UNIVERSITÀ
DI TRENTO

Master's Degree in Data Science

Mapping bikeability in Berlin to promote sustainable travel

Supervisor
Luigi Amedeo Bianchi

Candidate
Marta Fattorel

Co-supervisors
Diego Giuliani
Maurizio Napolitano

Academic year
2020/2021

Acknowledgements

First and foremost I would like to thank my research supervisor, Professor Luigi Amedeo Bianchi, for his continuous support throughout this project and his insightful advices. I would also like to thank my co-supervisors Professor Diego Giuliani and Maurizio Napolitano.

I would like to express my gratitude to Targomo and all my colleagues, for the warm welcome to their team. Special thanks to Henning Hollburg, Hermann Schwarting and Florian Bersch, who made my internship experience at Targomo possible. In particular, my research interest and some methodologies and tools applied in this work are inspired by my working tasks and reflect what I have learned during the past 10 months at Targomo.

My gratitude extends to the University of Trento and all my professors, who introduced me to the data science world and conveyed fundamental theoretical and technical knowledge to accomplish this research. I would also like to thank my course mates, for the cooperative and collaborative atmosphere we built. Special thanks to Anna, who since the first day shared with me worries and concerns that the transition to a scientific field of studies implied, and to Fabio, my study and projects mate.

I would like to thank my friends and roommates, who supported me closely and from afar and who shared with me this important stage of life. Thanks to Adua, Riccardo, Micol, Sofia, Laura, Chiara, Marta, Romina, Chiara and Elisa.

Last but not least, I would like to thank my parents, who always supported my choices, my sister and my grandmothers. Thank you for your love and encouragement.

Table of contents

Chapter 1
Introduction

Supporting and promoting cycling is a key factor in order to achieve the UN Sustainable Development Goals (SDG), approved in 2015 within the 2030 Agenda for Sustainable Development [69]. Cycling represents a sustainable mode of transportation of people and goods and thus it helps fight climate change. Indeed, it reduces traffic congestion and as a consequence also air and noise pollution caused by motorised vehicles. Furthermore, cycling helps preserve biodiversity and habitat since it does not require new land development [80]. When combined with e-cycling technologies and the use of other public transport, cycling increases safety, sustainability and inclusion of cities [69]. Moreover, it is an affordable and cheap mode of travelling, which allows people who cannot afford alternative means of transportation, to satisfy their daily needs. Furthermore, the cycling industry promotes economic growth through the creation of new jobs and services in various sectors such as bicycle retail, bicycle infrastructure, bicycle tourism and bicycle services. In particular, studies [9] state that more local jobs, for lower skilled workers and more jobs per euro spent could be created by cycling compared to other sectors. This would promote the local economy as cyclists go more to local restaurants, shops and other local businesses and increase social inclusion since lots of bike-related jobs do not need high qualification and can thus represent opportunities for people who are disadvantaged in the labour market due to their low skills [9]. Concerning health benefits, scientific studies proved that cycling enhances a healthy individual lifestyle. On the one hand, cycling decreases the risk of Cardiovascular and Coronary heart diseases, cancer mortality among adults and elderly, diabetes, overweight and obesity. On the other hand, it improves cardiorespiratory fitness [22, 52, 58]. Research demonstrates that overall, health benefits of cycling outweigh possible risks due to air pollution and accidents [13, 22, 80]. Moreover, more bicycles on the streets would lead to more visibility for cyclists and thus fewer cars and traffic speed. This implies a reduction in the number of accidents, which increases safety [80].

In recent years, cycling has become a socio-cultural phenomenon since it is not only perceived as a transport mode but also as an individual choice that mirrors certain values and a certain way of life [60]. Lots of initiatives such as the *CITY CYCLING* campaign and the *Berlin Mobility Act*, which will be presented in the next chapters, have been organised by both institutions and citizens in order to raise

public awareness towards sustainable travel.

Nowadays, traffic engineers, urban planners and architects have more awareness of the benefits of cycling and of the importance of promoting a bike-friendly environment through urban design and planning measures [47]. Researchers found that more walkable and bikeable environments have a positive influence on physical activity and people who live in such neighbourhoods are less likely to have chronic diseases and suffer from obesity and travel less by car. In addition, cities with more cycle lanes have higher cycle rates [77]. In fact, human behaviour is the result of attitudes, preferences and perceptions of the surrounding environment, which implies that people seek places and engage in activities that make them feel safe and comfortable. Therefore, urban planning and design choices have a significant impact on transport choices and lifestyle [39, 59].

Taking the aforementioned evidence into consideration, this work recognises the importance and the urgent need for cities to promote healthy and sustainable alternatives to motorised traffic. In particular, the aim of this study is to explore and investigate bikeability through social and geospatial data science tools. As will be discussed in the following chapters, bikeability is a multidimensional concept that assesses the perceived comfort, safety and convenience in accessing important destinations [44]. Accessibility and safety are the two bikeability dimensions, which will be measured focusing on the city of Berlin. To be more specific, accessibility denotes here areas characterised by a high bicycle flow, which most probably represent important connections to reach relevant destinations. Safety refers instead to areas where the crash risk for cyclists is limited. The main goal of this work is to detect the most significant drivers of accessibility and safety, meaning the key factors contributing to making streets popular with cyclists on the one hand, and less dangerous for them on the other hand. Bicycle flow data will be used as a proxy for measuring accessibility, while bike accidents data will be a proxy for safety. After an exploratory analysis aiming at investigating the characteristics of bike accidents and their relation with bike flow, statistical models are built to predict bicycle flow and bicycle accidents. In particular, bike lane score, hill score, POIs (Points of Interest) score, parks, intersection density, traffic lights, bicycle parking capacity, car traffic, car speed, betweenness centrality, network coverage and population will be the independent variables selected to predict bicycle flow and bicycle accidents through regression and classification models. Detecting and understanding the main factors that contribute to increasing bike flow and reducing bike accidents, and developing models that allow predicting such phenomena, are key elements to suggest coherent policies to promote bikeability.

This specific research interest comes from my internship experience at Targomo [28], a tech start-up based in Berlin, which is specialised in location intelligence. Working at Targomo has raised my awareness of the fundamental role played by geospatial data and analytics in driving private and public decisions. Moreover, I had the chance to work on several projects and this gave me valuable insights on the current hot topics regarding spatial analytics and the corresponding analysis techniques. In particular, the city of Berlin is characterised by a quite strong cycling culture and recently, as will be discussed later, several initiatives have been carried out in order to further develop the bicycle network. According to the 2019 Copenhagenize index, Berlin is the 15^{th} most bike-friendly city after cities like Copenhagen, Paris and Bogotà [31]. Even if it is not a bad position, targeted measures based on social and urban studies would improve the current situation. In addition, a deep understanding of the cycling dynamics of a specific city through a combination of social science and data science methodologies could provide useful insights and models that should be adapted and applied to different contexts. We live in a globalised, interconnected and fast-moving world, where more than ever before, events and actions that happen in a specific place have consequences on other places. It is therefore important to explore and interpret local phenomena in order to address global challenges. Thus, the bi-dimensional bikeability index presented in this work could first of all guide urban planning decisions by informing on which areas of Berlin lack adequate infrastructure and need therefore targeted policies. In the second place, the same approach could be applied in order to adapt the Berlin model to new cities, regions or countries based on their specific characteristics.

After this first introduction on the benefits of cycling, the presentation of the research topic and aim of the study and a brief discussion on the relevance of the chosen topic and the origin of my research interest, the next chapters are structured as follows. In the second chapter, the scientific literature related to bikeability and the specific focus of this research are discussed. A review of various approaches to measure bikeability is presented, including a theoretical definition of the concept. The third chapter describes the origin of the data used to investigate the research topic and the construction of the variables of interest for the analysis. The fourth chapter consists of an exploratory analysis in order to gain a deeper understanding of the available data, the presentation of the statistical techniques and models applied to investigate the research topic and the results obtained. The fifth chapter is dedicated to the discussion of the results and suggested measures to promote bikeability in Berlin. A final chapter is dedicated to the conclusion.

Chapter 2

Literature review and research focus

2.1. Bikeability definition

The concept of bikeability is rather new, it gained interest only at the beginning of this century [36]. Indeed, research on bikeability was inspired by the walkability movement, which began in 2005 and led to the development of a walkability index [11, 36, 56]. Even though the word "bikeability" still has no place in official dictionaries, various definitions of this rather broad concept can be found in the scientific literature. According to the scientific community, bikeability relates to the concepts of bicycle-friendliness, bicycle-suitability, bicycle level of service and cycling quality and it is mainly applied to the fields of transport, urban planning, public health, and well-being. Concerning public health and well-being, bikeability is defined in terms of environmental factors associated with cycling like comfort, convenience, safety, but also physical activity and body weight management [29]. In the field of urban planning, bikeability is concerned with the analysis of a road network and streets characteristics using measures such as connectivity, centrality or space syntax analysis [47]. The field of transport assesses how environments are suitable for cycling, how different geographic areas are accessible by various transportation modes and how many important destinations are reachable within a given threshold. Finally, few studies define bikeability as the actual ability to ride a bike following the traffic rules [11]. Connected to this aspect, the so-called *Cycle Training Program* or simply *Bikeability*, is the Department for Transport's national cycle training program for schoolchildren in England aiming at promoting a bicycle culture by teaching children how to properly ride a bike [72].

Among the existing definitions of bikeability, the one that better suits the research interest of this work is provided by Lowry et al. [44]. Indeed, their definition encompasses both accessibility of important locations and cycling safety, which are the aspects that will be measured in this research. According to the authors, bikeability is "an assessment of an entire bikeway network for perceived comfort and convenience and access to important destinations". In this respect, accessibility, which refers to the ease of reaching important destinations and suitability, which is "an assessment of the perceived comfort and safety of a linear section of bikeway", are indicators to measure bikeability [44].

