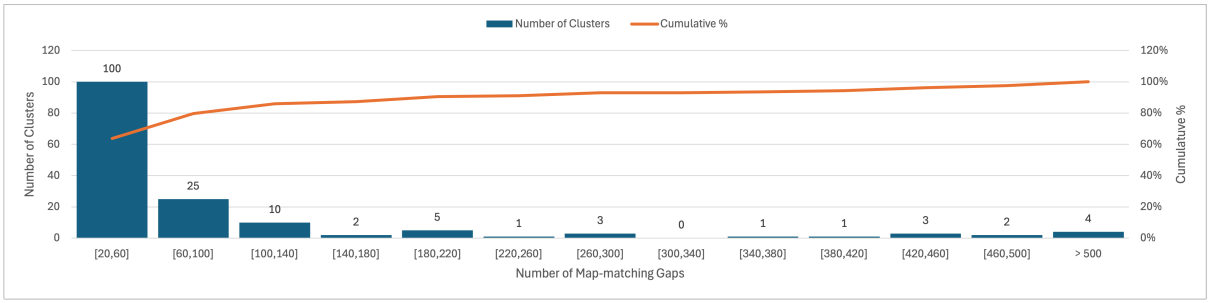


Figure 20: The distribution of the number of gaps per anomaly.

Anomaly Segmentation

The DBSCAN clusters corresponding to the anomalies were formed solely based on the geographical location of the midpoints of gap traces. Consequently, an anomaly may include gap traces with similar locations but differing movement patterns (e.g., direction, origin, destination, distance), as depicted in Figure 19b. In some cases, these differing patterns jeopardise the interpretation process. To simplify the analysis of an anomaly, we could visualise each gap trace individually. However, searching for similar or different patterns within an anomaly containing thousands of gaps with varied movement patterns is time-consuming and inefficient.

To calculate these trace patterns, we implemented an additional step using the DTW measure to subdivide the gap traces into groups matching the same movement pattern, as described in algorithm 1. This algorithm takes as input a set of gap traces that constitute an anomaly and groups them into one or more patterns. It has three parameters: the minimum number of gap traces that have a match (M), the maximum DTW value to consider two gap traces "similar" (T), and the number of "similar" gap traces to consider when computing DTW (N).

During this study, we empirically explored various parameter configurations, specifically testing $M \in \{3, 5\}$, $T \in \{15, 25, 30\}$, and $N \in \{10, 12\}$. It should be emphasized that the primary objective of this segmentation step was not to perform a rigorous quantitative hyperparameter optimization, but rather to heuristically partition the data in a manner that maximizes the interpretability of the subsequent cluster

Algorithm 1: The algorithm to divide an anomaly into different gap trace patterns.

Function DivideCluster (C, T, M);

Input : Set of gap traces C, Threshold T, Minimum Matches M

Output: Sets of gaps traces with similar movement patterns

begin

- **while** *there are gap traces left in C* **do**
  - Select 1 gap trace (A);
  - **for** *gap trace A' in C* **do**
    - Compute DTW(A, A') and divide by the minimum number of points of both geometries;
  - Select up to N gap traces on which the resulting value is below a threshold T;
  - **if** *number of selected anomalies* $<$ *M* **then**
    - Remove A from C;
    - Jump to the beginning;
  - Px $\leftarrow$ a new empty list to store the gap traces of the new trace pattern. x represents the trace pattern number;
  - **for** *Gap A" in C* **do**
    - **for** *Gap N' in N* **do**
      - Compute the DTW (A", N') and divide by the minimum number of points of both geometries;
    - **if** *A" has a value below T with more than M gap traces* **then**
      - Add A" to Px;
      - Remove A" from the cluster C;
- Return the resulting gap trace patterns;

analysis.

Results: Figure 21 shows the number of gap trace patterns obtained per cluster. The image shows that the majority of the clusters (64%) have 1 or 2 trace patterns.

Figure 22 provides 3 of the patterns obtained from applying the algorithm 1 to the anomaly ID 12. Each line represents a gap trace, and each point represents the start location of each gap trace.

By applying the algorithm to other anomalies and visually inspecting the resulting patterns, we could observe that the two patterns with the most gaps usually represent the same path but in opposite directions,

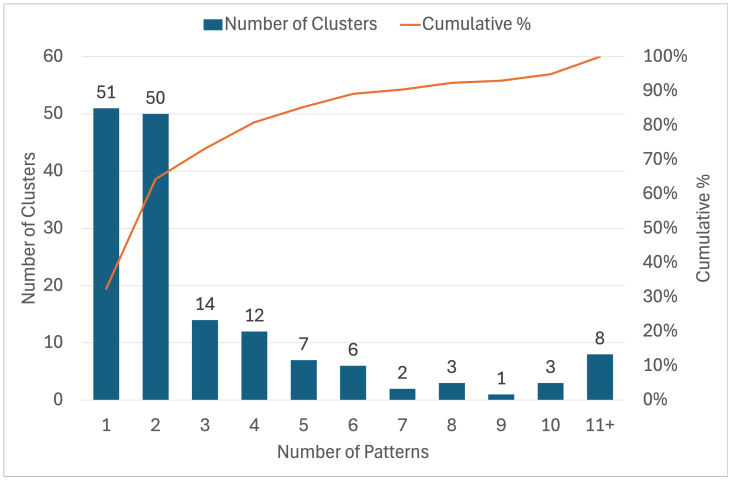


Figure 21: Anomaly Segmentation Results: The number of gap trace patterns per anomaly.

while the remaining trace patterns may represent small variations in the path or a completely different path.



(a) Pattern 1 (b) Pattern 2 (c) Pattern 3

Figure 22: Example of three patterns emerging from segmenting an anomaly.

6.3.4 Anomaly Analysis

To assess the result of our technique, we analyse the results produced in the previous steps. For this purpose, we randomly selected 50 anomalies and visually inspected them to assess their cause of failure.

Table 18 shows the results of this assessment process.

Table 18: Evaluation of the anomalies.

	Count	%
Total Anomalies	157	100%
Anomalies Inspected	50	32%
True Anomalies	42	84%
GPS Errors	7	14%
Internal Errors	1	2%

In total, we identified 42 cases corresponding to true anomalies, cases in which the map-matching algorithm failed to produce the correct representation of cyclists' traces due to limitations in the road network model. This corresponds to a 84% precision rate. This was especially common when cyclists used sidewalks to avoid travelling along or against traffic, changed roads at unconventional locations, or rode through parks and open spaces.

Another significant source of anomalies was the behaviour of the Mapbox map-matching algorithm, which, when configured for a cycling profile, excludes roads tagged as "trunk" from the set of eligible paths for cycling. In such cases, the algorithm either failed to return any match or inaccurately snapped the route to the nearest eligible road segment.

We identified 7 cases where map-matching failed due to strong GPS interferences. Some of them were caused by stay points, situations where the cyclist remained stationary at a location for a period of time. The data cleaning process did not account for these cases, but they could be mitigated by applying specific pre-processing techniques.

We also identified one case where the error was due to the map-matching algorithm. Specifically, this case was caused by two parallel roads. When there are two or more parallel roads nearby, even small interferences in the GPS signal can lead the map-matching algorithm to errors, and these types of errors are difficult to identify just by comparing the geometry of both polylines. Even when visually inspecting the results, without the ground truth about the cyclists' paths, we cannot assert whether they rode by the road along motorised traffic or used the sidewalk.

6.4 Overcoming Anomalies to Improve Cycling

In the previous sections, we addressed the objective of developing an unsupervised method to extract map-matching gaps and identify anomalies in a city. In this section, we revisit the second objective by analysing how map-matching anomalies can help cities uncover novel insights into cycling behaviour. We begin by breaking this research objective into two perspectives: The macro or global analysis, a city-scale exploration of cycling behaviour, and the micro or local analysis, a detailed investigation of specific locations or individual cyclists.

Macro analysis: In a macro analysis, the stakeholders may wish to obtain an overview of the severity of map-matching anomalies, monitor how the characteristics of anomalies (e.g. the number of gaps) vary during the day, across city areas or different cities, and benchmark the overall alignment between cycling activity and the road network model. They may wish to examine how anomalies relate to road accidents, road types, motorised traffic intensity, district characteristics (e.g., residential or commercial) or critical locations (e.g., schools or hospitals). This broad perspective enables cities to identify priority areas and benchmark the design of their cycling infrastructure.

To conduct macro-level analysis, it is ideal to establish a set of metrics that can be computed automatically on a large scale. This approach ensures that the methodology is replicable across different cities, allowing effective benchmarking and comparison. Some of these metrics may be computed at the city level or at a finer spatial scale. The latter describes a neighbourhood or district in finer detail. Additionally, it should be possible to aggregate them into a global metric to represent the overall conditions of the city. Examples of metrics may include the average number of gaps per cycling trip, the average number of anomalies per neighbourhood, or the average detour ratio per anomaly. These metrics help identify how frequently cyclists search for informal travelling paths and evaluate the impacts of current cycling conditions and the potential for improvement.

Micro-analysis: In a micro-analysis, the focus shifts to identifying the root causes of specific anomalies to plan effective solutions. To achieve a systematic resolution of the main causes, stakeholders may need to combine both quantitative and qualitative methods. In other words, they need to integrate both metrics and visualisations to provide comprehensive insights.

The metrics will enable the establishment of a ranking based on the impact or potential improvement. This could help stakeholders obtain systematic and actionable feedback, allowing them to prioritise efforts effectively. These metrics may quantify factors such as the volume of gaps within an anomaly, the frequency of gap occurrences, or the impact of an anomaly on cyclists in terms of safety, detours, and

comfort. The visualisations provide a qualitative perspective, allowing for comparison of anomalies with the road network and local knowledge.

6.4.1 Workflow Analysis

In this work, we explored four visualisation approaches to perform the analysis. Figure 23 illustrates the workflow used for analysing a hotspot and identifying its root cause. This workflow has 4 steps, namely the global, anomaly, pattern and gap views.

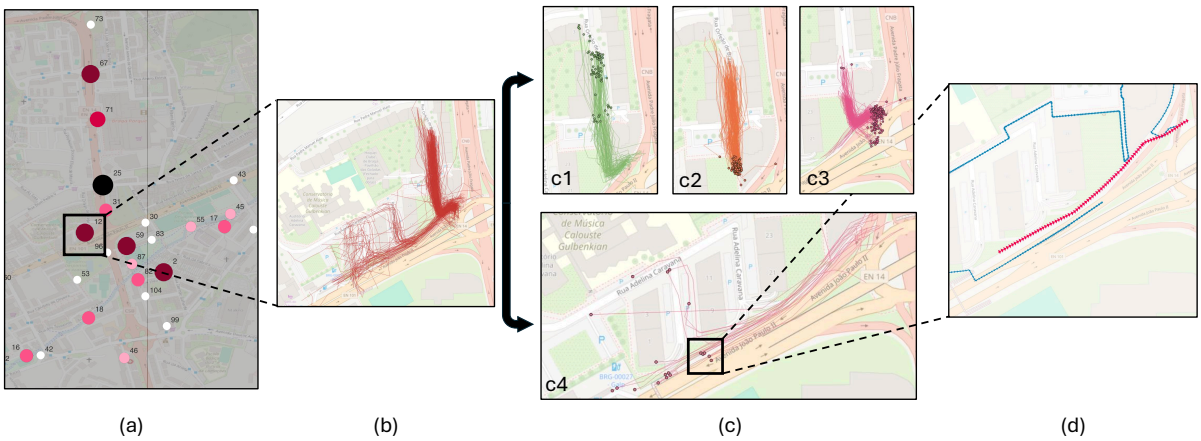


Figure 23: Different levels of analysis: a) global, b) anomaly, c) pattern, and d) gap.

Step 1: Global View: We usually start by visualising the locations and the volume of anomalies on the map. Figure 23(a) shows an example where each circle represents the location of an anomaly. The colour and size of a circle indicate the number of gaps. This map highlights the spatial distribution of anomaly hotspots, allowing stakeholders to identify critical areas in need of intervention.

Step 2: Cluster View: Then, after identifying critical hotspots, we zoomed into individual anomalies. Figure 23(b) shows a close-up of an anomaly composed of thousands of anomaly traces. This visualisation reveals several trace patterns emerging from the data, indicating the existence of a recurring problem. However, distinguishing between these patterns can be challenging, especially when trying to characterise the direction or the beginning or the end of the anomalies.

Step 3: Pattern View: To simplify the visualisation, we segmented the anomaly into isolated gap trace patterns. Figure 23(c) shows four distinct patterns, with the points representing the starting positions of each anomaly trace. This step clarifies the movement behaviours within the cluster.

Step 4: Anomaly View: Finally, we visualise individual anomalies. Figure 23(d) shows a single

anomaly trace along with its corresponding map-matching anomaly.

By applying this workflow, we were able to identify the root cause of several anomalies. In the example provided by Figure 23, we identified several errors in the OSM road network, namely, OSM had no representation for the walkway along the main road, and some connections between residential streets and pedestrian facilities were also missing. These omission errors caused the map-matching algorithm to create detours to connect the map-matched segments and could negatively impact the quality of route suggestions, especially for pedestrians. The probable root cause of this error was the cyclist's attempt to avoid sharing the road with cars at high speeds, which forced the cyclist to use a pedestrian facility, which is an illegal behaviour.

These insights can be valuable to various stakeholders: First, they can help OSM collaborators improve OSM road network data by highlighting missing roads or incomplete tags in the OSM database. Secondly, they can help urban planners detect zones needing improvement, namely, where cycling has less support and where cyclists engage in illegal behaviour. Finally, addressing the causes of these anomalies can help cyclists move better across the city by providing new paths, which other cyclists use to avoid unfavourable situations.

6.5 Summary

We started this chapter by introducing the concept of map-matching anomaly as a systematic mismatch between cycling traces and the road network data model to which the traces are expected to be matched. Contrary to sporadic map-matching errors, e.g., caused by GPS interferences, these anomalies are systematic gaps likely to occur with all similar traces and accumulate at the specific location where the source problem may exist.

Detecting the anomalies between cycling traces and the road network data model is an essential first step to improve these models and, consequently, offer better insights to urban planners and better routes to cyclists. For this purpose, we then proposed a four-step methodology to automatically identify urban zones where map-matching gaps are prevalent and systematic, which correspond to the concept of map-matching anomalies. In the first step, we preprocess and map-match GPS traces of cycling activities into the OSM road network. In the second step, we use the DTW to compare the GPS traces with the corresponding map-matched segment and extract map-matching gaps. In the third step, we apply a DBSCAN to cluster the midpoints of each gap to identify the anomaly locations. Finally, in the fourth step, we visualise the results.