2.2. MEASURING BIKEABILITY

In the scientific literature, there are several works that present different approaches and methodologies to measure bikeability depending on its specific purpose, research area and context [11]. According to these aspects, different methodological procedures can be implemented. Firstly, concerning the type of factors analysed, a choice needs to be made among hard factors such as bike infrastructure and facilities or soft factors like events and public policies or a combination of both. Secondly, different types of data and information can be considered, namely objective measures of the built environment gathered from geodatabases or audit tools, and subjective perceptions of that environment investigated through interviews, focus groups and surveys. Thirdly, the model can be a broader city level model or a high-resolution grid-based model. Finally, the overall approach could be quantitative aiming at collecting lots of data in a systematic way for generalisation purposes or qualitative, which focuses on a restricted area to allow a rich and detailed data collection aiming at obtaining an in depth understanding of such area. For the purpose of this work seventeen different approaches have been reviewed and compared.

One of the most famous indexes is the *Bike Score* developed by the *Walk Score* initiative [8], which provides commercial web-based measurement tools of walkability and bikeability together with data and other products. This index measures how a location is good for biking on a scale from 0 – 100 considering four equally weighted factors: bike lanes, hills, destinations and road connectivity and bike commuting mode share. This bike score has been used by several researchers to explore and predict relationships with economic and socio-demographics variables. *Bike Score* was proven to have some influence on cycling to work behaviour in different US and Canadian cities [78]. Furthermore, income inequalities have been observed in several Canadian cities when it comes to bikeability. In other words, higher income areas are characterised by greater cycling and more cycling infrastructure than lower income ones [19, 78]. Another well-known bikeability score is developed by Lowry et al. [44]. It is a novel modification of an existing equation used to measure accessibility, which in turn is defined as the ease of reaching important destinations. McNeil [48] explored a methodology to assess bikeability starting from the concept of a *20-min neighborhood*. This concept comes from an innovative idea for urban planning called *20-Minute Living* credited by Mark Edlen, the CEO of Portland-based Edlen & Co. The idea is to create areas where people can get around by walking, bike, public transport or, as last option, by car within 20 minutes. The aim is to offer citizens comfortable and healthy neighbourhoods that enable them to

satisfy all their needs in a reasonable distance, preferably by walking and cycling. This implies building new businesses close to existing houses, but also the other way around. It is also a matter of building more public transit stations and stops and planning more mixed land use to host various activities in a relatively small area [30, 48]. Bike Ottawa implemented a *Stress Map* inspired by the Level of Traffic Stress (LTS) model formalised by Furth et al. [20]. Bike Ottawa is a not-for-profit organisation entirely run by volunteers, aiming at making Ottawa a bike friendly city. The team developed a stress map, which categorises all streets of Ottawa based on their estimated level of traffic stress from a cyclist perspective. Data were retrieved from OpenStreetMap and four categories of traffic stress were built namely "LTS 1 – Suitable for children", "LTS 2 – Low stress", "LTS 3 – Medium stress" and "LTS 4 – High stress" [20, 55]. Hamidi et al. [25] developed a bikeability index, which is more properly an accessibility metric, in association with the concept of intermodality, meaning the integration of two or more transportation modes during the same trip. Grigore et al. [23] started from the assumption that cyclists, when deciding the route to take, look for a balance between minimal distance and maximal quality of streets and intersections. Thus, their bikeability index is a "measure of the ability and convenience in reaching important destinations by bike, based on the travel distance weighted by the perceived safety, -comfort, and -attractiveness of the streets and intersections along the routes" [23]. Lin and Wei [42] developed an area-wide bikeability assessment model (ABAM), which measures how a street block, neighbourhood, community, or village is bike friendly. Their zone-based bikeability index is developed as an analytic network process (ANP), that is a multiple-criteria assessment method that ranks zones considering interdependence among neighbouring zones and among the criteria that define the index itself. Kamel et al. [34] built the so-called Bike Composite Index, which is the combination of two sub-indexes namely Bike Attractiveness Index (BAI) and Bike Safety Index (BSI). A city-level index was developed by the Copenhagenize Design Company in 2011 [12] and has been updated over the years. The index provides a ranking of cities based on their level of bicycle friendliness, aiming at encouraging a growing number of cities around the world to promote a bike friendly lifestyle. Other approaches to measure bikeability are described by Hardinghaus et al. [26], Krenn et al. [40], Eliou and Galanis [15], Zayed [80], Porter et al. [56], Kang et al. [35], Koh and Wong [39] and Winters et al. [77].

Similarities and differences can be highlighted among these approaches considering the four aspects mentioned above. Starting with the type of factors, all indexes consist of a combination of hard factors such as bike infrastructure, hill

score, land use mix, network centrality measures, POIs accessibility, green spaces, bike parking slots and traffic except for the Copenhagenize index and the score computed by Kang et al. [35], which include also soft factors such as cycling policies. Regarding the type of data and information, apart from Kang et al. [35] and Eliou and Galanis [15], who select subjective indicators to measure bikeability, all other approaches consider objective measures of the environment. Subjective indicators represent people's perception towards cycling and the built environment and are measured through semi-structured interviews [35] and surveys [15]. Objective indicators are chosen considering existing research [8, 12, 23, 25, 26, 34, 42, 44, 48, 55, 77, 80], surveys [39, 56, 77], path analysis [40], focus groups [77] and face-to-face interviews [39, 42]. In addition, it is worth highlighting that, since human behaviour is the result of their interpretation of the reality based on attitudes, preferences and perceptions, both objective and subjective measures should be considered in order to fully understand individual and collective behavioural patterns. In this respect, some research focuses on analysing the match and mismatch of those two measures and their relationship with walking and cycling behaviour. The mismatch can explain why in highly walkable and bikeable neighbourhoods not everyone decides to adopt these travelling modes [45]. Once defined, indicators are combined, usually with a weighted summation in order to compute the final bikeability measure. Equal weights can be assigned to all indicators [8, 12, 39, 40, 77] or weights can differ based on the level of importance of each indicator according to theoretical assumptions [23, 42, 44, 48], experts' opinion [26], decision makers and planners' evaluations [25] or statistical models such as PCA and multiple linear regression [34]. Concerning model resolution, the index can measure bikeability at a city level [12, 15, 80], for a limited area surrounding the survey respondents [56] or can provide high resolution estimates [23, 25, 26, 34, 39, 40, 42, 44, 48, 55, 77]. Among the described approaches, only the ones by Kang et al. [35] and Eliou and Galanis [15] can be considered qualitative insofar their data and findings are meant to provide a deep understanding of a specific geographic area, rather than presenting a generalisable approach. The other methodologies could potentially be applied to several different scenarios if data are available. However, a good practice may be to conduct a preliminary investigation of the area of interest in order to check whether the pre-selected criteria adapt well to such a context. Indeed, the level of importance of bikeability indicators may change according to the country and city, but it may also vary from urban to suburban areas. In addition, the extension of the bikeability computational framework to new areas is limited by data availability. For instance, the mathematical approach proposed by Lowry et al. [44] requires a lot of specific measurable indicators that

may not be available for all areas. An example of a generalisable procedure applied to a restricted area due to lack of data is presented by Schmid-Querg et al. [60]. The bikeability index is computed for a part of a district of the inner city of Munich considering existence and type of bike path, speed limit, parking facilities for bicycles, and quality of intersection infrastructure for bicycles. Since OpenStreetMap data were not complete or classified consistently, the authors integrated the dataset with assessed local knowledge and field observations [60]. Precision of the estimate is in this case preferred over generalisation of the findings. Another aspect worth considering when computing the bikeability index is the cycling purpose. Only two out of the seventeen approaches presented above explore it. In particular, Porter et al. [56] and Kang et al. [35] show a significant difference between cycling for transportation and cycling for leisure. Objective measures of the built environment appear to be correlated only to the first purpose, while cycling for recreation is more correlated with subjective perceptions of the environment and personal preferences. New indexes should be built to measure bikeability for recreation considering these findings. Finally, some studies present a validation procedure for the presented bikeability index. Researchers investigate travel behaviour of people in order to assess the predictive validity of their score. The assumption, confirmed by their findings, states that travel behaviour is positively correlated to the bikeability index [35, 40, 56, 77].

2.3. Research focus

Similar to the methodology used to build the Bike Composite Index by Kamel et al. [34], in this work bikeability will be estimated through two sub-dimensions of the concept namely accessibility and safety. This choice is motivated by the fact that bikeability is a rather broad concept and a one-dimensional measure could easily lead to misleading results. For instance, an area that is characterised by comfortable and safe bike infrastructure, parks and low car traffic, but at the same time due to its location does not represent an important link to access relevant destinations, could have a bikeability score which is similar to an area with opposite characteristics: central but unsafe. In fact, in the first case bike infrastructure, parks and low car traffic would increase the bikeability score, while the lack of POIs, population density and low centrality measures would decrease it. The second case represents the opposite scenario. These statements are supported by the literature as studies found that wide and well separated bike lanes, green areas, network betweenness, bike network coverage, population and POIs density increase bikeability, while car traffic, intersection density and traffic lights decrease it [26, 34, 46, 47, 80].

This example demonstrates the need to further investigate the broad bikeability concept in order to detect strong and weak points of each specific area. Thus, creating a bi-dimensional bikeability index would lead to a better understanding and interpretation of reality in order to develop efficient and effective policies. Recalling the categorisation illustrated in the literature review, it is possible to state that the bi-dimensional bikeability index presented in this work is a high resolution grid-based model, which considers hard factors and objective measures of the built environment. Furthermore, the overall approach is quantitative as it is meant to be adapted and applied to new cities and countries. Moreover, this work investigates cycling for transportation purposes, as opposed to cycling for leisure. In particular, accessibility and safety are the two bikeability sub-dimensions to be estimated throughout this work. As already mentioned, accessibility denotes areas characterised by a high bicycle flow, which most probably represent important connections to reach relevant destinations. Safety refers instead to areas where the crash risk for cyclists is limited. Therefore, bicycle flow and bicycle accidents are going to be the proxies to determine accessibility and safety respectively. Based on the literature, bike lane score, hill score, POIs score, parks, intersection density, traffic lights, bicycle parking capacity, car traffic, car speed, betweenness centrality, network coverage and population are selected as independent variables, which are expected to have an impact on accessibility and safety. The data analysis presented below focuses on identifying the most relevant factors among the aforementioned ones to estimate the bikeability sub-dimensions of an area through both descriptive statistics and statistical models. In particular, predictive models allow us to estimate accessibility and safety even when the proxy variables namely bicycle flow and bicycle accidents are missing. Once bikeability sub-dimensions are estimated and their main drivers are identified, it would be possible to detect which areas need to be targeted by policies to promote cycling and make it a safer activity.