We applied our methodology to a dataset containing 16,292 traces in the city of Braga, and it achieved 84% precision in identifying locations prone to map-matching anomalies. By analysing the resulting anomalies, we identified several cases where the OSM was incorrect or incomplete and identified several locations where a deeper intervention is needed to improve the road network infrastructure to meet cyclists' desire lines.

Finally, we discussed how the information about the anomalies and map-matching gaps can be used by cities to assess, compare, and improve their support for cycling. We defined a workflow that cities can follow to first obtain an overview of critical locations with more potential for improvement and then analyse concrete cases to identify and solve their root cause. Namely, the information about the anomalies can be used by OSM collaborators to improve the OSM road network and by urban planners to target their efforts in improving mobility.

Chapter 7

Stakeholders' Perspectives on the Gaps Between Cycling Traces and Their Representation on OSM

7.1 Motivation

In the previous chapter, we developed an unsupervised method for identifying city-scale areas prone to map-matching anomalies. These anomalies reveal information about the discrepancies between the road network and cycling practices. Identifying these anomalies and characteristics thus can be very valuable to cyclists, urban planners and OSM mappers.

In this chapter, we aim to gather new insights into the perspectives of the various stakeholders involved, specifically municipalities, cyclists and OSM mappers, about how the information about those behaviours can help improve the cycling experience and urban mobility in general. We build on the assumption that the core issue is not technical. It is mainly an issue of diverse and contradictory views about the same problem, which results in a conundrum. Our goal is to understand the perspectives of these key stakeholders in order to identify common ground among their seemingly conflicting positions. This study may open new avenues for reconciling some of these differences and unlock OSM's full potential as a key enabling technology for cycling systems.

7.2 Research Methods

This study is designed to characterise how the stakeholders involved in cycling mobility perceive the gaps between cycling routes and the road network model and explore how the knowledge about those gaps may help to reconcile their views.

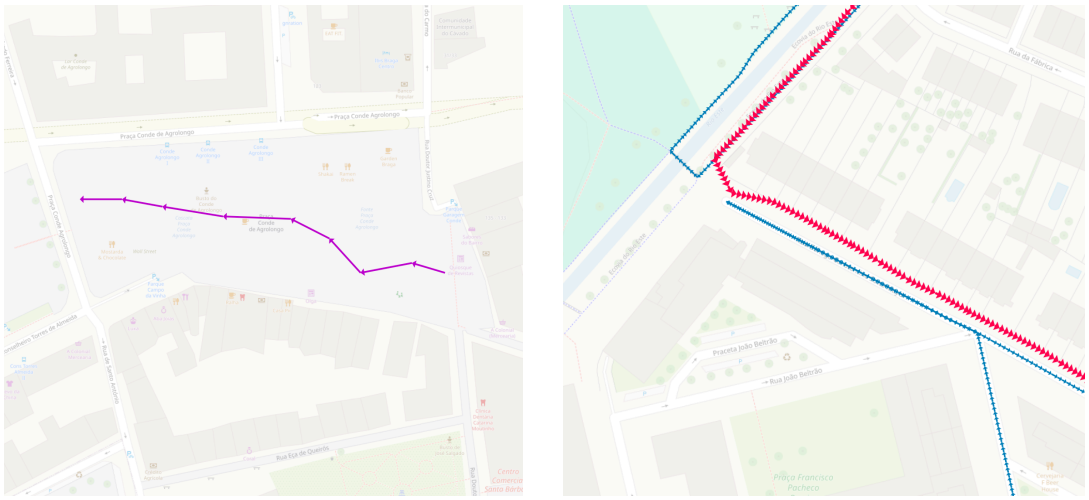
The research methodology comprises two primary components. The first involves the identification and categorisation of specific map-matching anomalies. The second employs a multi-method qualitative

approach, consisting of semi-structured interviews and focus group sessions conducted with three primary stakeholder groups: cyclists, OSM contributors, and local mobility authorities. During the sessions, the anomalies selected provided important focal points for the discussions between the various stakeholder groups, helping to uncover possible bridges and confront key differences between those stakeholder groups or even within the participants of the same group. The results of the interviews and focus groups were then used as input to a qualitative data analysis, from which we were able to identify relevant emerging topics and aggregate the results as key findings for this work.

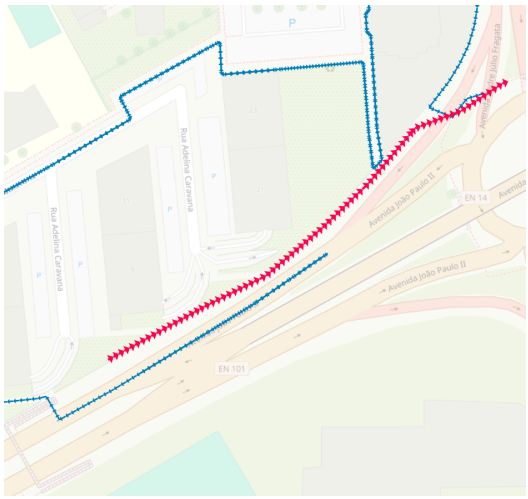
7.2.1 Gap Cases

To foster a rich and focused discussion, we selected six map-matching gaps representative of common scenarios in which OSM fails to represent real urban cycling traces. These gaps were sampled from the anomalies identified in the previous chapter. While all selected cases were related to Braga, they illustrate generic situations that can be observed in most urban environemnts. Figure 24 illustrates the selected examples. The red arrow lines represent the gap traces extracted from the GPS trace, while the blue arrow lines represent the gap segment extracted from the map-matched route.

Case A: According to the GPS signal shown in figure 24a, the cyclist crossed through the centre of a square. In this case, the map-matching algorithm failed to match this route segment. Consequently, the original GPS trace is represented in purple to denote the absence of a corresponding map-matched route. In other similar situations, the map-matching algorithm snapped the GPS trace into the perimeter of the square. This example illustrates common scenarios in which cyclists traverse open areas, such as parks or squares, where only the boundaries are represented in OSM.



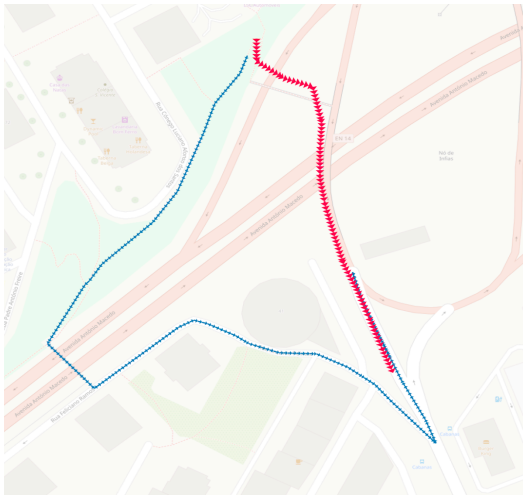
(a) Open Area (b) Missing Segment



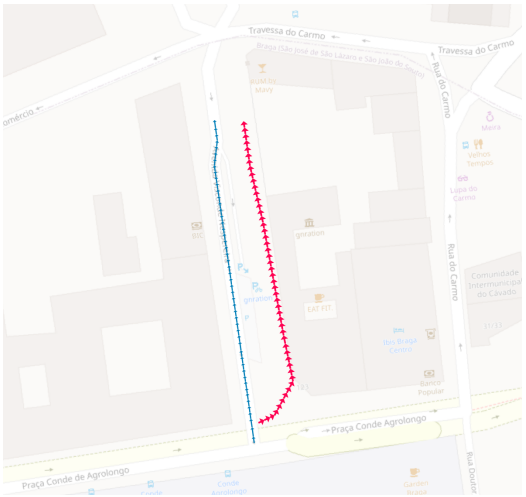
(c) One Way



(d) Unconventional Crossing



(e) Map-Matching



(f) One Way

Figure 24: Examples of six map-matching gaps.

Case B: According to the GPS trace in figure 24b, the cyclist used a small ramp to change directions, going from a cycleway to a residential road. However, this ramp is not represented in OSM. Consequently, the map-matching algorithm tried to find the nearest path connecting the cycleway to the residential road, resulting in a route approximation around the block. This is representative of cases where cyclists use roads omitted in OSM.

Case C: Figure 24c shows a complex but typical case. In this example, the cyclist used the sidewalk to avoid riding along high-speed cars on a major and busy road. Considering the direction of movement, the cyclist went against traffic. By manually inspecting the results, we inferred that map-matching failed for two reasons: Firstly, the sidewalk was only partially represented in OSM and was omitted along the segment on which the map-matching failed. Secondly, the Mapbox map-matching algorithm does not

consider main roads (e.g., tagged as "trunk") eligible to map-match cycling traces. Consequently, the map-matching initially snapped the GPS trace into the main road (although the cyclist used the sidewalk). Then, when the cyclist reached the road tagged as "trunk", map-matching tried to find an alternative path to connect the last map-matched point to the roundabout sidewalk. This resulted in a path around the block.

Case D: Figure 24d is representative of cases where cyclists cross the road in unconventional places. In this case, to continue his route by the cycleway and avoid a short detour to use the crossing section, the cyclist crossed the road in a zone where no official crossing exists. By snapping the GPS trace into the OSM network, map-matching made a detour along the zebra crossing.

Case E: Figure 24e illustrates a common situation where cyclists ride along dangerous roads characterised by high-speed vehicular traffic. In this case, the cyclist travelled along a busy road, tagged as "trunk", which was not considered for map-matching purposes. This road segment is located at one of Braga's major intersections, known for its fast traffic flow and high vehicle volume. We observed multiple occurrences where cyclists followed the same path in the opposite direction, suggesting they likely used the adjacent sidewalk. However, the sidewalk is extremely narrow and not suitable for safe use, particularly when two users need to pass each other.

Case F: Figure 24f represents a typical scenario where cyclists use wide sidewalks to avoid unnecessary detours. In this case, the cyclist used the sidewalk to, first, avoid a longer detour around the block and, second, to avoid riding against traffic, as the shortest road was unidirectional. The sidewalk is not particularly busy and is relatively wide. However, since it is not represented in OSM, the map-matching algorithm inferred that the cyclist travelled along the adjacent road segment.

7.2.2 Focus Groups and Interviews

After selecting the map-matching gaps representative of anomalies in Braga, we proceed to schedule and perform the sessions with the different stakeholders.

We considered three main stakeholders: cyclists, OSM mappers, and local mobility authorities. To represent cyclists, we recruited participants who frequently use bicycles to move around the city daily. These cyclists were recruited in collaboration with a local cycling mobility association. To represent mappers, we selected participants who contribute to OSM by proactively inserting new elements or updating existing ones. These participants were recruited within the local community of OSM mappers. Finally, to represent local authorities, we selected participants actively involved in planning the cycling road network or defining cycling policies within their cities. Even though participants were selected to represent

one of these stakeholder groups, there were some cases where representatives of the local authorities or mappers were also cyclists, at least occasionally. We asked them to focus on their primary role, but we assumed that their experience on both sides would be a valuable contribution to the study.

After identifying interested participants, we conducted the following sessions:

One in-person focus group with five cyclists: The session aimed to capture why cyclists desire informal paths, their daily challenges, and what type of routes could improve their cycling experience.

One online focus group with five mappers: The session aimed to capture the mappers' perspectives on informal paths and the feasibility of including them in the OSM database. We also focused on their most common challenges when mapping cycling routes.

Four in-person solo interviews with local authorities: During each session, we aimed to assess their awareness of the problem, analyse the feasibility of adapting these paths to accommodate cycling traffic, determine whether the information about those behaviours should be made public to other stakeholders, and explore potential collaborations with OSM mappers. The rationale for conducting individual interviews rather than focus groups was twofold. First, we had to recruit local authorities from different cities. In each municipality, there were no more than one or two potential participants - i.e., individuals with direct responsibility over the decisions under study. Second, due to the politically sensitive nature of some of the topics in this study, we feared a group setting could fail to provide the safe and controlled environment needed to elicit honest and open responses.

Despite the different formats, we followed a very similar structure for all the sessions, comprising three sequential phases: introduction, main discussion, and closure. The focus groups were planned for 90 minutes, while individual interviews were limited to 60 minutes.

Introduction: In the introductory phase, we started by explaining the problem and objectives of our study. We informed participants that the sessions would be recorded and asked them to express their consent. Then, we started the recordings and asked several introductory open-ended questions to assess the participants' awareness of the problem, their perspectives about why these informal paths are used, and whether the information about their usage should be made public to other stakeholders.

This phase included the following questions:

1. What are your views on informal paths used by cyclists?

2. What drives the demand for informal cycling paths?

3. Should information about informal cycling paths be shared with other stakeholders? If so, which stakeholders should be informed, and why?

4. Do you think informal paths should be added to OSM? If so, how should they be represented, and under what conditions? If not, why?

Main discussion: In the main discussion, we presented the six map-matching gaps and incentivised a short discussion around each case. For each situation, we provided a brief description of the cyclist's trace and the characteristics of the surrounding area. Then, we asked several questions to assess how each stakeholder perceived the case, whether the information about the path should be made public, and if so, how it should be represented in OSM.

This phase included questions such as:

1. What factors might have influenced the cyclist to choose this specific path?
2. What specific actions could be taken to resolve this gap?
3. Do you think cyclists should be redirected to another road, or is the chosen path more suitable?
4. Should the area be redesigned or adapted to legally accommodate bicycle traffic?
5. Should this information be made available to other cyclists, and if so, under what conditions?
6. Should this information be incorporated into OSM? If so, how should it be represented?

Closure: In the final phase, we presented an exaggerated scenario in which each stakeholder's perspective was taken to an extreme. In this hypothetical situation, we assumed the following: local authorities would strictly enforce official cycling paths, prohibiting cycling anywhere else; cyclists would always choose the most comfortable and direct routes, even if they were illegal; and OSM mappers would always map the paths that most accurately reflected the actual routes made by cyclists. Considering this scenario, we asked participants how such differences could be balanced. Then, we presented a set of final questions to wrap up the discussion, reviewed the main findings, and assessed whether the participants changed their point of view during the session.

During this phase, we asked questions such as:

1. (All) What are the main challenges you face when cycling, planning cycling paths, or mapping?
2. (Cyclists) Do you use navigation tools to find route suggestions? How often do you use them?
3. (Cyclists) Do you feel that route suggestions from navigation tools match the routes that you usually take?

4. (Cyclists) What type of solution or information would help cyclists discover new routes?

5. (All) How can OSM balance the perspectives and needs of different stakeholders?

6. (Mappers) What type of solutions could benefit mappers or enhance their mapping experience?

7. (Mappers and Local Authorities) How would you describe the interactions between the OSM community and municipal authorities?

Some of the questions were slightly adapted accordingly to the session, while others were asked only to one or two stakeholders groups.