Chapter 3

Data

3.1. Research context

This work uses the city of Berlin as a case study to map bikeability and gain a deeper understanding on the phenomenon. On December 31st, 2021, Berlin registered 3,775,480 residents [1] and according to 2020 data the city has a population density of 4,112 inhabitants per square kilometre [7] and a total area of 89,112 hectares [4]. Berlin was almost completely destroyed during the Second World War and re-built as a car-first city with wide streets and lots of car parking facilities. However, the city offers wide sidewalks and streets to accommodate cyclists, lots of parks and green spaces, which cover 27% of the total area [27], forests and lakes in most of its neighbourhoods. As mentioned before, the 2019 Copenhagenize index assigned a score of 56.3% to Berlin, which positions the city to the 15th place of bicycle friendly cities compared to other cities in the world [31]. Recently, several initiatives have been organised by citizens, associations and institutions in order to promote the transition to a walk and bike-centric city. They will be discussed later.

3.2. Data sources

Several types of data and data sources have been used in order to conduct the analysis presented in this work. A relevant work of data retrieval and harmonisation has been done in order to build the final dataset.

OpenStreetMap (OSM) [53] data were retrieved to categorise Berlin's bike network based on the type of bike infrastructure namely bike paths, lanes and shared infrastructure. POIs locations such as schools, grocery stores and transit stations were also extracted from OSM. This choice is motivated by the fact that OpenStreetMap is considered the most successful crowdsourced geographic information project, which offers a free and editable map of the entire world and the possibility to extract a huge variety of free geographic data. However, it has to be said that, while

in some areas data are complete and accurate as it is the case for Germany [81], other areas instead such as developing countries, or some rural areas still miss important information. Concerning Berlin, FIS-Broker [16] provides bike infrastructure data as well. However, when compared, OSM data were more accurate and complete.

Population data come from the Federal Statistical Office (Destatis) of Germany and refer to 2018 [66].

Car traffic data are provided by the tech start-up Targomo based in Berlin. The dataset consists of traffic data extracted from car GPS trajectories, coming from smartphone applications and collected during a week (2019-10-21 – 2019-10-27) in Germany.

High resolution elevation data (20 m grid) were found in the official portal for European data *data.europa.eu* [14].

Bicycle traffic data come from the *MOVEBIS* project. This dataset contains bicycle traffic volume recorded from 2018 to 2020 in Germany by users via the *CITY CYCLING* app. The app records data from smartphone sensors such as GPS and accelerometer during the so-called *CITY CYCLING* campaign organised by the Climate Alliance. The campaign aims at encouraging a growing number of people to travel by bike in their daily life and asks participants to complete as many bicycle trips as possible over a 21-day period. Data from 2019, before the Covid-19 pandemic outbreak, were chosen for this analysis. 1,001,931 trips in Germany completed by 77,049 users, for a total of 7.8 million km were recorded [10, 65, 68]. Bicycle traffic data are also offered by the 26 bike counting stations in Berlin [79]. Even though GPS data from *MOVEBIS* were preferred for this analysis due to their broader space coverage, counting stations data produced useful insights as well. They will be presented in the next section.

Bike accidents data (2019) were downloaded from the city data portal [67], which offers open data about road traffic accidents including several attributes such as type of vehicle(s) involved, day and hour of the day, type of accident and road conditions.

3.3. VARIABLES

First of all, H3 hexagons resolution 9 [70] i.e. with an average edge length of around 174 m were generated in order to cover the entire city of Berlin. H3 [24, 32] is an open source hierarchical geospatial indexing system developed by Uber, which partitions the world into hexagonal cells. In this work, the purpose is to calculate

the accessibility and safety indicators for each cell where bicycles are allowed to travel. In this way, it is possible to generate a high-resolution map of the city bikeability made of equally sized units. Moreover, resolution 9 represents a good trade-off between high resolution and computational complexity. The total number of hexagons covering the city is 9460, however only 7583 units are taken into consideration as areas where bicycles are not allowed are excluded from the analysis. Therefore, the data set is characterised by 7583 spatial units, which represent the total number of observations.

In order to measure bikeability, the following indicators have been chosen and will be described in the next sections: bike lane score, hill score, POIs score, parks, intersection density, traffic lights, bicycle parking capacity, car traffic, car speed, betweenness centrality of the bike network, network coverage and population. Bicycle flow and bicycle accidents are instead used as proxies to estimate accessibility and safety respectively.

3.3.1. Bike lane score

Bike lane score (BLS) is an estimate of how an area is bike friendly in terms of bike infrastructure. The computational approach follows the methodology applied to compute the *Bike Score* provided by the *Walk Score* initiative discussed above [8]. To compute such a score, the total length in metres of bike paths, bike lanes and shared infrastructure was weighted and summed up for each hexagon. In particular, bike paths were considered three times more valuable than shared infrastructure and twice more valuable than bike lanes. In addition, a penalisation was added for those shared infrastructure with a bad surface to cycle according to OSM tags such that half of the total amount in metres of shared infrastructure with a bad surface per hexagon were subtracted from the total amount of shared infrastructure.

```
BLS = bike_paths * 3 + bike_lanes * 2 + (shared_infra -
bad_shared_infra * 1/2)
```

To be more specific, bike paths refer to cycle lanes, which are physically separated from the motorised traffic, for example with a bollard, kerb or trees. This is the best and the safest configuration of cycle lanes, which explains why it is assigned more weight in the bike lane score. Bike lanes are instead cycle infrastructure, which are usually separated from the rest of the traffic flow only by a painted line on the street. Finally, shared infrastructure are roads where bikes have access, but no cycle lanes

are provided. Tables 3.1, 3.2, 3.3 illustrate the specific OSM tags that were used in order to create the streets' categorisation. Concerning bad surface, the following tags were considered: *''surface''=''sett''*, *''surface''=''cobblestone''* and *''surface''=''unhewn_cobblestone''*.

CATEGORY	OSM TAGS
Bike paths	"highway"="cycleway" "cycleway"="track" "cycleway"="opposite_track" "cycleway:right"="track" "cycleway:right"="opposite_track" "cycleway:left"="track" "cycleway:left"="opposite_track" "cycleway:both"="track" "cycleway:both"="track" "sidewalk:left:bicycle"≠"no" + "sidewalk:left:segregated"="yes" "sidewalk:left:bicycle"="yes" "sidewalk:left:bicycle"="designated" "sidewalk:right:bicycle"≠"no" + "sidewalk:right:segregated"="yes" "sidewalk:right:bicycle"="yes" "sidewalk:right:bicycle"="designated" "sidewalk:both:bicycle"≠"no" + "sidewalk:both:segregated"="yes" "sidewalk:both:bicycle"="yes" "sidewalk:both:bicycle"="designated" "highway"="cycleway" + "separation:both"="grass_verge" "highway"="cycleway" + "separation:left"="grass_verge" "highway"="cycleway" + "separation"="kerb" "highway"="cycleway" + "separation:left"="kerb;parking_lane" "highway"="cycleway" + "separation:left"="parking_lane" "highway"="cycleway" + "separation:left"="kerb" "highway"="cycleway" + "separation:left"="bollard" "highway"="cycleway" + "separation:left"="vertical_panel" "highway"="cycleway" + "separation:left"="separation_kerb" "highway"="cycleway" + "separation:left"="planter" "highway"="cycleway" + "separation:left"="railing" "highway"="cycleway" + "separation:left"="guard_rail" "highway"="cycleway" + "separation:left"="structure" "highway"="cycleway" + "separation:left"="tree_row" "highway"="cycleway" + "separation:both"="tree_row" "cycleway:both"="separate" "cycleway:right"="separate" "cycleway:left"="separate" "highway"="track" + "motor_vehicle"="no" "highway"="path" + "bicycle"="designated" "highway"="path" + "bicycle"="yes" "highway"="service" + "bicycle"="yes" + "motor_vehicle"="no" "highway"="service" + "bicycle"="designated" + "motor_vehicle"="no" "highway"="pedestrian" + "bicycle"="yes" "highway"="pedestrian" + "bicycle"="designated" "highway"="footway" + "bicycle"="designated" "highway"="footway" + "bicycle"="yes" "highway"="bridleway" + "bicycle"="yes" "highway"="bridleway" + "bicycle"="designated" "bicycle_road"="yes"

Table 3.1. Bike paths definition

CATEGORY	OSM TAGS
Bike lanes	"cycleway"="lane" "cycleway:right"="lane" "cycleway:left"="lane" "cycleway:both"="lane" "cycleway"="opposite_lane" "cycleway:right"="opposite_lane" "cycleway:left"="opposite_lane" "cycleway:both"="opposite_lane"

Table 3.2. Bike lanes definition

CATEGORY	OSM TAGS
Shared infrastructure	"highway"="primary|primary_link|secondary|secondary_link|tertiary|tertiary_link|residential|road|unclassified|ferry|living_street" + !"cycleway" + !"cycleway:right" + !"cycleway:left" + "access"≠"no|private" + "bicycle"≠"no|private|discouraged" + "motor_vehicle"≠"no"

Table 3.3. Shared infrastructure definition

3.3.2. Hill score

The hill score was computed following the procedure described by Kamel et al. [34]. A subset of LiDAR Digital Terrain Model data was created for the city of Berlin and the Berlin bike network, which is the union of bike paths, bike lanes and shared infrastructure retrieved before was considered. Two algorithms were run in QGIS [75]: `Drape (set Z value from raster)` allows to set the z values (i.e. elevation) of every vertex of each street segment; `Extract Z values` extracts z values from feature geometries into feature attributes. The absolute difference between the z values corresponding to the vertexes of each segment, is a measure of the slope of the segment. Finally, the hill score for each hexagon, is given by the ratio between the aggregate lengths weighted slope and the length of the network in each hexagon:

$$\text{Hill score} = \frac{\sum_{i=1}^{n} l_i \, s_i}{\sum_{i=1}^{n} l_i}$$

where l_i represents the i-th link's length, s_i represents its slope and n represents the number of links in the given hexagon [34].

3.3.3. POIs score

The POIs score is estimated with OpenStreetMap data and Targomo Travel Times

API [71]. Similarly to the categories identified by McNeil [48], the following POIs (Points of Interest) were selected: schools, libraries, transit connections (public transport stops), grocery stores, clothing stores, general goods, beauty services, banks, laundry and cleaners, gyms, drinking establishments, cinemas and theatres, food and drink (restaurants, cafes, bars, pubs, fast food and biergartens) and places of worship. Table 3.4 presents the specific OSM tags used to retrieve the POIs. After data cleaning, to remove duplicate locations, the POIs score was calculated per hexagon as the sum of the weighted distance of POIs reachable in 10 minutes by bike from the centroid of each hexagon. For each category, each POI belonging to the category, was divided by a decay function based on its travel time by bike (set to a maximum of 10 minutes catchment area) from the centroid of the hexagon. The decay function considered was the square root of travel time. In this way, the closer the POI, the higher the score. Then, all POIs' weighted scores were summed up and finally all categories' results were summed up to compute the overall POIs score. Therefore, the POIs score ps of each category of POIs c is computed as:

$$\mathrm{ps}_c = \sum_{i=1}^{n} \frac{1}{\sqrt{t_i}}$$

where t_i is the travel time in seconds from the centroid of the hexagon considered to the i-th POI belonging to category c and n is the number of POIs in category c. The overall POIs score PS for a hexagon is given by:

$$\mathrm{PS} = \sum_{i=1}^{n} \mathrm{ps}_i$$

where n indicates the number of POIs categories.

CATEGORY	OSM TAGS
Schools	"amenity"="childcare" "amenity"="kindergarten" "amenity"="school" "amenity"="university" "amenity"="college"
Libraries	"amenity"="library"
Transit connections	"railway"="tram_stop" "highway"="bus_stop" "amenity"="bus_station" "highway"="platform" "railway"="subway_entrance" "railway"="halt" "railway"="station"
Grocery stores	"shop"="grocery" "shop"="supermarket" "shop"="convenience" "shop"="greengrocer"
Clothing stores	"shop"="clothes" "shop"="fashion" "shop"="boutique" "shop"="shoes"
General goods	"shop"="general" "shop"="variety" "shop"="department_store"
Beauty services	"shop"="beauty" "shop"="hairdresser" "shop"="tanning_salon" "shop"="massage" "shop"="cosmetics"
Banks	"amenity"="bank"
Laundry and cleaners	"shop"="laundry" "amenity"="dry_cleaning"
Gyms	"amenity"="gym" "leisure"="fitness_centre"
Drinking establishments	"amenity"="drinking_water"
Cinemas and theatres	"amenity"="cinema" "amenity"="theatre"
Food and drink	"amenity"="restaurant" "amenity"="fast_food" "amenity"="cafe" "amenity"="biergarten" "amenity"="bar" "amenity"="pub"
Places of worship	"amenity"="place_of_worship" "building"="church" "building"="cathedral" "building"="chapel" "building"="mosque" "building"="synagogue"

Table 3.4. POIs definition

3.3.4. Parks

Parks are considered as a separate category and instead of counting the number of parks within a given distance, the square meters of reachable area, from the centroid of each hexagon to a maximum travel time of 10 minutes by bike was considered. This choice is motivated by the fact that a simple count of parks reachable in 10 minutes could not always be our desired output. A big park like Tiergarten (2.1 km^2) should have more importance than two very little ones. Therefore, considering the square meters of reachable parks, makes it possible to overcome this issue. To be more specific, parks are defined by the following OSM tags: *''leisure''='park''*, *''leisure''='dog_park''* and *''landuse''='recreation_ground''* and the routing is computed using Targomo Travel Times APIs.

3.3.5. Intersection density

Intersection density of the bike network, which is the combination of bike paths, bike lanes and shared infrastructure, was computed in QGIS. First of all, the bike network was dissolved and a unique MultiLineString object was created out of all street segments. The algorithm `Multipart to singleparts` was used to merge all adjacent segments. Bike streets' intersections were computed with `Line intersections`. After having deleted duplicate geometries and created a spatial index to allow the algorithm to run faster, the number of intersection points per hexagons were computed with `Count points in polygon`.

3.3.6. Traffic lights

As mentioned before, traffic lights were retrieved from OpenStreetMap using the *''highway''='traffic_signals''* tag and the total number per hexagon was computed.

3.3.7. Bicycle parking capacity

Similarly to traffic lights, bike parking facilities were retrieved from OSM and the absolute number of bike slots, namely the capacity of each bike parking space per hexagon was considered (*''amenity''='bicycle_parking''*). The median capacity value was assigned to bike parking facilities with missing capacity attributes. Furthermore, a 50 m buffer was created around each parking point or polygon and the bike parking's capacity was proportionally assigned to intersecting hexagons based on the percentage of intersection area. The reason for this step is that parking facilities that are located next to the border of an hexagon are also partially assigned to the very close neighbour.

3.3.8. Car traffic and car speed

Car traffic data, processed and provided by Targomo, consists of the car network of Germany divided into street segments with the daily average car traffic and average speed of cars as attributes for each edge. In order to obtain the average car traffic and speed per hexagon, the mean average traffic and speed of all segments that intersect the hexagon are computed. One the one hand, it must be said that the mean of the traffic is not the optimal aggregation criterion as it underestimates car traffic when busy streets and small roads with few cars belong to the same hexagon. On the other hand, alternative aggregation criteria such as the maximum or the sum would lead to biased estimates as well. For instance, considering the sum, since streets are divided into multiple segments, the same cars are counted several times. Moreover, since street segments have different lengths, the same cars are counted more times in some hexagons and less in other ones. Alternative aggregation approaches could be explored and applied in future research.

3.3.9. Betweenness centrality

The betweenness centrality of a node is the number of shortest paths between pairs of nodes in the graph that pass through that specific node. Betweenness centrality is computed for each bike network's node using the `momepy` Python library [50]. The values for each hexagon are then computed by averaging the values of the nodes inside it.

3.3.10. Network coverage

Network coverage is calculated as described by Kamel et al. [34] as the ratio of the number of bike links, meaning all streets where bikes have access, to the number of all street network links in each hexagon. The whole bike and street network of the city of Berlin are retrieved from OSM via the `osmnx` Python package [54].

3.3.11. Population

Population data coming from the census were processed by Targomo and a H3 grid resolution 10 was generated. Our indicator was built following the same approach as for POIs score. Indeed, the sum of the weighted number of people reachable in 10 minutes from each hexagon's centroid was calculated. As before, weights are defined by the inverse of the square root of the travel time.

3.3.12. Bicycle flow

The *MOVEBIS* project offers bicycle count data coming from GPS trajectories coordinates, represented by the centroid of H3 hexagons in resolution 12. In order to get the bike flow estimate per resolution 9 hexagons, the sum of the bike count within each hexagon and the average bike count were considered as aggregation criteria. Both of them represent biased estimates of the total bike count per hexagon. On the one hand, the sum overestimates the real value as bikes passing through an hexagon are counted multiple times. On the other hand, the average underestimates it. For instance, if a popular street that counts a high number of cyclists is close to a small road with only few cyclists recorded, the final average bike count would be significantly penalised. However, for the purpose of this work, the sum is selected as the preferred aggregation criterion as it allows to assess bicycle flow more precisely. In fact, bicycles in transit are counted multiple times in every hexagon in the same way, while bicycles that stop somewhere in the hexagon are only counted before they stop.

3.3.13. Bicycle accidents

Bike accidents data were filtered from road accidents data provided by the city data portal. Similarly to what was done for bike parking facilities, a 50 m buffer was created around each accident and the score was proportionally assigned to the intersecting hexagons based on the percentage of intersection area. This smooth assignment procedure avoids giving full responsibility to a specific hexagon when the accident happened close to the borders.

Chapter 4

Data analysis

4.1. Data exploration

Before fitting statistical models on the prepared data, it is a good practice to perform exploratory analysis in order to gain a deeper understanding of the data. Figure 4.1 shows Berlin's bike network. In particular, there are around 4082 km of bike paths, 540 km of bike lanes and 6923 km of shared lanes.

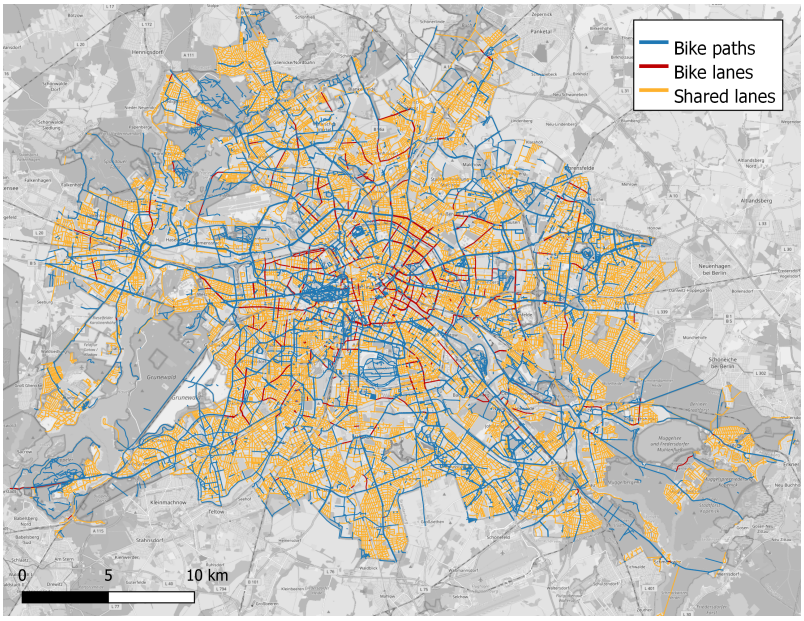

Figure 4.1. Bike network by lanes category

Concerning the variables of interest, from the kernel density plots (Figure 4.2), it is possible to observe that the majority of them namely bike lane score, intersection density, hill score, parks, traffic lights, POIs score, population, betweenness centrality, car traffic, bicycle parking capacity, bicycle flow and bicycle accidents, present right-skewed distributions, meaning that lots of observations have low values, while

there are outliers that display high values. The opposite is true for network coverage, which is characterised by a left-skewed distribution. Average car speed presents two peaks, one of which is at 0 km and represents streets where cars do not have access. Furthermore, a correlation matrix (Figure 4.3) was built in order to display pairwise correlations among variables. It can be seen that there are no relevant negative correlations, while the highest positive correlations can be observed between POIs score and population and between intersection density and bike lane score.