7.2.3 Qualitative Analysis

The next step was the qualitative analysis of the results from stakeholder sessions. The recordings of the interviews and focus groups were transcribed and anonymised. Participants were identified using letter-based codes indicating their stakeholder group: C for cyclists, M for OSM mappers, and A for local authorities. The transcriptions were analysed using NVivo[1], a qualitative data analysis software. The analysis followed an unstructured coding process to identify emerging topics relevant to our research questions, producing 201 coded segments associated with 21 unique codes. These codes were progressively consolidated and aggregated into a set of emerging topics. These topics were then arranged into two major dimensions. The first dimension encompasses topics focusing on the root causes (mostly technical) behind the gaps between real cycling routes and their representation in OSM, factors that are intrinsic to OSM technology and related practices. The second dimension encompasses topics focusing on the diverse ways in which stakeholders perceive these gaps and how this influences their conceptualisation of possible solutions. Figure 25 summarises the main topics emerging under these two dimensions.

These dimensions and their key topics provide relevant insights about the many perspectives of the three stakeholder groups in our study. The next two sections will provide a detailed analysis of these topics. The results presented in Section 7.3 focus on the limitations of OSM in representing cycling paths. The results in section 7.4 focus on the politics of route representation and the different interpretations that the various stakeholder groups can have about the same concrete cases.

[1] https://lumivero.com/products/nvivo/

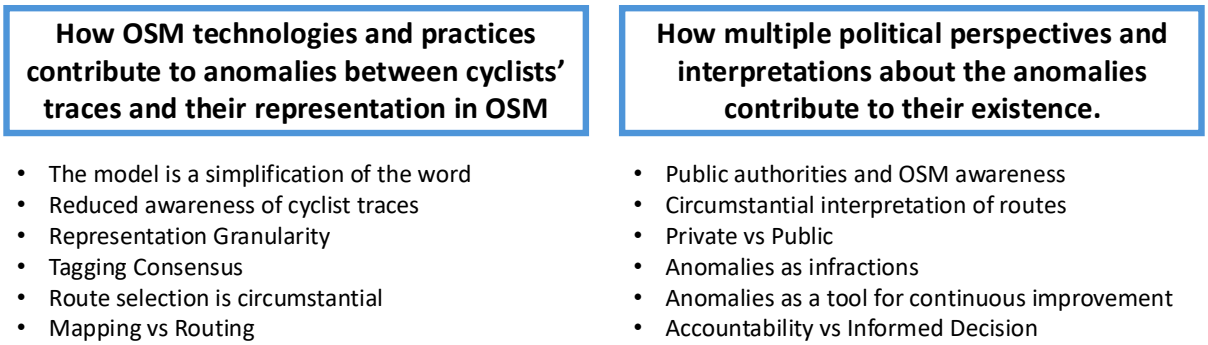


Figure 25: Dimensions emerging from the qualitative analysis.

7.3 Limitations of OSM Technology and Mapping Practices

The first dimension of analysis is related to the root causes of these discrepancies between common cyclists' traces and their representation in OSM. Our analysis showed that these discrepancies are mostly caused by OSM technical limitations in representing reality and accommodating different routing profiles and by people's ambiguity in understanding the roles of mapping and routing.

7.3.1 The Model is Only a Simplification of the World

The results show that a significant part of the problem is somehow linked to the broader challenges of world representation. Like any other model, the OSM road network model is just a simplified representation of a complex world. In most cases, they are, at best, based on partial data. Furthermore, as we will see in the next section, they must be interpreted by humans, who bring their own world views and political agendas.

When we accept that not all cycling traces are officially classified as cycling routes, we no longer have objective criteria to determine which routes should be mapped as cycling routes. This might not be such a major problem in countries with major cycling infrastructures because official paths are much more pervasive and properly represented [M1]. In other countries, particularly in the Portuguese cities where we made this study, the official network can be very small, and a significant part of cycling trips occur outside the official cycling network. This will always become a profound source of ambiguity with important implications for route representation.

7.3.2 Reduced Awareness of Cycling Traces

A specific challenge for mappers is awareness about the cycling traces themselves. Mappers may use satellite images for clearly marked paths, but for cycling traces, they often need to go directly to the field to observe and map existing paths [M2]. However, even on the field, many paths followed by cyclists may not be visible from observation, which means mappers would not normally represent them, even if they become aware that cyclists use them.

The principle is that if the path is not perceivable to cyclists or pedestrians on the ground, it should not be represented [M1]. If, on the contrary, there is a visible trail on the ground, even if only informally created by the previous passages by cyclists or pedestrians, then it should be mapped [M1]. However, even if the trail is visible, its usage by cyclists may not be perceptible during the field mapping process. Therefore, mappers will not become aware of those cycling traces' existence and will not represent them. In large, open urban squares with endless possible paths, mappers are also unaware of which specific path might be more common. However, they may create virtual lines representing logical connectivity more than concrete paths followed by cyclists [M1].

7.3.3 Representation Granularity

From a more practical perspective, there is also an issue with the granularity of representations. City roads are expected to be represented, and they already provide the essential knowledge about the connectivity offered by those roads. In these cases, adding a parallel path to represent the path commonly followed by cyclists, possibly a sidewalk or a safer road margin, may be seen as a refinement of information that, in its essence, is already there. From a practical point of view, many mappers would simply add cycling-related attributes to the road itself, rather than create a whole new representation for a side path that is not officially a cycling path [M1].

Participant M1 reported: *"The reason there are both approaches is due to scalability. Ideally, every sidewalk should be represented independently, but adding an attribute to a road is easier than creating a new feature"* [M1], and added: *"There is a lot that needs to be improved; however, there are not enough hands"* [M1].

7.3.4 Tagging Consensus

The tags used in OSM to classify OSM roads have been repeatedly mentioned as another major source of ambiguity. OSM includes tagging mechanisms that can be combined to create rich representations of

road characteristics and usage. While very flexible, this flexibility comes at a price: there is no consensus or widely accepted practice regarding their concrete meaning or applicability [M3]. Different mappers will have different strategies, meaning that the tags that could help to differentiate the subtle differences between paths will not offer a clear interpretation that can be used to provide cyclists and routing systems with a solid description of their properties. These ambiguities and the misuse of tags are recognised as problems within the community of mappers; however, no easy solution seems to exist, as it would involve having many people agree on a single way to do things [M3]. These ambiguities can impact other systems that could expect to use OSM as their source data for presenting or suggesting cycling routes [M1].

7.3.5 Route Selection is Circumstantial

Participants highlighted very consistently how very diverse types of vehicles and very diverse cycling profiles may strongly affect the perception of the most suitable route. Scooters, electric bikes or cargo bikes were referred to as types of vehicles that could lead to different choices. For example, cargo bikes were described as less suitable for less conventional paths [C1]. Likewise, participants also mentioned how different cycling profiles could significantly impact route choices. For example, travelling with children would be a key driver in choosing sidewalks over roads [C1]. In contrast, sports cyclists using mountain bikes were described as more adventurous and creative in selecting their urban routes [M2].

The key implication of these observations is the need to selectively address the various profiles. This poses a significant challenge for mappers, as it adds significant complexity to the mapping process and requires very fine-grained data about the cycling activity of those various profiles. These data are not easily available and are not obtainable from on-site observations.

7.3.6 Mapping vs Routing

Another source of ambiguity arises from conflicting views associated with mapping and routing purposes. Mapping is concerned with the objective representation of the territory, in this case, the representation of the various road types that people may follow to move from one point to another. A map provides static geographical information that helps to categorise roads (e.g., highways, main streets, bike lanes, pedestrian paths) but does not offer personalised recommendations. In contrast, routing is an interactive and dynamic decision support system aiming to help users find routes tailored to their needs and profiles. A routing decision may consider multiple parameters (e.g., fastest route, safest route, least traffic, most scenic) to find the best match, and many of them will involve assessing diverse properties of the roads that compose the road network.

The interviews revealed that the assumptions about how routing engines should interpret map representations are also a source of ambiguity. By mapping more roads, even roads with obstacles where a cyclist may have to dismount, e.g., urban stairs, pedestrian passage or an elevator, mappers are increasing the range of possibilities and, hopefully, helping routing engines to offer better route suggestions. However, routing applications may use complex algorithms and diverse parameters to suggest a cycling route. This sophistication can be very powerful, but can only be effective with sound information about route possibilities and a strong understanding of cyclist profiles. The ambiguities in road representation can quickly become a problem. For example, routing engines may only consider roads marked as "bicycle"="yes", or they may also consider roads that are not explicitly marked as such [M1]. Mappers may represent more roads and signal their limitations, but it is up to routing engines to decide whether to include them in their route suggestions [M1]. This division of responsibility can be problematic because mappers may not fully understand how their information will be used, and routers may not fully understand all the assumptions and interpretations made by mappers. Cyclist could help with their situated interpretation, but they will also have the same problem interpreting the information represented in OSM.

7.4 The Politics of Route Representation

The second dimension of analysis emerging from the interviews concerns the politics of route representation. Unlike the previous section, this is not about technical limitations. This is essentially about ethical, legal, safety, accountability, and freedom interpretations of cycling activity. While in the previous section, the issues could be mainly addressed within the OSM community, regarding these issues, there are very different perspectives, all of which may lead to very different conceptions about possible solutions for these gaps.

7.4.1 Public Authorities and OSM Awareness

The level of familiarity and adoption of OSM by local authorities was a topic of discussion in the respective interviews. All local authorities had had previous contact with OSM, but none were actively involved in managing or using OSM data. They also had no interaction with the mapper community, which creates OSM representations within their territory [A1]. Mappers mentioned this as a problem. While not necessarily a condition, a stronger familiarity with OSM was described as important to create awareness about its data and its relevance for the ecosystems of cycling services and applications.

Two main arguments justify this distance between local authorities and OSM. First, as part of their

many municipal activities, these local authorities must use other Geographic Information Systems (**GIS**). These are often high-end GIS systems, making OSM seem more limited [A2]. Their familiarity with these systems and the activities they already have to use them do not leave much space for other tools, including OSM. The second reason is the lack of trust in the data. Participant A1 stated: *"This is open data, but it is not our information. It is not produced by us. We are not the ones who process and work on it to provide it to the target audience"* [A1]. Similarly, participant A3 emphasised the concerns about data reliability, stating: *"You often hear about making decisions based on data. The problem is that most of the data used for decision-making is sometimes incorrect or, at the very least, does not accurately represent reality as it should"* [A3].

Despite their reduced awareness about OSM, local authorities were not surprised by these representation discrepancies. They were even aware of some of the specific cases used in the interviews, especially those more directly linked to infrastructure problems. They also recognised that the cyclist practices represented by the GPS traces are common and occur frequently across different locations. Their previous knowledge was mostly based on their empirical knowledge as attentive citizens or occasional cyclists [A2, A1]. However, there were also reports of field studies conducted to produce detailed insights about cyclist behaviour in certain areas, especially as part of urban redevelopment initiatives [A2].

7.4.2 Circumstantial Interpretations of Routes

A recurring theme among participants was the circumstantial interpretation of the reasons that led cyclists to paths that were not supposed to be cycling paths. For example, regarding the option to ride on the sidewalk instead of the road, several participants highlighted circumstantial elements and how they could affect that assessment. Participant A2 noted: *"On a road with heavy traffic, I understand that there could be some hesitation about using the road and that people might prefer to ride on the sidewalk"* [A2]. Participant A1 emphasised the role of sidewalk characteristics, stating: *"It always depends on the characteristics of a sidewalk (...) if it is 4 metres wide and it does not have much pedestrian traffic, it will be perfectly suitable for a cyclist to occasionally coexist with pedestrians."* [A1], and added: *"When it comes to cyclists coexisting on the sidewalk, we're never talking about a high density of cyclists"* [A1].

This led participant A2 to mention cities that have implemented schedules as a solution: *"very early in the morning and at the end of the day, between the evening and the following morning, cyclists can ride freely, but outside of these hours, they must walk with their bicycles"* [A2]. Such an approach, especially if based on dynamic schedules, would represent another challenge for mappers and routing systems.

These frequent references to the situated nature of these decisions have highlighted the ambiguity of

the whole process and the challenges of trying to embed that situated knowledge into the formal representation of the cycling road network.

7.4.3 Private vs Public

An important boundary for route mapping is whether the property is private or public. Many routes will cross private lands simply because there has never been any problem, but that does not mean they can be considered official. These cases might differ substantially, from remote forest trails to a University Campus.

This is a major issue for local authorities. As expressed by participant A4: *"When a local authority marks a route or trail, they have to ensure that there is a formal agreement with the landowner allowing access or else that it is, in fact, a public path... even if it's a forest track, it still has to be public. That's why, normally, when we identify cycle paths, we take that care"* [A4]. A4 added that, in cycling applications, many paths trespass private land, but they result from private initiatives, associations, or similar; they don't bear the same responsibility for land management that a local authority has [A4].

Mappers also recognise these issues with private property as a relevant problem. Even if they do not have the same responsibility as local authorities, they would not want to map routes that trespass on private property and may become a source of conflicts between cyclists and owners. However, the maps they rely on for field mapping do not always include ownership information, making it difficult for them to properly assess whether a path is private or public, or if there is some type of permission to use it, even if it is private. Even a path marked as unofficial could be interpreted as suggesting trespassing. As noted by participant A3: *"advising people to trespass on private land would be illegal, even if it's based on the behaviour of previous cyclists"* [A3]. This places an additional burden on mappers who may be defensive and restrained from representing paths for which they cannot clarify ownership.

7.4.4 Anomalies as Infractions

Local authorities have expressed diverse perspectives on the meaning and the potential value of the information about the anomalies. A common perspective is to frame those anomalies as traffic infractions.

In the specific case of the cities in our study, with only incipient cycling infrastructure, most of these anomalies are essentially a consequence of cyclists' quest for safer and more efficient routes. To a certain extent, local authorities are aware of this reality, but avoid dealing with it through what may be described as wilful ignorance. In these cases, authorities pretend not to know to avoid responsibility or the political consequences of stronger interventions. From this perspective, increased awareness about these anoma-

lies becomes evidence of a reality being avoided, which will only exacerbate a feeling of frustration with the generalised nature of these unresolved situations.

Awareness about these anomalies will also inevitably call attention to the root problems behind the behaviours that lead to those anomalies. It brings a quantitative perspective often not perceptible from on-site observations and may contradict narratives where only a few rule-breaking cyclists do not respect traffic rules. Furthermore, it may emphasise the ambiguity of cycling policies, where local authorities have the legal obligation to enforce rules yet acknowledge a lack of proper alternatives. This exposes the contradictions of those policies and may lead to a surge for solutions that may reduce those infractions, such as additional signage, suggestions for cyclists, stronger enforcement or fines for any violations.