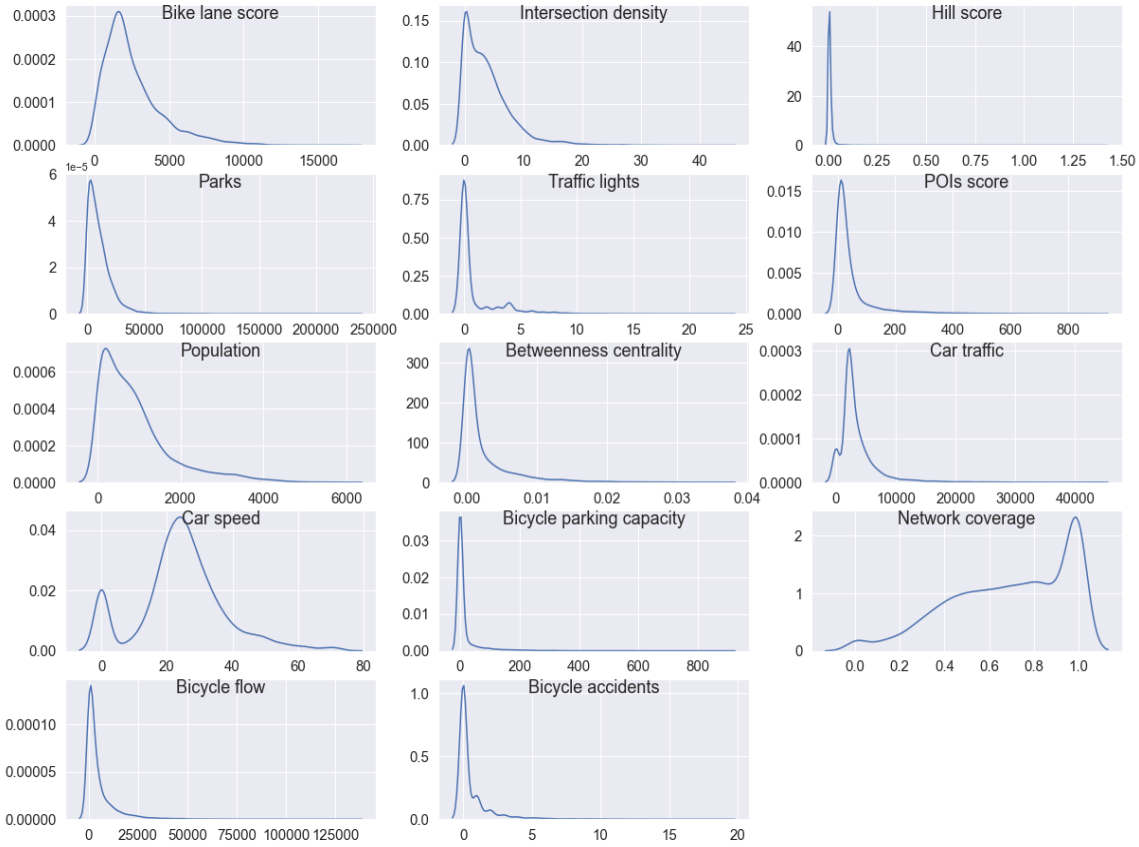

Figure 4.2. Density plots

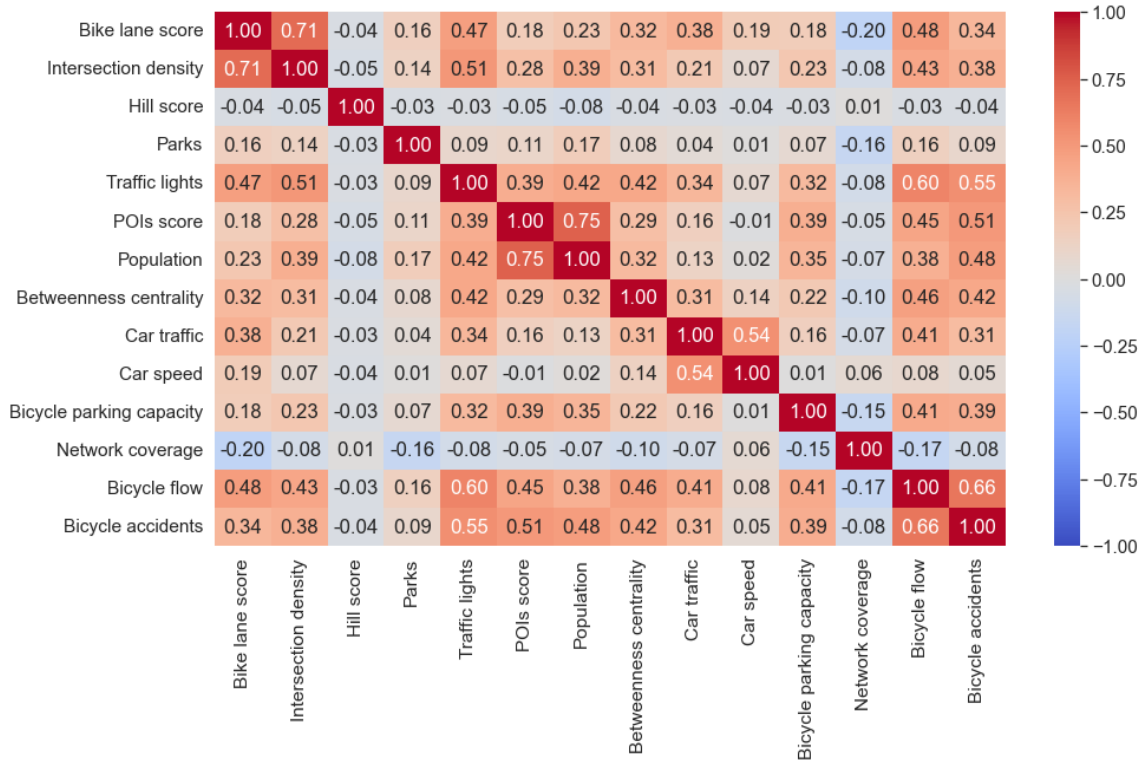

Figure 4.3. Correlation matrix

Concerning bicycle accidents data, together with bicycle accidents count, the dataset also provides some interesting features about each accident's circumstances, which are worth discussing here (Figure 4.4). First of all, out of the 5005 bike accidents registered in 2019, 87% are accidents with minor injuries. Regarding the time period, the number of bike accidents is higher from April to September. This is probably connected to weather conditions and bicycle volume. In other words, warmer and sunnier months might encourage more people to cycle and as a consequence, more bicycles on the streets increase the number of accidents. Moreover, the majority of accidents seem to happen between 8 am and 10 am and between 4 pm and 7 pm. This time period coincides with the usual start and end of the working day when the traffic is usually higher. The victims could be commuters, meaning people that are cycling from home to the working place and vice versa. Working days, from Monday to Friday, are characterised by a higher number of accidents compared to weekends. Regarding accidents' circumstances, the vast majority of them happened with turning and crossing vehicles and more precisely with turning and crossing cars. In addition, the largest number of accidents register a car as the second vehicle involved (3405 cars involved as second vehicle). Finally, the road conditions were mostly dry rather than wet or slippery when the accident happened.

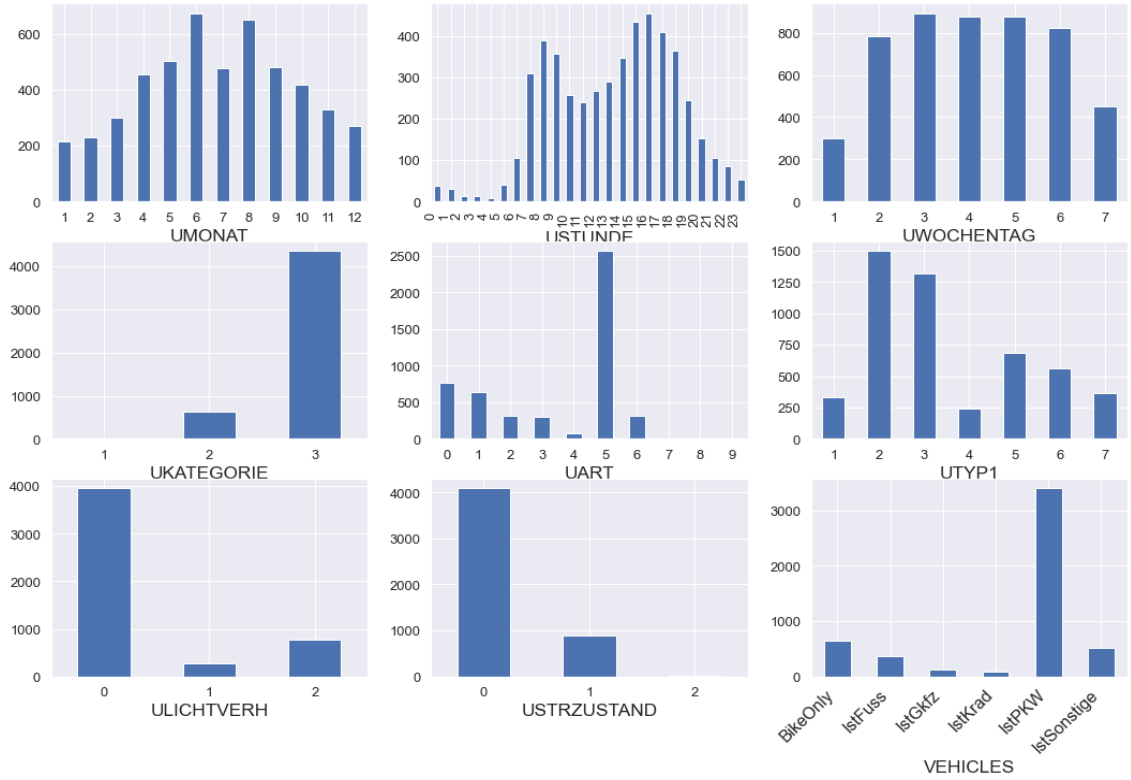

Figure 4.4. Bicycle accidents features

UMONAT = month (1=January, 12=December); USTUNDE = hour; UWOCHENTAG = day of the week (1=Sunday; 7=Saturday); UKATEGORIE = category (1=accident with fatalities, 2=accident with seriously injured people, 3=accident with minor injuries); UART = type of accident (1=collision with starting / stopping / stationary vehicle, 2=collision with one driving ahead / waiting vehicle, 3=collision with laterally in the same direction moving vehicle, 4=collision with oncoming vehicle, 5=collision with turning / crossing vehicle, 6=collision between vehicle and pedestrian, 7=collision with a road obstacle, 8=departure from the lane to the right, 9=departure from the lane to the left, 0=other type of accident); UTYP1 = type of accident (1=driving accident, 2=accident when turning, 3=turning / crossing accident, 4=crossing accident, 5=accident caused by stationary traffic, 6=accident in longitudinal traffic, 7=other accident); ULICHTVERH = light conditions (0=daylight, 1=twilight, 2=darkness); USTRZUSTAND = road conditions (0=dry, 1=wet / damp / slippery, 2=winter smooth); VEHICLES = other vehicles involved in the accident