Another element emerging from the interviews was the perception of the impact of the detours that cyclists may need to make to comply with regulations by restricting their paths to official routes. Participant A1 observed: *"We're talking about a very short route. In fact, it was a shortcut that isn't really that necessary, unlike others that might be more useful..."* [A1]. Meanwhile, participant A2 framed the issue as a balance between convenience and civic responsibility, stating: *"The cyclist has to go around. It's a matter of convenience and civic responsibility. I don't think adding a few extra metres to a bike journey is really a problem. But we know people will always take the most convenient route if allowed."* [A2]. This seems to reinforce the idea of a conflicting view about the impact of these detours, as cyclists often describe them as a major hassle.

7.4.5 Anomalies as a Tool for Continuous Improvement of the Cycling Network

An opposite perspective was to embrace anomalies as hints for the continuous improvement of the cycling network. Participants suggested that some of these situations could be seen as desire paths that local authorities should consider as the most natural paths for cyclists [A3]. They recognised that this information could help local authorities make more informed decisions by balancing their understanding of public space with the cyclists' perspectives, potentially leading to changes in the network infrastructure [A1]. It should inform multiple decision levels and decision-making processes, whether in planning changes in public space or managing them effectively [A2].

This perspective, in which information is not made public but is made available to local authorities, was seen as potentially very positive. Local authorities would get all the data and references to anomalies, representing highly valuable insights about cycling behaviour in their city. They could also filter paths involving serious safety concerns or the crossing of private property and return more reliable data for

cycling services.

Regarding interventions, participant A1 suggested that this information could provide the context for small interventions aiming to reroute cyclists to safer routes. The rationale is that sometimes local authorities can analyse on a larger scale and possibly identify equivalent routes that avoid problematic points. Information about gaps could drive their efforts, helping them identify specific locations where alternative routes could be signalled.

Information about these anomalies is also recognised as a form of pressure for addressing well-known problems. These are situations where the need for physical intervention is widely recognised, and potential solutions may already have been discussed, but a concrete solution remains out of sight. In these cases, the anomalies are not in the mapping process. They are anomalies in the city design, and evidence about their existence should provide local authorities with a clear sense of priority [A3].

7.4.6 Accountability vs Informed Decision

The distribution of responsibility and, particularly, the legal implications of representing unofficial paths were repeatedly mentioned by all stakeholders in diverse ways. The discussions highlighted a delicate balance between two opposite perspectives: Accountability and informed decision-making.

The accountability perspective emphasises that publishing a route suggests that route as suitable for cyclists and therefore conveys specific safety, convenience and legal expectations. As expressed by participant A4: *"There has to be some level of Accountability. You can't map every path that cyclists take, because if they're paths that are not suitable from a safety point of view, and I'm not even talking about legality here, I don't think they should be mapped as a cycle route"*.

On the contrary, an informed decision perspective emphasises that those paths could be represented as long as they were properly signalled as unofficial, community, or some other designation that made it clear they were not official cycling routes. Cyclists should then be informed about the properties and potential risks of all possible routes and decide the most suitable route for their circumstances. As participant A3 noted: *"as long as the more serious situations [...] are clearly and previously identified as undesirable, [...] I do believe that yes, this information should be made publicly available so that informed choices can be made"* [A3]. This seems to embrace a vision of society where *"with proper context for the information and data we provide, [...] the decision-making citizen, the person who makes choices about their life every day, will always make the best choices"* [A3]. The assumption would be that, in this way, cyclists would be able to make better decisions and assume the risks associated with their choices.

The results have revealed a significant divide between these two perspectives, sometimes even within

the opinions of the same participant. Participant A2 was very critical of the idea of informed decision, describing it as inappropriate because it could induce people to commit an offence: *"Even if it is a commonly tolerated offence, it would still be wrong to suggest to people that they could do it"* [A2]. Participant A4 suggested that a mapper could be legally accountable for mapping unsuitable routes: *"Suppose there's an accident, and the insurance company realises the cyclist took that route because it appeared on an app. Could they sue the app for giving the cyclist misleading directions?"* [A4]. Experienced cyclists have felt that some of their routes should not be shared. They are willing to take the risk of using specific paths, but do not wish to take the burden of suggesting them to others, who are potentially less experienced: *"There are unofficial paths I take that are quite adventurous, so the idea of sharing them ..."* [C1].

The informed decision perspective was not defended so vigorously. However, many participants acknowledged that most cycling traces are very likely to include roads with some type of limitation, and therefore, excluding them all would be too restrictive. Participant A3 argued that representation was essential and suggested an informal perspective that acknowledges the reality of the situation while highlighting all possible risks. A3 explained: *"We make these decisions every day, all the time, and I lean towards my belief that people are intelligent. That is, we make decisions based on information, and the more information we have, the better decisions we are likely to make. Perhaps representation should include a few warnings, indicating that something is not the most natural, formal, or advisable in various ways..."*. Participant M1 suggested that specific markings could be used to signal the informal nature of the path and the specific limitations that cyclists should be aware of if they decide to take those paths. Participant A4 suggested a hybrid approach, involving *"mapping those paths as a pedestrian route or as a car route, and then if a cyclist chooses to take a pedestrian or car route, that's their responsibility"* [A4].

The essence of these discussions ends up in the many interpretations of what it means to include a road in the OSM model. Participant A3 was receptive to representing those roads, but ended up in a challenging discussion about the semantics of such indications. The key idea is that representing those roads, with a proper description, should not express that cyclists "could" take those alternative roads, as that would imply official permission. *"I would view it negatively if the municipality had official plans and then suddenly, the very entity responsible for managing the roads and enforcing the rules were to say, 'It's better to take the sidewalk' ... I would find that message incoherent and inconsistent"* [A3].

If a route is interpreted only as a sign of the frequent passage of previous riders, then the mapper's accountability would be minimal because he would only be providing information that was accurate and had to be interpreted only for what it was. However, that is not how most people seem to interpret represented routes, and it is not how mappers see their work. For mappers, representing dangerous or private paths

in OSM is controversial. From their perspective, people will interpret such paths as usable based on their expectations about the meaning of "it can be used". This is a responsibility that mappers will not assume. Even if they know that many people take that path, no mapper is willing to take on the responsibility of suggesting to people to cross such an area [M3].

This divide between accountability and informed decision-making represents a significant challenge for OSM. It highlights a fundamental clash between a traffic management culture rooted in formal regulations and a culture of active mobility and community-based knowledge, which often relies on common sense to define acceptable behaviours.

7.5 Implications in the OSM's Role as an Enabling Technology

The results obtained from the focus groups and interviews have clearly shown the diversity of perspectives various stakeholder groups can have about the same set of anomalies in the OSM representation of common cycling routes. In this section, we discuss the implications of these results for the role of OSM as an enabling technology for cycling information systems.

7.5.1 Implications of Technical Uncertainty

Regarding the more technical challenges, many of them could be addressed through improvements in the OSM technology and practices. Still, the core issue is the disconnection between the various stakeholder groups, which does not create the right incentives for convergence within and between communities. For example, tagging ambiguity can be complex to address within the OSM community alone. The community favours flexibility and openness, instead of more constrained representations. Without the external pressure of concrete and relevant usage practices, convergence will be very hard to emerge.

A stronger connection between local authorities and mappers, through municipal information systems, could provide an important external push for convergence. However, local authorities, at least those considered in this study, do not take OSM very seriously. Thus, they will not complain about misrepresentations or tag ambiguity, they will not feed OSM with their own data, or they will even be aware of its potential role for local cycling services and applications.

Likewise, a stronger connection between cyclists and OSM, through OSM-based applications, could also play a key role in promoting convergence by helping to close the information cycle. Applications could value and benefit from the OSM road network by suggesting more suitable routes, while also generating additional information to feed the evolution of the models. However, the ambiguity of road network models

is also a problem for applications, as they cannot find the consistency they would need to offer cyclists the same consistency, particularly across many different locations. Consequently, cyclists' expectations about the representativeness of data provided by those applications and the perceived value of those applications also seem very low. Cyclists report that they only use route recommendation systems when travelling to new destinations. Once they develop situated knowledge about local routes, they gradually adjust their journey based on their experience [C4]. Finally, a more thorough representation of the road network data model is also challenging when considering the voluntary nature of the contributions and the lack of data most contributors face.

Therefore, the key conclusion is that without more intensive adoption, OSM may not be able to reach the level of demand that could push for more solid solutions and overcome the issues identified by the various stakeholders.

7.5.2 Implications of Conflicting Perspectives Between Stakeholders

Regarding the issues associated with the politics of mapping representations, we have seen how even very concrete situations can elicit very different interpretations of the respective causes and possible solutions. A key point is therefore to discuss to what extent these different views might be reconciled.

We should acknowledge that no single stakeholder will have a fully comprehensive view of reality. There might be some cases where the problem is just about cyclists misbehaving, and further road signage or vigilance may be needed. There might be cases where the problems originate from poorly designed infrastructures that need to be addressed by the local authorities, and, finally, there are cases where there is just a problem of representation, and OSM mappers can quickly make the necessary corrections. For these cases, knowledge about the gaps is valuable mainly in how it can create awareness about the issues and prompt quicker action.

However, we have seen from the analysis of the 6 cases by the three stakeholder groups that, for most cases, any realistic solution will have to consider elements from the perspectives of each of these stakeholder groups. An appropriate assessment will always involve experience with mapping procedures, a deep understanding of cycling practices, and analysis of many technical variables, such as demand volume, road characteristics, and available alternative routes.

A well-balanced solution should thus consider the safety and accessibility needs of all citizens, ensuring that the infrastructure serves cyclists and promotes their safety while also protecting pedestrians, children, and individuals with reduced mobility. As participant A2 emphasised, local authorities need to work with the cyclists to *"build a solution that is useful for everyone... It may not be the best for cyclists, but it has*

to be a middle ground that is good for them, but does not endanger pedestrians and does not endanger cyclists.... We have to work on a solution together, even if it is not the best for either party" [A2]. Supporting this need for collaboration, participant A1 noted: *"balance is a good tool... the balance between what is ideal and what is possible... Knowing that, from the point of view of a municipality, if we try to impose the ideal, it will not be used, and what will happen is that cyclists will end up using routes that pose risks to them"* [A1].

Despite widespread recognition of the importance of coordination between technical knowledge, users, and managing entities, this collaboration is not happening in practice. Mappers highlighted the lack of institutional cooperation: *"I wish there was more cooperation with the authorities. It is a bit disheartening to map cycling conditions when they don't actually exist"* [M1]. The simple "No" response to the question about interaction with municipalities reveals a gap in coordination between mappers and authorities, something that technical staff like participant A1 considered essential for safer decisions. Without stronger coordination between the parties involved, no relevant solution is likely to emerge.

7.6 Summary

This chapter addresses the often neglected existence of relevant and frequent gaps between common cycling paths and the cycling road network representation in OSM. The relevance of this study concerns the potential role of OSM as an enabling technology for cycling information systems. To support this study, we conducted two focus groups and four in-person interviews with three key stakeholder groups: cyclists, OSM mappers and local authorities. These sessions were centred on the analysis of six concrete cases of representation gaps by these different stakeholders.

The results have provided relevant insights to answer the research objective defined for this study. The research objective aimed to provide a deeper understanding of these representation gaps and how they are perceived by the various stakeholder groups. We identified two major categories of issues relevant to this question. The first concerns the technical challenges of representing the inherent informality of many cycling routes. This is challenging because cyclists' paths are very fuzzy. Not only because interpretations differ across stakeholder groups, but also across elements within a stakeholder group, and even the same individual may have different interpretations in different circumstances. We have also shown how this inherent ambiguity is exacerbated by additional ambiguity in OSM practices, particularly regarding the use of tags to classify roads and how they should be used for selecting cycling routes.

The second category concerns the politics of route representation and the different interpretations that

the various stakeholder groups can have about the same concrete cases. The results have confirmed the existence of significantly different perspectives and provide valuable insights into the concrete elements that make those differences. While we were able to identify significant divides, for example, between anomalies as infractions versus anomalies as a tool for improvement, and also between accountability versus informed decision, we also identified considerable recognition that the best answer is always strongly dependent on specific circumstances, which is yet another source of ambiguity, but an important sign for the need of congregation of efforts between stakeholders.

Our research objective also aimed to explore how knowledge about these anomalies could help to congregate the diverse and often opposite perspectives that these stakeholders can have. In section 7.5, we provide our answer to this question by revisiting the results described in sections 7.3 and 7.4 and discussing what they mean to the role of OSM as an enabling technology for cycling information systems. The first conclusion is that many of the problems identified in this study are associated with the disconnection between the various stakeholder groups. This disconnection reinforces ambiguity and does not produce the necessary incentives to promote convergence within and between communities. This may severely undermine the ability of OSM to reach the level of demand that could push for more solid solutions and overcome the issues identified by the various stakeholders.

Despite widespread recognition of the importance of coordination between these stakeholders, this collaboration is not happening in practice. However, a positive outcome of this work is the confirmation that shared knowledge about concrete cases of representation anomalies can help to bridge across these stakeholder groups. By framing the dialogue between those stakeholders, anomalies help focus on concrete situations and promote a problem-solving mindset where it may become easier for participants to exchange ideas, perceptions, and practices to uncover balanced solutions that meet the sensitivities of all parties.

Chapter 8

Conclusions

8.1 Contributions

In this section, we revisit the research question that frames this thesis's work, introduced in Chapter 1:

"Can OpenStreetMap serve as a reliable reference system for cycling analysis and an enabling technology for cycling information systems?"

This broad research question was addressed through five specific research objectives. Here, we summarise the key contributions and findings associated with each objective. Following this discussion, we present our formulation of an answer to the research question based on the findings of our work.

Objective 01. Assess the existence of generic functionality across cycling models, characterise the opportunity for horizontal services, and analyse how a digital platform strategy could bring many of those models into a common system with added value for different cycling.

We started this thesis by identifying a critical limitation of analytics models for cycling in urban environments: existing models have not yet fulfilled their potential as drivers for positive change. In chapter 1, we posited that the underlying root cause relies on their development paradigm, described as monolithic, and then we proposed that a platform strategy should be able to overcome these limitations and allow models to deliver their full potential.

Through a structured analysis of three mobility models, as described in Chapter 3, we demonstrated that these models, yet simple, can be decomposed into several atomic code blocks, with numerous possibilities for creating horizontal features. Crucially, these models have strong domain knowledge embedded during their development, which remains siloed and is not shared upon publication. These findings supported our initial hypothesis regarding the incompatibility between the prevalent development paradigm and the requirements of digital innovation, such as distributed and combinatorial innovation.

Also in Chapter 3, building on general principles of platform innovation, we articulated several essential properties for a platform dedicated to analytics in urban cycling. Such a platform should be specific to cycling, multi-sided, and global, should support collective intelligence, horizontal innovation, and the facilitation of convergence for models and data formats. These properties are expected to foster significant progress in cycling mobility model development by optimising the development effort and improving their overall quality.

Objective O2. Characterise the concrete causes and the main properties associated with the discrepancies between real cycling routes and how they are mapped into the OSM road network model.