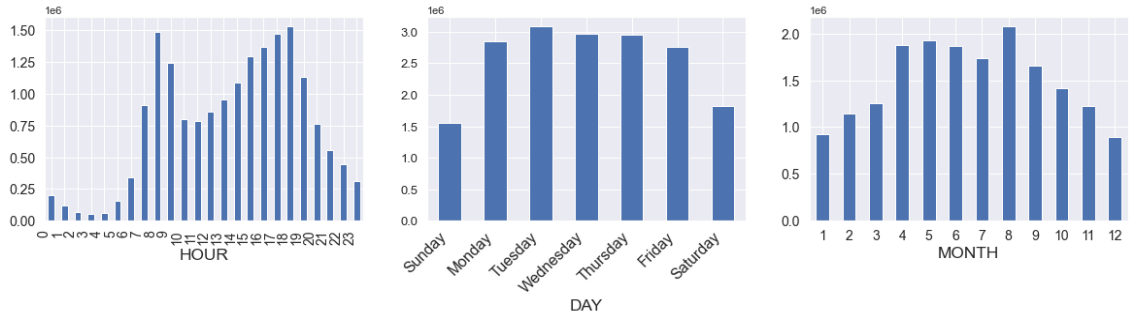

Figure 4.5. Bicycle count over hours, days and month

From these statistics, three main hypotheses arise: bicycle accidents are correlated with bicycle volume, bicycle accidents are correlated with streets' intersections and bicycle accidents are correlated with the presence of cars.

4.1.1. Bicycle accidents are correlated with bicycle volume

As presented above, the data show a higher number of accidents in Spring and Summer and in specific hours of the day that are usually the most crowded. Therefore, a positive correlation is expected between bike accidents and bicycle volume. Since the data provided by *MOVEBIS* do not have a time-stamp, data coming from 26 bike counting stations in Berlin are used. Only 2019 data are included in the analysis and the total number of bikes per hour of the day, day of the week and month is computed. The bar plots (Figure 4.5) show the same trend observed for bike accidents, which means a higher bicycle flow was registered from 8 am to 10 am and from 4 pm to 7 pm, during weekdays and in Spring and Summer. In addition, a strong and positive Pearson's correlation was found between bike accidents and bike count considering hour ($r = 0.96$), day ($r = 0.99$) and month ($r = 0.92$). Furthermore, a positive Pearson's correlation of 0.66 can be seen between bike flow data coming from *MOVEBIS* and bike accidents per hexagon. This result further confirms the initial hypothesis, meaning bicycle accidents and bicycle volume are correlated.

4.1.2. Bicycle accidents are correlated with streets' intersections

The bar plots described before, showed that the vast majority of accidents happened with turning and crossing vehicles. Therefore, a positive correlation is expected between bike accidents and streets' intersections and traffic lights. Indeed, a Pearson's correlation of 0.38 is found between accidents and number of intersections, while $r = 0.55$ was obtained between accidents and traffic lights. Traffic lights may have a higher positive correlation with bike accidents than the total number of intersections as they are usually placed in big and dangerous crossings, where accidents are also more likely.

4.1.3. Bicycle accidents are correlated with car traffic

As discussed before, the vast majority of bike accidents involved a car as well.

Therefore, a positive correlation is expected between car traffic and bike accidents. In addition, crashes with cars should happen when cycle lanes are not well separated or not separated at all from the motorised traffic. Therefore, a negative correlation is expected between bike lane score (the higher, the better the cycle lane) and bike accidents. Bike accidents and car traffic show a moderate and positive Pearson's correlation ($r = 0.31$), while average car speed is not correlated ($r = 0.05$). Unexpectedly, bike accidents are positively correlated with bike lane score ($r = 0.34$), meaning that places with more and better cycle infrastructure are also characterised by more bike accidents. A possible explanation could be that, as the majority of accidents happened close to crossings, the bike lanes' quality did not really play a role in these cases. If on the one hand it is possible to find cycle lanes at crossings, on the other hand they cannot be physically separated from the rest of the traffic flow. Moreover, even if a cycle lane is present at crossing places, the latter are still dangerous. An alternative explanation could be that better cycle infrastructure attracts more cyclists and as a consequence bike accidents become more likely.

4.2. Spatial autocorrelation

In 1970 Waldo Tobler promulgated the so-called first law of geography, "everything is related to everything else, but near things are more related than distant things" [49]. This law represents the core of spatial analysis and modelling. Indeed, spatial analysts recognise that every location has a degree of uniqueness due to its situation with respect to the rest of the spatial system [49]. This is the reason why, when analysing geospatial data, it is important to choose models and techniques that consider their spatial structure. Therefore, in this work global and local spatial autocorrelation analysis are performed before fitting statistical models on the data.

First of all, a spatial weights matrix needs to be defined. The spatial weights matrix is a representation of the spatial structure of the data and can be built in different ways according to the chosen notion of proximity and the type of spatial relations among observations. There exist approaches to compute spatial weights according to adjacency relations or distance-based relations. Distance bands weights is the approach chosen here. It consists of choosing a distance threshold up to which weights based on inverse distance are computed, while for all other units weights are set to 0 [64]. A threshold of 3000 m, which is about 10 minutes travel time by bike, was chosen. The spatial weights matrix allows us to compute the spatial lag

for our independent variables namely bicycle flow and bicycle accidents. The spatial lag of a variable is basically a weighted sum of the values observed at neighbouring locations, where the neighbouring relation is defined by the spatial weights matrix. Figure 4.6 shows the spatial distribution of bike flow and bike accidents together with their spatial lag. From the choropleth map, both variables seem to be spatially autocorrelated with high values concentrated in the city centre and lower values close to the borders.

In order to detect whether the global spatial autocorrelation is statistically significant, Moran's I index is implemented. The index ranges from -1 indicating that the variable of interest is perfectly dispersed to 1 that means perfect autocorrelation. Both bicycle flow and bicycle accidents are spatially autocorrelated as their Moran's I values are 0.47 and 0.34 respectively. Moreover, both p-values are lower than 0.01, which means that the global autocorrelation is statistically significant [21].

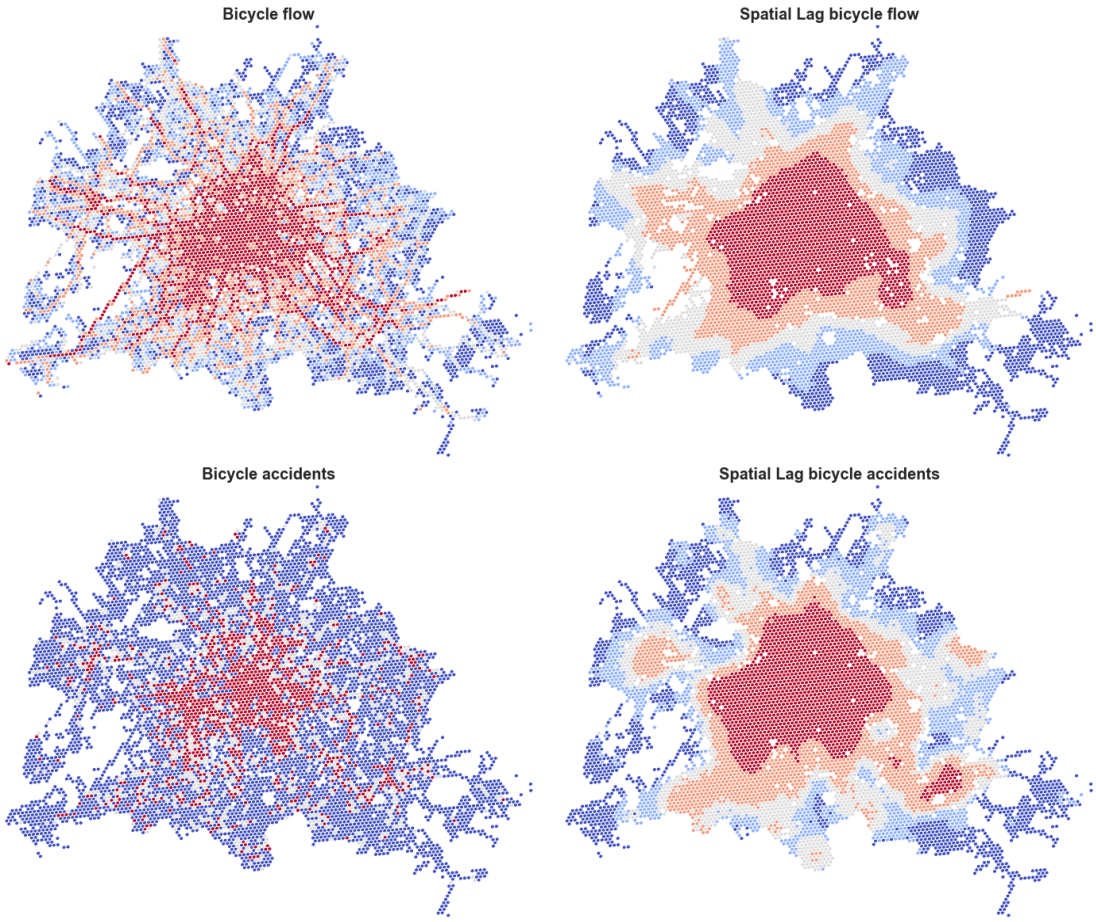

Figure 4.6. Spatial distribution of bicycle accidents and bicycle flow and their spatial lag

Concerning local spatial autocorrelation, local Moran's I index is computed. The index allows us to assess whether each location represents a statistically significant cluster by comparing the results with the expected scenario in case of randomly allocated data. Unlike the global index, local Moran's I outputs as many values as the observations we have. The scatterplots depicted in Figure 4.7, show in red areas characterised by units with high values of bike flow or bike accidents close to units with high values, blue areas represent units with low values close to units with low values, the rest are units with low values close to units with high ones and the other way around. From the scatterplots it can be seen that the majority of units are red and blue, which means affected by local spatial autocorrelation. However, we need to look at the Moran cluster maps in order to assess whether the observed local autocorrelation is statistically significant. Concerning bicycle flow, only 21% of hexagons show a non-significant correlation, while the value increases to 25% considering bicycle accidents. We can conclude that bicycle flow and bicycle accidents are affected by both global and local spatial autocorrelation [43].