In chapter 1 we posited that, despite OSM's recognised good coverage and data quality, the OSM road network may not represent some of the actual cycling paths because cyclist routes are fuzzier and include many unconventional and unofficial paths. To characterise the concrete causes and the main properties associated with such discrepancies, we conducted the study described in the chapter 4. We designed and recorded a cycling route in Braga, riding through paths commonly used by cyclists. In that route, we intentionally included segments likely to challenge map-matching algorithms due to their unofficial nature. The resulting GPS trace was then map-matched into OSM.

The analysis of the map-matching results revealed that, while the map-matching service succeeded for the majority of GPS segments, it often fails in situations where the path of the cyclist diverges from the official road network, resulting in map-matching gaps. Based on these results, we were able to characterise the concrete causes of map-matching gaps. Many categories of these gaps stemmed not just from GPS or algorithmic errors, but from genuine misalignments between cyclists' traces and road network representation in OSM. This includes map-matching gaps caused by missing OSM road segments, incorrect tagging, missing connections between two or more OSM roads, riding against traffic on a one-way road, situations where the cyclist changes its direction to avoid complex crossings, or situations where two parallel roads exist near each other.

Objective O3. Develop a supervised methodology to detect relevant discrepancies between cyclists' GPS traces and the routes produced by a map-matching algorithm and classify them according to their root cause.

To address the third objective, in chapter 5, we studied the feasibility of using supervised machine learning classification models to automatically detect and classify map-matching gaps, allowing us to identify the

GPS traces that cannot be map-matched into OSM. More specifically, we compared the performance of several binary and multi-class classification models using data for four European cities. For each city, we mapped into OSM a set of cyclists' GPS traces obtained from Wikiloc, manually inspected each map-matching result to establish ground-truth data, and computed several metrics to assess the similarity between each GPS trace and its corresponding map-matched route.

Regarding the binary classifiers, XGBoost with ANOVA feature selection was most effective in identifying the presence of such gaps. This model achieved a 0.906 F1-score, with accuracy close to 0.89, showing a reliable ability to identify the presence of these gaps in a GPS trace. This method also outperformed other methods in the literature, namely the Error Indicator method. The EI method was found to be very vulnerable to outliers, reducing its sensitivity to gaps where the differences between the GPS trace and the map-matching are very subtle, such as the anomalies classified as "Competing Road".

However, classifying the gaps' root cause remains a challenge. The performance of multi-class classification models was unsatisfactory: the random forest classifier with no feature selection achieved the best performance with an F1 score of 0.59 and an accuracy of 0.71. Despite being able to distinguish between cases with and without gaps, in most cases, it failed to classify those gaps according to their root causes.

Overall, results have shown that it is viable to identify the presence of gaps. The categorisation may help people identify the gaps' root cause, thus potentially helping OSM collaborators maintain the OSM database and helping municipalities plan a better road infrastructure.

Objective 04. Develop an unsupervised computational method to process sets of cycling traces and identify city areas prone to map-matching anomalies, and analyse how the knowledge about the map-matching anomalies can help cities uncover novel and actionable insights about real cycling behaviour.

Map-matching gaps correspond to cases where the map-matching route is not representative of the GPS trace. However, this concept is associated with a single GPS trace. To move beyond the scope of a single trace and look at the global perspective of urban mobility, in Chapter 6, we introduced the concept of map-matching anomalies. An anomaly is a severe mismatch between the GPS traces and road network models to which they are expected to be matched. As a result, map-matching algorithms will systematically produce map-matching gaps near that location. Anomalies are not a consequence of GPS or other algorithmic errors, but stem from the limitation of the road network model in that area. By considering the global view provided by many GPS traces, gaps produced by GPS interferences or other occasional events become

distributed and diluted.

We also proposed an unsupervised methodology to identify these anomalies in a city. By employing Dynamic Time Warping (DTW) to compare GPS traces against map-matching results and applying DB-SCAN clustering over 16,292 GPS traces, we successfully identified urban areas susceptible to recurrent anomalies. This method achieved 84% precision for anomaly identification.

Based on our experience of analysing the results, we proposed several metrics that may be useful to automatically benchmark cities' propensity to anomalies and a visual approach with four levels of analysis, where urban planners and OSM mappers may identify anomalies' root causes. The broader level includes an overview of the anomalies' spatial distribution over a city, along with their frequency or impact on the cycling experience. Then, the individual can focus on a specific anomaly, inspect the different cyclists' movement patterns, and inspect individual GPS trajectories. This makes it easier to understand potential conflicts between cyclists, the existing road network and its representation on OSM.

Objective 05. Characterise the perspectives of key stakeholder groups about the discrepancies between common cycling paths and the cycling road network model in OSM, and explore how knowledge about these discrepancies can become a tool for congregating the perspectives of these stakeholders and reducing the representation gaps.

Chapter 7 addressed the fifth research objective by characterising the perspectives of three key stakeholder groups about the gaps between OSM and cycling practices, and exploring how the knowledge about them can become a tool for congregating the perspectives of these stakeholders. For this purpose, we conducted two focus groups and four interviews with key stakeholder groups to foster discussion and collect their perspectives.

Two main topics emerged from the interviews: how OSM technology and practices contribute to the discrepancies between real routes and their representation, and how the multiple political perspectives and interpretations about the anomalies contribute to their existence. In the first dimension, we found that anomalies emerge because the OSM model is only a simplification of the world. Moreover, they also emerge because mappers have reduced awareness about many cycling routes, employ different mapping techniques, and have an ambiguous interpretation of the roles of mapping and routing. In the second dimension, we found that while anomalies can be seen as traffic infractions and as an excuse to promote law enforcement, they can also be seen as a tool for continuous improvement of the cycling network by revealing cyclists' desired paths. However, the circumstantial interpretations of routes were a recurring theme. The best solution to resolve an anomaly will always depend on several factors, including the

volume of cyclists and pedestrians, or the characteristics of the infrastructure.

Another recurring theme was the dichotomy between the accountability of mapping a route versus allowing cyclists to make informed decisions. From one perspective, publishing a route means that it is suitable for cyclists and therefore conveys specific safety, convenience and legal expectations. From the other, informal paths could be represented as long as they were properly signalled as unofficial cycling paths. Cyclists should then be informed about all the potential risks and decide the most suitable route for their circumstances.

The sessions with the stakeholders also revealed that the perspective of a single stakeholder group is not enough to create a suitable solution that benefits all citizens, as each group does not have a comprehensive view of reality. Moreover, there is a lack of co-creation processes for stakeholders to share their knowledge, perceptions, and ideas about concrete problems. Anomalies can be the bridge between stakeholders. By framing the dialogue between those stakeholders, it focuses on concrete situations and promotes a problem-solving mindset where it may become easier for participants to exchange ideas, perceptions, and practices to uncover balanced solutions that meet the sensitivities of all parties.

Research Question: To what extent can OpenStreetMap serve as a reliable reference system for cycling analysis and an enabling technology for cycling information systems?

This thesis confirmed that OSM possesses several limitations in representing de facto urban cycling routes, often resulting in discrepancies between cycling traces and the OSM road network. These discrepancies between actual cycling trajectories and the OSM road network directly impact the quality of cycling services, applications, and predictive models. Based on the findings of this research, we can conclude that these limitations undermine the OSM's capacity to serve as the definitive convergence mechanism for large-scale cycling initiatives.

We can also discuss the limitations arising from limitations in the OSM model itself or from common practices around OSM mapping. In general, it can be argued that the OSM model is so generic that it should not be a limitation in itself. However, abstraction also has its price, which, in this case, is the added challenge of promoting convergence on specific practices. This is not a major problem for car road networks, as they have always been the major use case for OSM, and their mapping practices are well established. However, when considering other types of road networks, there are no special constraints on how to represent them, but there are also no well-defined practices on how to do it. This is a problem for any road network, which is expected to be co-created by multiple entities and serve multiple independent services. Without some level of convergence on basic principles, such as which tags to use or which

roads to represent, a road network model will not be able to serve this core role in mapping services. The challenge of achieving consistent tagging across locations with heterogeneous cycling infrastructures is likely less pronounced in commercial road network datasets. Since these are typically produced and maintained by centralised entities with full control over the mapping process, they may be less prone to tagging ambiguities and inconsistencies in how information is interpreted by routing engines or other applications.

The perception of this problem and the need to confront this mismatch between road network representations and cycling traces raise the obvious question of what can be done about it. The answer may not be so obvious because the root of the problem lies in a fundamental flaw of the current model. The results show a clear conflict between cyclists and mobility authorities about what constitutes an official or a de facto cycling route. For cyclists, many of these unofficial routes may simply be obvious choices based on their field knowledge and their situated assessment of current circumstances. Mobility authorities might tolerate some of these de facto routes because they understand their value for cyclists, but they are bound by regulatory and legal responsibilities that may stop them from acknowledging those routes as suitable cycling routes.

In the end, this dilemma ends up as a content management challenge for OSM. The idea that everyone can create and enrich OSM so that OSM data may become a faithful representation of the physical world as seen by the collective view of multiple contributors is at the core of the OSM model. It also matches common cycling practices, where route sharing is very popular. The insider knowledge that cyclists have about the best ways to cycle in their city could allow the OSM road network to evolve continuously, as any viable routes would eventually become part of the model. However, this open paradigm relies on the implicit assumption that, if people go to the field, they can assert the correction of information, and it should thus be simple to agree about the existence of a road and its basic properties. In practice, this rarely occurs because many cyclists' paths are not physically visible on the ground and cannot be discovered by surveying the area. Moreover, even when the markings are visible on the ground, mappers often lack data indicating the level of demand by cyclists. The idea that there are de facto cycling routes that cannot be officially classified as cycling routes creates an ambiguity that challenges those basic assumptions and the simplicity of OSM abstractions.

Additionally, this thesis identifies a significant lack of adoption of OSM across different stakeholders. Local authorities do not take OSM very seriously and prefer to use proprietary GIS tools, which they are often already used to, potentially underestimating the utility of crowdsourced platforms. Simultaneously, cyclists exhibit low expectations regarding the accuracy and representativeness of data in existing applications.

Without broader institutional and user adoption, OSM may struggle to reach the level of maturity required to drive more robust solutions and address the issues identified by the various stakeholders.

Nevertheless, despite the current limitations, OSM remains an immensely valuable resource for mobility research and practical applications, and its data can still be leveraged to enhance cycling mobility. Maintained by a vast and active community of volunteers who strive to keep it up to date, OSM offers global coverage of the road network and open access to its data. This accessibility enables the replication of services and models almost anywhere in the world without substantial financial costs. The primary recommendation, particularly for cycling-related studies, is therefore to use OSM data with caution, acknowledging that it does not capture every cycling route, especially those not officially designated for cyclists but which in practice are used by them. Regarding these unofficial paths, the OSM community has not reached consensus on whether they should be mapped based on usage demand. In line with the "on the ground" mapping principle, even when physical traces or markings indicate real desire lines, such routes should, at most, be classified for pedestrians.

In this context, map-matching anomalies derived from cycling activity can serve as a point of convergence for cycling stakeholders, helping to frame discussions around concrete cases and optimise collective efforts. By focusing on specific instances, anomalies can become a valuable tool for reducing the gaps between real-world cycling practices, the existing road network, and its representation in OSM. More specifically, for OSM contributors, anomalies can reveal issues within the OSM road network, such as missing links or incorrect tags, while for municipalities and urban planners, they can identify areas with potential for improvement and guide effective interventions. These infrastructure changes would then be reflected and updated in OSM, reinforcing the alignment between the physical and digital representations of cycling routes.

If all cycling stakeholders could align their efforts and progressively close the gaps between actual cycling routes and their representation, OSM could become a far more reliable reference system for aggregating cycling data and fostering a new generation of mobility models and applications. In turn, the development of these new solutions could raise awareness of OSM and encourage its adoption among authorities and users, creating a virtuous cycle of continuous improvement and broader utilisation.

However, this collaboration is still far from becoming a reality. At present, the different stakeholders remain disconnected and do not cooperate. Without meaningful collaboration, achieving a point where OSM accurately represents real cycling routes, including what are now unconventional paths, remains a distant goal, if not an impossible one.

8.2 Limitations

A limitation of the work described in Chapter 3 is that we only considered three models for the model decomposition exercise. Although our objective was to confirm the existence of generic functionalities across cycling models and the limitations of the current model development paradigm, the review of a wider range of models would increase the robustness of the study and allow us to have a deeper understanding of the frequency of the problem.

A second limitation of this work is that we did not evaluate whether combining the machine learning models, as described in Chapter 5, with the gap extraction procedure described in Chapter 6 would improve the identification of anomalies. The main reason was that, in the work described in Chapter 5, we only considered GPS traces with an average of 1000 meters, and the machine learning models were trained using that dataset. In the work described in Chapter 6, we wanted to extract anomalies independently of the GPS trace length.

A third limitation of this thesis is that we only considered OSM as the source of road network data. It would be interesting to study how other road network models represent urban cyclist routes. More specifically, it would be interesting to adapt the anomaly identification method presented in Chapter 6 to other road network models and compare the anomalies produced.

A fourth limitation is that we didn't assess how different map-matching algorithms and algorithm parameterisation will affect anomaly identification results. Nevertheless, the process is expected to work independently of any map-matching algorithm, parameterisation. It is also expected to work with other road network models.

A final limitation is that, regarding the perspective of stakeholders about anomalies, as described in Chapter 7, all interviews were made in the context of Portuguese cities, where cycling culture is still very limited. It would be interesting to study the perspectives of people from other countries, especially with a different cycling culture.

8.3 Future Work

We can enumerate several research opportunities emerging from this thesis.

First, the core for reducing the gaps between cycling routes and the OSM road network relies on the anomaly detection method. Therefore, improving this method is mandatory to not only reduce the number of false positives and reduce the wasted effort, but also to detect occurrences that currently pass

undetected.

Reducing the ambiguities regarding cycling routes requires collaboration between stakeholders. Therefore, a second research opportunity is the establishment of concrete procedures to foster dialogue and exchange of information between stakeholders. An example could be developing a platform where cyclists collaboratively share their activities to help municipalities and OSM mappers detect anomalies in the road network and OSM.

A third research opportunity is the development of a technical platform to develop and execute new analytics models. This idea, presented in chapter 1 and validated in chapter 3, will make it possible to bring together the efforts of researchers, students and programmers interested in improving urban mobility, as well as the use of the resulting models by municipalities to improve the city and by cyclists to move around cities.

A fourth research opportunity is to repurpose the anomaly identification process with other road networks and develop a standardised approach to systematically benchmark how well each road network is aligned with real cycling practices.

Bibliography

Acs, Z., Song, A., Szerb, L., Audretsch, D., and Komlosi, E. (2021). The evolution of the global digital platform economy: 1971–2021. *Small Business Economics*, 57:1–31.