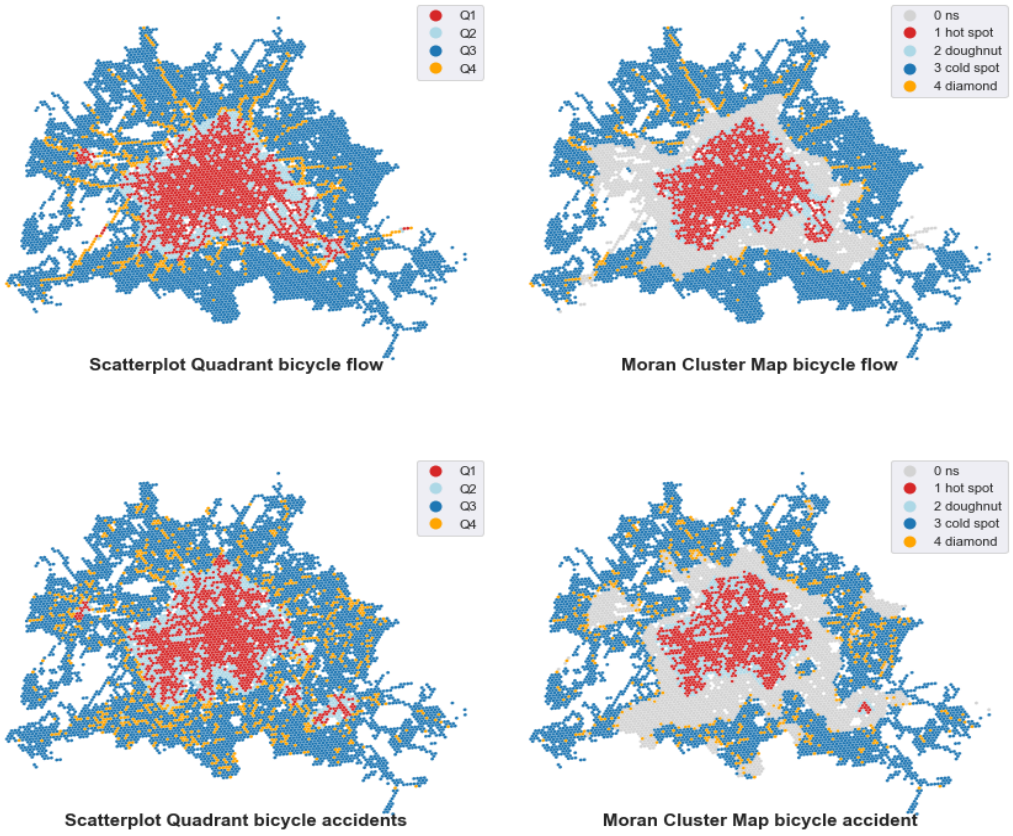

Figure 4.7. Local spatial autocorrelation of bicycle flow and bicycle accidents

4.3. Predictive analysis

4.3.1. Models

Statistical models to predict bicycle flow and accidents can be implemented and then used to estimate the sub-dimensions of bikeability, namely accessibility and safety when bike flow and bike accidents data are not available. Indeed, especially high resolution bike volume data are rarely available and free to use, while data to compute all other independent variables previously described are usually easy to retrieve from OSM or public datasets. In this section, the statistical models applied to predict bicycle flow and bicycle accidents are illustrated. As already mentioned, the independent variables or predictors considered are: bike lane score, intersection density, hill score, parks, traffic lights, POIs score, population, betweenness centrality, car traffic, car speed, bicycle parking capacity and network coverage. As they have different units of measure, they were normalised in order to have mean 0 and standard deviation equal to 1. Moreover, to properly assess the performance of each prediction model, data were randomly split into train (67%) and test set (33%).

4.3.1.1. Bicycle flow prediction

To predict bicycle flow and identify its most influential indicators, three regression models have been tested namely LASSO, ridge regression and Geographically Weighted Regression (GWR).

LASSO is one of the most used algorithms for feature selection and regularisation insofar it gives a penalty to the model for having too many predictors and shrinks some of the least contributive variables towards 0. 10-fold cross-validation was performed in order to select the best model and as a consequence the best lambda parameter, meaning the one that minimises the coefficient of determination R^2. When predicting bike flow, a lambda of 20.19 was chosen as best penalisation parameter and the only coefficient set to 0 resulted in hill score. A RMSE of 6014.44 and R^2 of 0.544 were obtained in the test set.

Ridge regression is a popular regularisation technique, which similarly to LASSO, penalises the magnitude of less contributive predictors in order to reduce overfitting. However, unlike LASSO, ridge regression does not set coefficients to 0, but rather to values close to it. For this analysis, the best model and best lambda were selected once again by 10-fold cross-validation. When predicting bike flow,

a lambda of 10 was chosen and hill score resulted in the least contributive coefficient. A RMSE equal to 6014.35 and R^2 equal to 0.544 were obtained in the test set.

According to both LASSO and ridge regression, bike lane score, intersection density, parks, traffic lights, POIs score, car traffic, betweenness centrality and bicycle parking capacity have a positive impact on bicycle flow, meaning that they contribute to increase it. Car speed, population and network coverage have instead a negative impact.

As already discussed in the exploratory data analysis section, our dependent variables are spatially autocorrelated and as such they violate the independence assumption of OLS regression models. Therefore, a spatial model is built in order to increase the performance. In particular, Geographically Weighted Regression (GWR) is performed. This model fits a local linear regression equation for each observation in the dataset. GWR considers the dependent and independent variables of neighbouring features in order to build these separate equations. Neighbouring relationships can be defined according to different criteria as explained in the exploratory analysis. In this case, the Python function `Sel_BW()` from the `mgwr` Python package [51] searches the best bandwidth for the kernel given train data and coordinates. Adaptive bandwidth was selected as the performance on the test set was higher than the one obtained with a fixed bandwidth. The best value resulted in 136 m. In particular, the variable "hill score" was excluded from the model as LASSO considered it a non-significant predictor for bicycle flow. A RMSE of 4623.54 was obtained on the test set. Since GWR fits a linear regression equation for each observation, it makes it possible to visualise the local R^2 (Figure 4.9). The average R^2 value was 0.865. Figure 4.8 shows a summary of the model's results. In addition, a Moran's I test [41] was performed in order to check whether the model residuals are autocorrelated. As required by the test, a row-standardised spatial weights matrix was built based on the critical cut-off distance criterion. Consistently to the threshold chosen before to test the independent variables' spatial autocorrelation, a cut-off distance of 3000 m was chosen. The Moran's I test displayed a value of -0.009581358 and a p-value below 0.01, meaning that the spatial autocorrelation among the residuals is statistically significant but extremely low ($I = 0$ indicates no autocorrelation). Indeed, the expected value under the null hypothesis of no spatial autocorrelation is equal to -0.000196889. Future research could consider an extension of the GWR model that takes into account the spatial dependence in the residuals in a more proper way. This could improve the parameters' estimates and the model's predictions.

```
Geographically Weighted Regression (GWR) Results
------------------------------------------------------------------------
Spatial kernel:                                     Adaptive bisquare
Bandwidth used:                                               136.000

Diagnostic information
------------------------------------------------------------------------
Residual sum of squares:                                62548937362.243
Effective number of parameters (trace(S)):                      948.932
Degree of freedom (n - trace(S)):                              4131.068
Sigma estimate:                                                3891.157
Log-likelihood:                                              -48676.625
AIC:                                                          99253.114
AICc:                                                         99690.656
BIC:                                                         105459.082
R2:                                                               0.865
Adjusted R2:                                                      0.834
Adj. alpha (95%):                                                 0.001
Adj. critical t value (95%):                                      3.420

Summary Statistics For GWR Parameter Estimates
------------------------------------------------------------------------
Variable                  Mean        STD        Min     Median        Max
-------------------- ---------- ---------- ---------- ---------- ----------
X0                    4764.964   3769.575 -15912.346   4014.774  36885.217
X1                    1855.605   1822.620  -3126.007   1255.507  10554.663
X2                    -152.382   1080.348  -7949.728   -148.498   5283.581
X3                     477.251   2040.195  -5826.621    148.826  13622.421
X4                    1646.907   2053.754 -43956.557   1283.588  76540.605
X5                     746.453   2434.531 -14522.863    565.742  17973.324
X6                    -697.665   1895.804  -9518.393   -385.197   8202.403
X7                    1100.575   1003.398  -1772.096    927.810   7359.992
X8                     930.794   1861.638  -4465.494    541.019   9126.176
X9                    -355.599   1450.246 -14502.259    -75.914   3985.116
X10                    322.595   1342.826  -9659.891    328.923  13322.777
X11                     37.784    799.433  -2242.764    -32.452   6307.873
========================================================================
```

Figure 4.8. GWR results

x0 = intercept; x1 = bike lane score; x2= intersection density; x3 = parks; x4 = traffic lights; x5 = POIs score; x6 = population; x7 = betweenness centrality; x8 = car traffic; x9 = car speed; x10 = bicycle parking capacity; x11 = network coverage

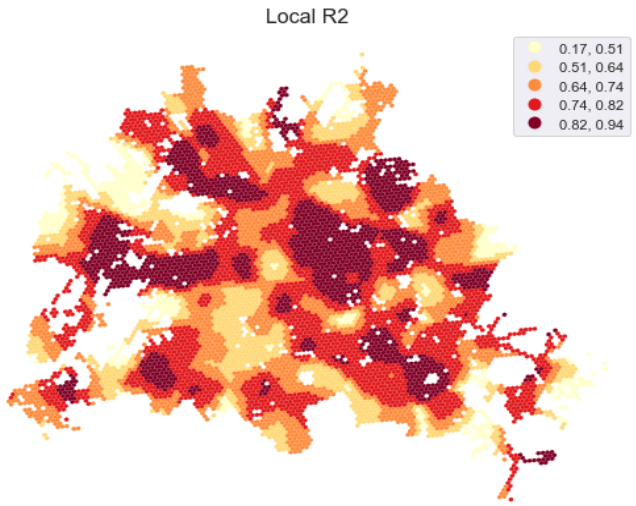

Figure 4.9. Local R^2 of GWR used to predict bicycle flow (all data)

Machine Learning Algorithm	RMSE	R2
LASSO	6014.40	0.544
Ridge Regression	6014.35	0.544
GWR	4623.54	0.865

Table 4.1. Bicycle flow prediction results

Table 4.1 presents the results obtained with the algorithms discussed above.