Alattar, M. A., Cottrill, C., and Beecroft, M. (2021). Modelling cyclists' route choice using strava and osmnx: A case study of the city of glasgow. *Transportation Research Interdisciplinary Perspectives*, 9:100301.

Anderson, J., Sarkar, D., and Palen, L. (2019). Corporate editors in the evolving landscape of openstreetmap. *ISPRS International Journal of Geo-Information*, 8:232.

Arellana, J., Saltarín, M., Larrañaga, A. M., González, V. I., and Henao, C. A. (2020). Developing an urban bikeability index for different types of cyclists as a tool to prioritise bicycle infrastructure investments. *Transportation Research Part A: Policy and Practice*, 139:310–334.

Barrington-Leigh, C. and Millard-Ball, A. (2017). The world's user-generated road map is more than 80 *PLOS ONE*, 12:e0180698.

Behr, T., van Dijk, T. C., Forsch, A., Haunert, J.-H., and Storandt, S. (2021). Map matching for semi-restricted trajectories. In Janowicz, K. and Verstegen, J. A., editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*, volume 208, page 12:1–12:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Berger, M. and Dörrzapf, L. (2018). Sensing comfort in bicycling in addition to travel data. *Transportation Research Procedia*, 32:524–534.

Bergman, C. and Oksanen, J. (2016). Conflation of openstreetmap and mobile sports tracking data for automatic bicycle routing. *Transactions in GIS*, 20:848–868.

Berjisian, E. and Bigazzi, A. (2023). Evaluation of map-matching algorithms for smartphone-based active travel data. *IET Intelligent Transport Systems*, 17:227–242.

Bittner, C. (2017a). Diversity in volunteered geographic information: comparing openstreetmap and wikimapia in jerusalem. *GeoJournal*, 82:887–906.

Bittner, C. (2017b). Openstreetmap in israel and palestine – 'game changer' or reproducer of contested cartographies? *Political Geography*, 57:34–48.

Blan, L. and Butler, R. (1999). Comparing effects of aggregation methods on statistical and spatial properties of simulated spatial data. *Photogrammetric Engineering and Remote Sensing*, 65:73–84.

Bock, J. D. and Verstockt, S. (2023). Road cycling safety scoring based on geospatial analysis, computer vision and machine learning. *Multimedia Tools and Applications*, 82:8359–8380.

BooST (2018). Gross potential for cycling. In BooST `https://boost.up.pt/en/roadmap/gpc/`. [Accessed 12-05-2021].

Boot, M., Ulak, M. B., Geurs, K. T., and Havinga, P. J. M. (2024). Using body sensors for evaluating the impact of smart cycling technologies on cycling experiences: a systematic literature review and conceptual framework. *European Transport Research Review*, 16:13.

Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu Raimond, A.-M., and De Runz, C. (2023). Analysis of cycling network evolution in openstreetmap through a data quality prism. *AGILE: GIScience Series*, 4:1–9.

Broach, J., Dill, J., and Gliebe, J. (2012). Where do cyclists ride? a route choice model developed with revealed preference gps data. *Transportation Research Part A: Policy and Practice*, 46:1730–1740.

Bulc, V. (2016). Cycling: green and efficient transport for the future. In European Commission `https://ec.europa.eu/commission/commissioners/2014-2019/bulc/blog/cycling-green-and-efficient-transport-future_en`. [Accessed 31-10-2019].

Cabral, R., Peixoto, E., Carvalho, C., and José, R. (2021). An ecosystem approach to the design of sensing systems for bicycles. In Paiva, S., Lopes, S. I., Zitouni, R., Gupta, N., Lopes, S. F., and Yonezawa, T., editors, *Science and Technologies for Smart Cities*, pages 580–595, Cham. Springer International Publishing.

Carvalho, C. and Jose, R. (2022). The gap between cycling practices and mapping services for smart cycling. In *2022 IEEE International Smart Cities Conference (ISC2)*, pages 1–7. IEEE.

Carvalho, C., Pessoa, R., and José, R. (2022). Cycling analytics for urban environments: From vertical models to horizontal innovation. In Martins, A. L., Ferreira, J. C., and Kocian, A., editors, *Intelligent Transport Systems*, pages 135–148, Cham. Springer International Publishing.

Carvalho, C., Ramires, M., and José, R. (2024). Automatic identification and classification of map-matching anomalies in cycling routes. In *Proceedings of the 13th International Conference on Smart Cities and Green ICT Systems*, pages 17–28. SCITEPRESS - Science and Technology Publications.

Caulfield, B., Brick, E., and McCarthy, O. T. (2012). Determining bicycle infrastructure preferences – a case study of dublin. *Transportation Research Part D: Transport and Environment*, 17:413–417.

Chauhan, A., Sarkar, D., Agrawaal, T. S., and Soden, R. (2024). Value tensions in openstreetmap: Openness, membership, and policy in online communities. *Proc. ACM Hum.-Comput. Interact.*, 8.

Chen, C.-F. and Chen, P.-C. (2013). Estimating recreational cyclists' preferences for bicycle routes – evidence from taiwan. *Transport Policy*, 26:23–30.

Chen, P., Shen, Q., and Childress, S. (2018). A gps data-based analysis of built environment influences on bicyclist route preferences. *International Journal of Sustainable Transportation*, 12:218–231.

Ciriello, R. F., Richter, A., and Schwabe, G. (2018). Digital innovation. *Business & Information Systems Engineering*, 60:563–569.

Citizens, B. (2017). The gps cycling data analysis tool - for evidence based transport planning. In bca `https://cyclingdata.net/`. [Accessed 18-05-2021].

Contributors, O. (2005a). Josm. In JOSM `https://josm.openstreetmap.de/`. [Accessed 18-04-2025].

Contributors, O. (2005b). Josm. In JOSM - OpenStreetMap Wiki `https://wiki.openstreetmap.org/wiki/JOSM`. [Accessed 18-04-2025].

Contributors, O. (2007). Josm/validator. In JOSM/Validator - OpenStreetMap Wiki `https://wiki.openstreetmap.org/wiki/JOSM/Validator#Source_code`. [Accessed 18-04-2025].

Contributors, O. (2023). In OpenStreetMap Wiki. `https://wiki.openstreetmap.org/wiki/Disputes`. [Accessed 02-07-2025].

Contributors, O. (2025a). Key:cycleway. In Key:cycleway - OpenStreetMap Wiki `https://wiki.openstreetmap.org/wiki/Key:cycleway`. [Accessed 13-10-2025].

Contributors, O. (2025b). Tag:highway=cycleway. In Tag:highway=cycleway - OpenStreetMap Wiki `https://wiki.openstreetmap.org/wiki/Tag:highway%3Dcycleway`. [Accessed 13-10-2025].

Copenhagenize (2019). 2019 copenhagenize index. In Copenhagenize https://copenhagenizeindex.eu/. [Accessed 28-07-2020].

Coya (2019). Global Bicycle Cities Index 2019. In Coya https://www.coya.com/bike/index-2019. [Accessed 18-05-2021].

Cyclers (2019). Cyclers app. In Cyclers App https://cyclers.app/plan. [Accessed 18-05-2021].

de Magalhães, D. J. A. V. and Rigatto, I. B. (2025). Individual perceptions of critical factors on route affecting the willingness of direct commuting trips by bicycle in a hilly city. *Transportation*, 52(1):127 – 153.

de Matos, F. L., Fernandes, J. M., Sampaio, C., Macedo, J., Coelho, M. C., and Bandeira, J. (2021). Development of an information system for cycling navigation. *Transportation Research Procedia*, 52:107–114.

Dey, S., Tomko, M., and Winter, S. (2022). Map-matching error identification in the absence of ground truth. *ISPRS International Journal of Geo-Information*, 11(11).

Duc-Nghiem, N., Viet-Phuong, N., and Kubota, H. (2020). Reasons for cyclists denying new-built bicycle lanes. In Ha-Minh, C., Dao, D. V., Benboudjema, F., Derrible, S., Huynh, D. V. K., and Tang, A. M., editors, *CIGOS 2019, Innovation for Sustainable Infrastructure*, pages 933–939, Singapore. Springer Singapore.

Edelman, B. (2015). In Harvard Business Review https://hbr.org/2015/04/how-to-launch-your-digital-platform. [Accessed 25-06-2025].

Eguiluz, A., Hernandez-Jayo, U., Casado-Mansilla, D., Lopez-de Ipina, D., and Moran, A. E. (2022). Design and implementation of an open-source urban mobility web service based on environmental quality and bicycle mobility data. In *2022 7th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–5.

Eiter, T. and Mannila, H. (1994). Computing discrete fréchet distance. Technical Report CD-TR 94/64, Information Systems Department, Technical University of Vienna.

Elia, G., Margherita, A., and Passiante, G. (2020). Digital entrepreneurship ecosystem: How digital technologies and collective intelligence are reshaping the entrepreneurial process. *Technological Forecasting and Social Change*, 150:119791.

Engler, N. J., Scassa, T., and Taylor, D. R. F. (2014). Chapter 4 - cybercartography and volunteered geographic information. In Taylor, D., editor, *Developments in the Theory and Practice of Cybercartography*, volume 5 of *Modern Cartography Series*, pages 43–57. Academic Press.

Estimate, A. (2022). The most bike-friendly cities in the u.s. (2022 data). In Anytime Estimate `https://anytimeestimate.com/research/most-bike-friendly-cities-us-2022/`. [Accessed 22-01-2024].

Fernandez, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905.

Ferreira, J. P., Isidoro, C., Sá, F. M., and Mota, J. C. B. D. (2020). The economic value for cycling – a methodological assessment for starter cities. *Hábitat y Sociedad*.

Ferster, C., Fischer, J., Manaugh, K., Nelson, T., and Winters, M. (2020). Using openstreetmap to inventory bicycle infrastructure: A comparison with open data from cities. *International Journal of Sustainable Transportation*, 14(1):64–73.

Ferster, C. J., Nelson, T., Robertson, C., and Feick, R. (2018). *Current Themes in Volunteered Geographic Information*, pages 26–41. Elsevier.

Figueiredo, A. P. and Vale, D. S. (2018). Bikefriendlyindex – um indíce para avaliação da amigabilidade de um concelho para a utilização da bicicleta enquanto modo de transporte urbano. In Bike Friendly Index `www.bikefriendlyindex.com`. [Accessed 02-07-2022].

Forum, O. W. (2024). In Oliver Wyman Forum `https://www.oliverwymanforum.com/mobility/urban-mobility-readiness-index.html`. [Accessed 04-04-2025].

Gao, T., Daamen, W., Krishnakumari, P., and Hoogendoorn, S. (2024). Map-matching for cycling travel data in urban area. *IET Intelligent Transport Systems*.

Gholamialam, A. and Matisziw, T. C. (2019). Modeling bikeability of urban systems. *Geographical Analysis*, 51:73–89.

Glasze, G. and Perkins, C. (2015). Social and political dimensions of the openstreetmap project: Towards a critical geographical research agenda. In Jokar Arsanjani, J., Zipf, A., Mooney, P., and Helbich,

M., editors, *OpenStreetMap in GIScience: Experiences, Research, and Applications*, pages 143–166. Springer International Publishing, Cham.

Gomez, J. (2017). Ride Report bike route tracking. In blog.mapbox.com `https://blog.mapbox.com/ride-report-bike-route-tracking-f0ecad0a08db`. [Accessed 04-05-2021].

Graser, G., Straub, M., and Dragaschnig, M. (2015). Is osm good enough for vehicle routing? a study comparing street networks in vienna. In Gartner, G. and Huang, H., editors, *Progress in Location-Based Services 2014*, pages 3–17. Springer International Publishing, Cham.

Grigore, E., Garrick, N., Fuhrer, R., and Axhausen, I. K. W. (2019). Bikeability in basel. *Transportation Research Record: Journal of the Transportation Research Board*, 2673:607–617.

Gu, P., Han, Z., Cao, Z., Chen, Y., and Jiang, Y. (2018). Using open source data to measure street walkability and bikeability in china: A case of four cities. *Transportation Research Record: Journal of the Transportation Research Board*, 2672:63–75.

Gyawali, S. and Qian, Y. (2019). Misbehavior detection using machine learning in vehicular communication networks. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–6.

Hachmann, S., Arsanjani, J. J., and Vaz, E. (2018). Spatial data for slum upgrading: Volunteered geographic information and the role of citizen science. *Habitat International*, 72:18–26. Regional Intelligence: A new kind of GIScience.

Hamidi, Z., Camporeale, R., and Caggiani, L. (2019). Inequalities in access to bike-and-ride opportunities: Findings for the city of malmö. *Transportation Research Part A: Policy and Practice*, 130:673–688.

Han, T., Peng, Q., Zhu, Z., Shen, Y., Huang, H., and Abid, N. N. (2020). A pattern representation of stock time series based on dtw. *Physica A: Statistical Mechanics and its Applications*, 550:124161.

Hardinghaus, M. and Weschke, J. (2022). Attractive infrastructure for everyone? different preferences for route characteristics among cyclists. *Transportation Research Part D: Transport and Environment*, 111:103465.

Heinen, E., van Wee, B., and Maat, K. (2010). Commuting by bicycle: An overview of the literature. *Transport Reviews*, 30(1):59–96.

Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., and Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap. *Nature Communications*, 14:3985.

Hochmair, H. H., Zielstra, D., and Neis, P. (2015). Assessing the completeness of bicycle trail and lane features in openstreetmap for the united states. *Transactions in GIS*, 19(1):63–81.

Hood, J., Sall, E., and Charlton, B. (2011). A gps-based bicycle route choice model for san francisco, california. *Transportation Letters*, 3:63–75.

Huber, S., Lindemann, P., and Schröter, B. (2024). Safety and bicycle route choice: To what extent do accident risk and perceived safety influence bicycle route choice? *Transportation Engineering*, 18:100240.

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., and Jo, O. (2020). Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health*, 8.

Jackson, P. A. A. P. M. S. A. W. T. M. A. B. M. and Zhang, L. (2016). Quality assessment of openstreetmap data using trajectory mining. *Geo-spatial Information Science*, 19:56–68.

Jiang, J., Lai, S., Jin, L., and Zhu, Y. (2022). Dsdtw: Local representation learning with deep soft-dtw for dynamic signature verification. *IEEE Transactions on Information Forensics and Security*, 17:2198–2212.

Jittrapirom, P., Caiati, V., Feneri, A. M., Ebrahimigharehbaghi, S., Gonzalez, M. A., and Narayan, J. (2017). Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. *Urban Planning*, 2.

Kamel, M. B., Sayed, T., and Bigazzi, A. (2020). A composite zonal index for biking attractiveness and safety. *Accident Analysis & Prevention*, 137:105439.

KAPPO (2016). Ciclistas - KAPPO Bike. In KAPPO Bike `https://www.kappo.bike/web/app.php`. [Accessed 05-05-2021].

Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378.

Krenn, P. J., Oja, P., and Titze, S. (2015). Development of a bikeability index to assess the bicycle-friendliness of urban environments. *Open Journal of Civil Engineering*, 05:451–459.

lab, U.-S. (2020). Declives-redeviaria. In Github `https://github.com/U-Shift/Declives-RedeViaria`. [Accessed 28-05-2021].

Lee, K. and Sener, I. N. (2021). Strava metro data for bicycle monitoring: a literature review. *Transport Reviews*, 41:27–47.

Leong, C., Lin, S., Tan, F., and Yu, J. (2024). Coordination in a digital platform organization. *Information Systems Research*, 35(1):363–393.

Li, R., Ji, S., and Wang, Y. (2019). Research on cycling distance based on cyclist perception. *IOP Conference Series: Materials Science and Engineering*, 688:022034.

Lin, J.-J. and Wei, Y.-H. (2018). Assessing area-wide bikeability: A grey analytic network process. *Transportation Research Part A: Policy and Practice*, 113:381–396.

Long, Y. and Zhao, J. (2020). What makes a city bikeable? a study of intercity and intracity patterns of bicycle ridership using mobike big data records. *Built Environment*, 46:55–75.

Lovelace, R., Goodman, A., Aldred, R., Berkoff, N., Abbas, A., and Woodcock, J. (2017). The propensity to cycle tool: An open source online system for sustainable transport planning. *Journal of Transport and Land Use*, 10:505–528.

Lowry, M. B., Callister, D., Gresham, M., and Moore, B. (2012). Assessment of communitywide bikeability with bicycle level of service. *Transportation Research Record: Journal of the Transportation Research Board*, 2314:41–48.

Ma, L. and Dill, J. (2015). Associations between the objective and perceived built environment and bicycling for transportation. *Journal of Transport & Health*, 2:248–255.

Maive, S. (2023). 2023's best biking cities in the u.s. - lawnstarter ranking. In LawnStarter `https://www.lawnstarter.com/blog/studies/best-biking-cities/`. [Accessed 22-01-2024].

Manum, B., Nordström, T., Gil, J., Nilsson, L., and Marcus, L. (2017). Modelling bikeability; space syntax based measures applied in examining speeds and flows of bicycling in gothenburg. *Proceedings - 11th International Space Syntax Symposium, SSS 2017*, pages 89.1–89.16.

Mapbox (2020). Map Matching API | API Docs. In `https://docs.mapbox.com/api/navigation/map-matching/`. [Accessed 25-03-2022].

Mekuria, M. C., Furth, P. G., and Nixon, H. (2012). Low-stress bicycling and network connectivity. Technical Report CA-MTI-12-1005, Mineta Transportation Institute.

Menghini, G., Carrasco, N., Schüssler, N., and Axhausen, K. (2010). Route choice of cyclists in zurich. *Transportation Research Part A: Policy and Practice*, 44:754–765.

Miah, M. M., Mattingly, S. P., and Hyun, K. K. (2023). Evaluation of bicycle network connectivity using graph theory and level of traffic stress. *Journal of Transportation Engineering, Part A: Systems*, 149.

Millard-Ball, A., Hampshire, R. C., and Weinberger, R. R. (2019). Map-matching poor-quality gps data in urban environments: the pgmapmatch package. *Transportation Planning and Technology*, 42(6):539–553.

Minghini, M. and Frassinelli, F. (2019). Openstreetmap history for intrinsic quality assessment: Is osm up-to-date? *Open Geospatial Data Software and Standards*, 4:1–17.

Muhs, C. D. and Clifton, K. J. (2016). Do characteristics of walkable environments support bicycling? toward a definition of bicycle-supported development. *Journal of Transport and Land Use*.

Murphy, J., Pao, Y., and Yuen, A. (2019). Map matching when the map is wrong: Efficient on/off road vehicle tracking and map learning. In *IWCTS 2019 - Proceedings of the 12th International Workshop on Computational Transportation Science*. Association for Computing Machinery, Inc.

Murray, C. (2021). New to ride report: Self-draw aois and multiple metrics charts. In `https://www.ridereport.com/blog/product-updates-april-2021`. [Accessed 18-05-2021].

Márquez, L. and Soto, J. J. (2021). Integrating perceptions of safety and bicycle theft risk in the analysis of cycling infrastructure preferences. *Transportation Research Part A: Policy and Practice*, 150:285–301.

Müller, M. (2007). Dynamic time warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg.

Nelson, T., Ferster, C., Laberee, K., Fuller, D., and Winters, M. (2021). Crowdsourced data for bicycling research and practice. *Transport Reviews*, 41:97–114.

Nikolaeva, A., te Brömmelstroet, M., Raven, R., and Ranson, J. (2019). Smart cycling futures: Charting a new terrain and moving towards a research agenda. *Journal of Transport Geography*, 79:102486.

Nunes, P., Moura, A., Santos, J. P., and Completo, A. (2021). A simulated annealing algorithm to solve the multi-objective bike routing problem. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 39–45. IEEE.

OMF (2021). About MDS. In Open Mobility Foundation `https://www.openmobilityfoundation.org/about-mds/`. [Accessed 04-05-2021].

Panek, J. and Netek, R. (2019). Collaborative mapping and digital participation: A tool for local empowerment in developing countries. *Information*, 10:255.

Pardos, Z. A. and Kao, K. (2015). Moocrp: An open-source analytics platform. *L@S 2015 - 2nd ACM Conference on Learning at Scale*, pages 103–110.

PCT.BIKE (2017). Welcome to the Propensity to Cycle Tool (PCT). In PCT.BIKE `https://www.pct.bike/`. [Accessed 04-05-2021].

PeopleForBikes (2023). Peopleforbikes 2023 city ratings. In PeopleForBikes `https://cityratings.peopleforbikes.org/`. [Accessed 22-01-2024].

Perkins, C. (2014). Plotting practices and politics: (im)mutable narratives in openstreetmap. *Transactions of the Institute of British Geographers*, 39:304–317.

Porter, A. K., Kohl, H. W., Pérez, A., Reininger, B., Gabriel, K. P., and Salvo, D. (2020). Bikeability: Assessing the objectively measured environment in relation to recreation and transportation bicycling. *Environment and Behavior*, 52:861–894.

Poudel, N. and Singleton, P. A. (2022). Preferences for roundabout attributes among us bicyclists: A discrete choice experiment. *Transportation Research Part A: Policy and Practice*, 155:316–329.

Prato, C. G., Halldórsdóttir, K., and Nielsen, O. A. (2018). Evaluation of land-use and transport network effects on cyclists' route choices in the copenhagen region in value-of-distance space. *International Journal of Sustainable Transportation*, 12:770–781.

Quinn, S. D. and MacEachren, A. M. (2018). A geovisual analytics exploration of the openstreetmap crowd. *Cartography and Geographic Information Science*, 45:140–155.

Reggiani, G., van Oijen, T., Hamedmoghadam, H., Daamen, W., Vu, H. L., and Hoogendoorn, S. (2022). Understanding bikeability: a methodology to assess urban networks. *Transportation*, 49:897–925.

Report, R. (2020). Ride Report: The platform to manage micromobility. In Ride Report `https://www.ridereport.com/`. [Accessed 04-05-2021].

Resch, B., Puetz, I., Bluemke, M., Kyriakou, K., and Miksch, J. (2020). An interdisciplinary mixed-methods approach to analyzing urban spaces: The case of urban walkability and bikeability. *International Journal of Environmental Research and Public Health*, 17:6994.

Romanillos, G. and Austwick, M. Z. (2016). Madrid cycle track: visualizing the cyclable city. *Journal of Maps*, 12:1218–1226.

Saghapour, T., Moridpour, S., and Thompson, R. G. (2017). Measuring cycling accessibility in metropolitan areas. *International Journal of Sustainable Transportation*, 11:381–394.

Sasaki, Y., Yu, J., and Ishikawa, Y. (2019). Road segment interpolation for incomplete road data. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8. IEEE.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis (texts in statistical science)*. Texts in Statistical Science Series. CHAPMAN & HALL/CRC.

Schmid-Querg, J., Keler, A., and Grigoropoulos, G. (2021). The munich bikeability index: A practical approach for measuring urban bikeability. *Sustainability*, 13:428.

Schröder-Bergen, S., Glasze, G., Michel, B., and Dammann, F. (2022). De/colonizing openstreetmap? local mappers, humanitarian and commercial actors and the changing modes of collaborative mapping. *GeoJournal*, 87:5051–5066.

Schweizer, J., Bernardi, S., and Rupi, F. (2016). Map-matching algorithm applied to bicycle global positioning system traces in bologna. *IET Intelligent Transport Systems*, 10:244–250.

Score, W. (2018). Bike score methodology. In Walk Score `https://www.walkscore.com/bike-score-methodology.shtml`. [Accessed 10-05-2021].

Sener, I. N., Eluru, N., and Bhat, C. R. (2009). An analysis of bicycle route choice preferences in texas, us. *Transportation*, 36:511–539.

Shan, Z., Wu, H., Sun, W., and Zheng, B. (2015). Cobweb. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 927–937. ACM.

Silva, C., Marques, J. S., Lopes, M., and Dias, A. M. (2021). The gross potential for cycling: planning for human scale urban mobility. In Mladenović, M. N., Toivonen, T., Willberg, E., and Geurs, K. T., editors, *Transport in Human Scale Cities*, pages 157–168. Edward Elgar Publishing.

Silva, C., Teixeira, J., Proença, A., Bicalho, T., and Cunha, I. (2018). The potential for cycling assessment method - final report of the generation. Mobi Project. CITTA.

Silva, C., Teixeira, J., Proença, A., Bicalho, T., Cunha, I., and Aguiar, A. (2019). Revealing the cycling potential of starter cycling cities: Usefulness for planning practice. *Transport Policy*, 81:138–147.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315.

Stamatiadis, N., Pappalardo, G., and Cafiso, S. (2017). Use of technology to improve bicycle mobility in smart cities. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 86–91. IEEE.

Stratta, P. (2018). Towards a smarter cycling. on the brink of a smart (r)evolution. In European Cyclists' Federation `https://ecf.com/what-we-do/cycling-new-technologies/towards-smarter-cycling`. [Accessed 30-06-2020].

Strava (2019). Strava Metro | Map. In Strava Metro `https://metroview.strava.com/map/demo#eyJtIjoicm91dGVzIn0=`. [Accessed 25-05-2021].

StreetLight (2019). Transportation analytics on demand. In StreetLight Data `https://www.streetlightdata.com/`. [Accessed 05-05-2021].

Tabet, F., Pentyala, S., Patel, B. H., Hendawi, A., Cao, P., Song, A., Govind, H., and Ali, M. (2021). Osmrunner : A system for exploring and fixing osm connectivity. In *2021 22nd IEEE International Conference on Mobile Data Management (MDM)*, pages 193–200.

Takahashi, K., Yamamoto, K., Kuchiba, A., and Koyama, T. (2022). Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied Intelligence*, 52(5):4961–4972.

Tang, J., Deng, M., Huang, J., Liu, H., and Chen, X. (2019). An automatic method for detection and update of additive changes in road network with gps trajectory data. *ISPRS International Journal of Geo-Information*, 8:411.

Tao, Y., Both, A., Silveira, R. I., Buchin, K., Sijben, S., Purves, R. S., Laube, P., Peng, D., Toohey, K., and Duckham, M. (2021). A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing*, 58:643–669.

Ton, D., Cats, O., Duives, D., and Hoogendoorn, S. (2017). How do people cycle in amsterdam, netherlands?: Estimating cyclists' route choice determinants with gps data from an urban area. *Transportation Research Record: Journal of the Transportation Research Board*, 2662:75–82.

Toohey, K. and Duckham, M. (2015). Trajectory similarity measures. *SIGSPATIAL Special*, 7(1):43–50.

Tran, P. T., Zhao, M., Yamamoto, K., Minet, L., Nguyen, T., and Balasubramanian, R. (2020). Cyclists' personal exposure to traffic-related air pollution and its influence on bikeability. *Transportation Research Part D: Transport and Environment*, 88:102563.

Trogh, J., Botteldooren, D., De Coensel, B., Martens, L., Joseph, W., and Plets, D. (2022). Map matching and lane detection based on markovian behavior, gis, and imu data. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):2056–2070.

Tsakalidis, A., Gkoumas, K., and Pekár, F. (2020). Digital transformation supporting transport decarbonisation: Technological developments in eu-funded research and innovation. *Sustainability*, 12:3762.

Uijtdewilligen, T., Ulak, M. B., Wijlhuizen, G. J., and Geurs, K. T. (2024). Effects of crowding on route preferences and perceived safety of urban cyclists in the netherlands. *Transportation Research Part A: Policy and Practice*, 183:104030.

Vale, D. (2020). Calculate accessibility from an od matrix on python (version 2.0). In Github `https://github.com/davidsvale/calculate-accessibility`. [Accessed 24-06-2021].

van der Waerden, P., van der Waerden, J., and Gebhard, S. (2024). Intersection based innovations and cyclists' route choice decisions in urban areas. *Transport Policy*, 146:205–214.

van Exel, M. (2012). Maproulette. In MapRoulette `https://maproulette.org/`. [Accessed 18-04-2025].

Vierø, A. R., Vybornova, A., and Szell, M. (2024a). Bikedna: A tool for bicycle infrastructure data and network assessment. *Environment and Planning B: Urban Analytics and City Science*, 51:512–528.

Vierø, A. R., Vybornova, A., and Szell, M. (2024b). How good is open bicycle network data? a countrywide case study of denmark. *Geographical Analysis*.

Wasserman, D., Rixey, A., Zhou, X. E., Levitt, D., and Benjamin, M. (2019). Evaluating openstreetmap's performance potential for level of traffic stress analysis. *Transportation Research Record*, 2673(4):284–294.

Wilmott, C. (2019). "mapping-with": The politics of (counter-)classification in openstreetmap. *Cartographic Perspectives*.

Winters, M., Brauer, M., Setton, E. M., and Teschke, K. (2013). Mapping bikeability: a spatial tool to support sustainable travel. *Environment and Planning B: Planning and Design*, 40:865–883.

Wysling, L. and Purves, R. S. (2022). Where to improve cycling infrastructure? assessing bicycle suitability and bikeability with open data in the city of paris. *Transportation Research Interdisciplinary Perspectives*, 15:100648.

Yang, C. and Gidófalvi, G. (2018). Fast map matching, an algorithm integrating hidden markov model with precomputation. *International Journal of Geographical Information Science*, 32:547–570.

Yang, C., Yue, P., Gong, J., Li, J., and Yan, K. (2024). Detecting road network errors from trajectory data with partial map matching and bidirectional recurrent neural network model. *International Journal of Geographical Information Science*, 38:478–502.

Yeboah, G. and Alvanides, S. (2015). Route choice analysis of urban cycling behaviors using open-streetmap: Evidence from a british urban environment. In Jokar Arsanjani, J., Zipf, A., Mooney, P., and Helbich, M., editors, *OpenStreetMap in GIScience: Experiences, Research, and Applications*, pages 189–210. Springer International Publishing, Cham.

Yoo, Y., Boland, R. J., Lyytinen, K., and Majchrzak, A. (2012). Organizing for innovation in the digitized world. *Organization Science*, 23.

Zayed, M. A. (2016). Towards an index of city readiness for cycling. *International Journal of Transportation Science and Technology*.

Zhang, H. and Malczewski, J. (2018). Accuracy evaluation of the canadian openstreetmap road networks. *International Journal of Geospatial and Environmental Research*, 5:Article 1.

Zimmermann, M., Mai, T., and Frejinger, E. (2017). Bike route choice modeling using gps data without choice sets of paths. *Transportation Research Part C: Emerging Technologies*, 75:183–196.

Zittrain, J. (2005). The generative internet. *Harvard Law Review*, 119.

Łukawska, M., Paulsen, M., Rasmussen, T. K., Jensen, A. F., and Nielsen, O. A. (2023). A joint bicycle route choice model for various cycling frequencies and trip distances based on a large crowdsourced gps dataset. *Transportation Research Part A: Policy and Practice*, 176:103834.

Appendix A: Binary Classifiers Extended Results

Table with the extended results from applying binary classification models to detect GPS traces containing gaps between OSM and cycling routes.

Table 19: Binary Classification Results

model	version	precision	recall	f1	accuracy
XGB	ANOVA	0,883	0,931	0,906	0,893
XGB	ANOVA-STD	0,883	0,931	0,906	0,893
DT	info gain	0,888	0,915	0,902	0,889
KNN	Corr	0,876	0,923	0,899	0,885
KNN	None	0,876	0,923	0,899	0,885
KNN	ANOVA	0,887	0,908	0,897	0,885
XGB	Corr	0,887	0,908	0,897	0,885
Ada	None	0,906	0,885	0,895	0,885
Ada	None-STD	0,906	0,885	0,895	0,885
LR	info gain	0,948	0,838	0,890	0,885
LR	None-STD	0,956	0,831	0,889	0,885
LR	ANOVA-STD	0,956	0,831	0,889	0,885
RF	Corr	0,892	0,892	0,892	0,880
Ada	ANOVA	0,898	0,885	0,891	0,880
Ada	ANOVA-STD	0,898	0,885	0,891	0,880
KNN	info gain	0,898	0,885	0,891	0,880
LR	None	0,947	0,831	0,885	0,880
DT	Corr	0,869	0,915	0,891	0,876
DT	None	0,874	0,908	0,891	0,876
DT	None-STD	0,874	0,908	0,891	0,876
Ada	Corr	0,885	0,892	0,889	0,876

Table 19 continued from previous page

model	version	precision	recall	f1	accuracy
Ada	info gain	0,911	0,862	0,885	0,876
SVC	ANOVA-STD	0,917	0,854	0,884	0,876
LR	ANOVA	0,939	0,831	0,882	0,876
SVC	None-STD	0,947	0,823	0,881	0,876
XGB	info gain	0,885	0,885	0,885	0,872
LR	Corr	0,939	0,823	0,877	0,872
LR	PCA	0,955	0,808	0,875	0,872
DT	ANOVA	0,872	0,892	0,882	0,868
DT	ANOVA-STD	0,872	0,892	0,882	0,868
XGB	None	0,872	0,892	0,882	0,868
XGB	None-STD	0,872	0,892	0,882	0,868
RF	info gain	0,884	0,877	0,880	0,868
SVC	PCA	0,945	0,800	0,867	0,863
RF	None	0,870	0,877	0,874	0,859
RF	None-STD	0,870	0,877	0,874	0,859
RF	ANOVA	0,870	0,877	0,874	0,859
RF	ANOVA-STD	0,870	0,877	0,874	0,859
KNN	ANOVA-STD	0,908	0,831	0,867	0,859
KNN	None-STD	0,961	0,754	0,845	0,846
Ada	PCA	0,883	0,815	0,848	0,838
RF	PCA	0,894	0,777	0,831	0,825
NB	info gain	0,978	0,700	0,816	0,825
XGB	PCA	0,885	0,769	0,823	0,816
DT	PCA	0,922	0,731	0,815	0,816
NB	Corr	0,968	0,692	0,807	0,816
KNN	PCA	0,906	0,738	0,814	0,812
NB	None	0,967	0,685	0,802	0,812
NB	ANOVA	0,967	0,677	0,796	0,808
NB	ANOVA-STD	0,957	0,677	0,793	0,803
NB	None-STD	0,967	0,669	0,791	0,803

Table 19 continued from previous page

model	version	precision	recall	f1	accuracy
SVC	info gain	0,735	0,938	0,824	0,778
NB	PCA	0,877	0,546	0,673	0,705
SVC	ANOVA	0,604	0,985	0,749	0,632
SVC	Corr	0,578	0,992	0,731	0,594
SVC	None	0,573	1,000	0,728	0,585

Appendix B: Multi-class Classifiers Extended Results

Table with the extended results from applying multi-class classification models to detect GPS traces containing gaps between OSM and cycling routes and classify each gap accordingly to its root cause.

Table 20: Multi-class classification Results

model	version	p ma	r ma	f1 ma	p mi	r mi	f1 mi	Acc
RF	None	0,63	0,57	0,59	0,71	0,71	0,71	0,71
RF	Info-Gain	0,57	0,54	0,53	0,68	0,68	0,68	0,68
RF	ANOVA	0,59	0,53	0,54	0,68	0,68	0,68	0,68
RF	Corr	0,47	0,41	0,41	0,68	0,68	0,68	0,68
DT	ANOVA	0,54	0,51	0,50	0,68	0,68	0,68	0,68
LR	PCA	0,45	0,40	0,39	0,68	0,68	0,68	0,68
KNN	ANOVA	0,44	0,40	0,38	0,67	0,67	0,67	0,67
KNN	Corr	0,42	0,38	0,37	0,66	0,66	0,66	0,66
SVC	PCA	0,40	0,40	0,37	0,66	0,66	0,66	0,66
DT	None	0,42	0,41	0,39	0,65	0,65	0,65	0,65
KNN	Info-Gain	0,39	0,36	0,34	0,65	0,65	0,65	0,65
RF	None-SMOTE-Under	0,54	0,60	0,54	0,65	0,65	0,65	0,65
LR	PCA-SMOTE-under	0,43	0,44	0,43	0,65	0,65	0,65	0,65
NB	None	0,53	0,44	0,45	0,64	0,64	0,64	0,64
LR	Corr	0,36	0,35	0,32	0,64	0,64	0,64	0,64
DT	Corr	0,42	0,42	0,41	0,64	0,64	0,64	0,64
DT	Info-Gain	0,52	0,48	0,48	0,64	0,64	0,64	0,64
RF	CORR-SMOTE-under	0,46	0,51	0,47	0,64	0,64	0,64	0,64
KNN	None	0,36	0,35	0,33	0,64	0,64	0,64	0,64
NB	Info-Gain	0,44	0,40	0,40	0,64	0,64	0,64	0,64
NB	CORR-SMOTE-under	0,49	0,44	0,45	0,64	0,64	0,64	0,64

Table 20 continued from previous page

model	version	p ma	r ma	f1 ma	p mi	r mi	f1 mi	Acc
NB	Corr	0,44	0,41	0,41	0,63	0,63	0,63	0,63
RF	PCA	0,39	0,36	0,34	0,63	0,63	0,63	0,63
NB	ANOVA	0,46	0,40	0,40	0,63	0,63	0,63	0,63
NB	PCA	0,51	0,38	0,40	0,63	0,63	0,63	0,63
KNN	PCA	0,40	0,35	0,35	0,63	0,63	0,63	0,63
RF	Info Gain-SMOTE	0,47	0,51	0,48	0,63	0,63	0,63	0,63
DT	PCA	0,37	0,37	0,34	0,63	0,63	0,63	0,63
NB	None-SMOTE-Under	0,47	0,41	0,42	0,63	0,63	0,63	0,63
NB	ANOVA-SMOTE-Under	0,44	0,39	0,39	0,63	0,63	0,63	0,63
NB	None-SMOTE	0,52	0,44	0,45	0,63	0,63	0,63	0,63
LR	ANOVA	0,23	0,30	0,25	0,62	0,62	0,62	0,62
NB	Info Gain-SMOTE	0,41	0,36	0,37	0,62	0,62	0,62	0,62
LR	None-SMOTE-Under	0,43	0,45	0,42	0,62	0,62	0,62	0,62
LR	CORR-SMOTE-under	0,31	0,38	0,34	0,62	0,62	0,62	0,62
RF	ANOVA-SMOTE-Under	0,51	0,60	0,52	0,62	0,62	0,62	0,62
RF	None-SMOTE	0,45	0,51	0,46	0,62	0,62	0,62	0,62
RF	Info Gain-SMOTE-under	0,43	0,46	0,44	0,61	0,61	0,61	0,61
NB	ANOVA-SMOTE	0,44	0,39	0,39	0,61	0,61	0,61	0,61
RF	ANOVA-SMOTE	0,50	0,57	0,51	0,61	0,61	0,61	0,61
LR	None	0,28	0,30	0,28	0,61	0,61	0,61	0,61
NB	Info Gain-SMOTE-under	0,52	0,34	0,35	0,61	0,61	0,61	0,61
NB	CORR-SMOTE	0,45	0,46	0,44	0,60	0,60	0,60	0,60
NB	PCA-SMOTE-under	0,45	0,37	0,39	0,60	0,60	0,60	0,60
SVC	PCA-SMOTE-under	0,44	0,45	0,44	0,60	0,60	0,60	0,60
RF	PCA-SMOTE	0,44	0,46	0,44	0,58	0,58	0,58	0,58
LR	ANOVA-SMOTE-Under	0,38	0,37	0,37	0,58	0,58	0,58	0,58
DT	ANOVA-SMOTE-Under	0,46	0,51	0,46	0,58	0,58	0,58	0,58
LR	PCA-SMOTE	0,42	0,45	0,42	0,57	0,57	0,57	0,57
RF	CORR-SMOTE	0,40	0,47	0,41	0,57	0,57	0,57	0,57
DT	Info Gain-SMOTE-under	0,41	0,48	0,43	0,57	0,57	0,57	0,57

Table 20 continued from previous page

model	version	p ma	r ma	f1 ma	p mi	r mi	f1 mi	Acc
DT	CORR-SMOTE-under	0,37	0,39	0,37	0,57	0,57	0,57	0,57
KNN	Info Gain-SMOTE-under	0,42	0,42	0,38	0,56	0,56	0,56	0,56
DT	Info Gain-SMOTE	0,35	0,35	0,34	0,56	0,56	0,56	0,56
KNN	None-SMOTE	0,37	0,39	0,37	0,56	0,56	0,56	0,56
RF	PCA-SMOTE-under	0,39	0,40	0,39	0,55	0,55	0,55	0,55
KNN	CORR-SMOTE-under	0,33	0,32	0,32	0,55	0,55	0,55	0,55
DT	None-SMOTE-Under	0,44	0,48	0,44	0,54	0,54	0,54	0,54
KNN	None-SMOTE-Under	0,37	0,41	0,38	0,53	0,53	0,53	0,53
DT	ANOVA-SMOTE	0,44	0,51	0,44	0,53	0,53	0,53	0,53
SVC	PCA-SMOTE	0,41	0,43	0,41	0,53	0,53	0,53	0,53
LR	CORR-SMOTE	0,37	0,36	0,34	0,53	0,53	0,53	0,53
DT	CORR-SMOTE	0,35	0,42	0,35	0,52	0,52	0,52	0,52
DT	PCA-SMOTE	0,36	0,36	0,35	0,52	0,52	0,52	0,52
KNN	PCA-SMOTE-under	0,40	0,41	0,40	0,52	0,52	0,52	0,52
DT	None-SMOTE	0,36	0,40	0,35	0,52	0,52	0,52	0,52
KNN	Info Gain-SMOTE	0,36	0,40	0,36	0,51	0,51	0,51	0,51
LR	None-SMOTE	0,34	0,34	0,33	0,51	0,51	0,51	0,51
SVC	CORR-SMOTE-under	0,25	0,18	0,13	0,50	0,50	0,50	0,50
SVC	None-SMOTE	0,25	0,18	0,13	0,50	0,50	0,50	0,50
SVC	CORR-SMOTE	0,17	0,18	0,13	0,49	0,49	0,49	0,49
DT	PCA-SMOTE-under	0,37	0,38	0,36	0,49	0,49	0,49	0,49
KNN	ANOVA-SMOTE-Under	0,32	0,33	0,32	0,49	0,49	0,49	0,49
SVC	None	0,41	0,18	0,14	0,48	0,48	0,48	0,48
SVC	Corr	0,41	0,18	0,14	0,48	0,48	0,48	0,48
SVC	ANOVA	0,41	0,18	0,14	0,48	0,48	0,48	0,48
KNN	CORR-SMOTE	0,29	0,28	0,27	0,48	0,48	0,48	0,48
SVC	None-SMOTE-Under	0,41	0,18	0,14	0,48	0,48	0,48	0,48
KNN	PCA-SMOTE	0,37	0,40	0,37	0,48	0,48	0,48	0,48
KNN	ANOVA-SMOTE	0,34	0,38	0,35	0,48	0,48	0,48	0,48
SVC	ANOVA-SMOTE-Under	0,29	0,18	0,14	0,47	0,47	0,47	0,47

Table 20 continued from previous page

model	version	p ma	r ma	f1 ma	p mi	r mi	f1 mi	Acc
LR	ANOVA-SMOTE	0,23	0,32	0,25	0,47	0,47	0,47	0,47
SVC	Info-Gain	0,31	0,19	0,15	0,47	0,47	0,47	0,47
SVC	Info Gain-SMOTE	0,29	0,34	0,26	0,47	0,47	0,47	0,47
SVC	Info Gain-SMOTE-under	0,19	0,19	0,16	0,47	0,47	0,47	0,47
NB	PCA-SMOTE	0,44	0,39	0,37	0,42	0,42	0,42	0,42
SVC	ANOVA-SMOTE	0,40	0,26	0,17	0,25	0,25	0,25	0,25
LR	Info-Gain	0,25	0,27	0,16	0,23	0,23	0,23	0,23
LR	Info Gain-SMOTE-under	0,10	0,26	0,06	0,07	0,07	0,07	0,07
LR	Info Gain-SMOTE	0,04	0,26	0,05	0,06	0,06	0,06	0,06