4.3.1.2. Bicycle accidents classification

As described before, bicycle accidents was a continuous variable, which represented the amount of accidents within each spatial unit. However, as shown in the descriptive analysis, bike accidents are positively correlated with bike volume. Thus, in order to get rid of the noise that this correlation would add to the model, bicycle accidents (`acc_share`) was converted into a binary variable accounting for the presence (`1`) or absence (`0`) of accidents within each unit. The new binary variable's distribution can be seen in Figure 4.10. In order to identify the main factors responsible for bike accidents and to make predictions, three classification models were performed namely LASSO, Decision Trees and Random Forest.

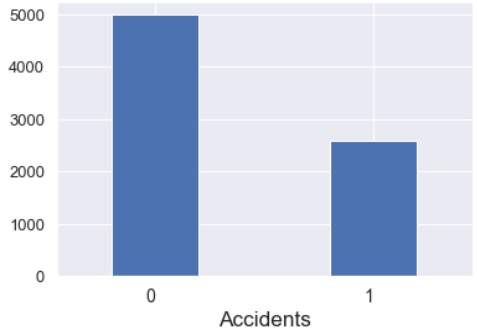

Figure 4.10. Distribution of accidents

A LASSO logistic regression model was fitted on the train set with the best lambda chosen by 10-fold cross-validation and equal to 0.0016. An accuracy score of 0.798 was obtained in the test set. Moreover, hill score and network coverage resulted in having a negative impact on bicycle accidents, while all other predictors positively affect the presence of accidents except for parks, whose coefficient was set to 0 by the model. Thus, parks will not be used as a predictor in the following statistical models.

Decision Trees are supervised-learning methods, which allow to predict a target variable by learning decision rules based on the input features. Moreover, these models are easy to visualise and interpret and for this reason they were chosen to predict the absence or presence of bicycle accidents. Hyperparameter tuning was performed in order to choose the best criterion between entropy and gini to measure the quality of the splits and the best tree's depth to avoid overfitting. The parameters `criterion=entropy` and `max_depth=4` were used to fit the model as they displayed the highest accuracy in the test set (Figure 4.11, on the left). The tree reached an accuracy score equal to 0.787.

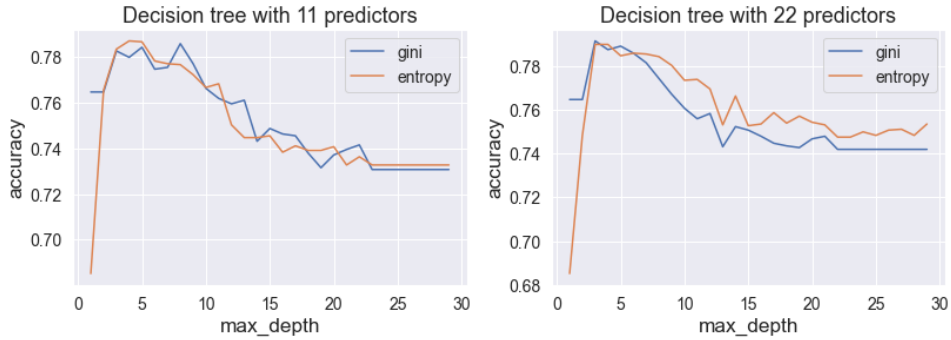

Figure 4.11. Decision trees: hyperparameters tuning

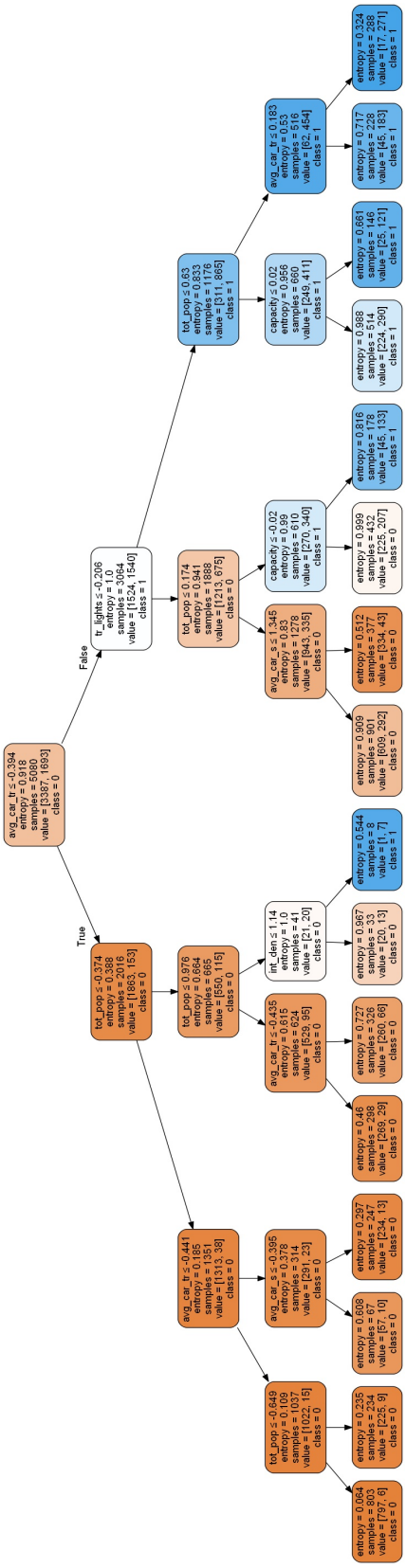

Figure 4.12. Decision tree with 11 predictors

1 indicates accidents = yes; 0 indicates accidents = no

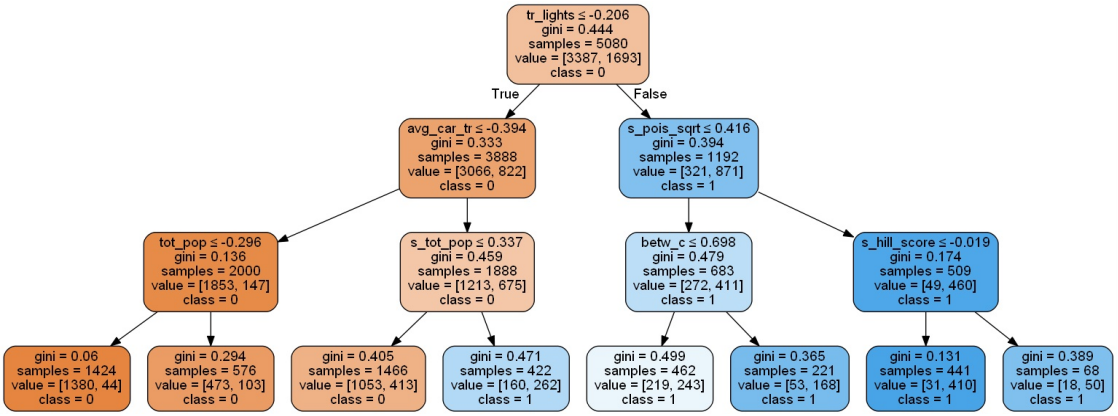

Figure 4.13. Decision tree with 22 predictors

1 indicates accidents = yes; 0 indicates accidents = no

As discussed above, in order to increase the model's performance it would be a good practice to consider the spatial structure of the data. An easy method to include it in the decision tree model is to add as many features as the one we already have in the dataset with the average of neighbouring values. To do so, a distance threshold of 3 km, which was the same criterion used to build the POIs score, parks and population variables and to compute the spatial weights matrix in the exploratory analysis, was chosen. Then, for each spatial unit and for each predictor, the average values of all units within 3 km from the centroid of the unit considered, were used to compute the new variables. As before, hyperparameter tuning was performed and `criterion=gini` and `max_depth=3` were chosen (Figure 4.11 on the right). An accuracy of 0.791 was obtained in the test set. Figures 4.12 and 4.13 illustrate the trees with 11 and 22 predictors respectively.

On the one hand, trees are easy to visualise and interpret, on the other hand they are also known to be less accurate than other traditional statistical models and have high variance. This motivates the choice to perform Random Forest. The Random Forest algorithm grows multiple trees during training and when predicting the target class, it chooses the class with the highest number of "votes". `RandomizedSearchCV()` algorithm from the `sklearn` library [61] was used to perform hyperparameter tuning. By 3-fold cross-validation, the following parameters were chosen considering 11 predictors: `n_estimators=311`, `min_samples_split=2`, `min_samples_leaf=4`, `max_features='sqrt'`, `max_depth=10` and `bootstrap=True`. They indicate the number of trees to be considered by the algorithm, the minimum number of samples required to split a node, the minimum number of samples required at each leaf node, the number of features to consider at every split, the maximum depth of the tree and

the bootstrap method to select samples for training each tree. The accuracy score resulted in 0.805. Moreover, Random Forest also provides estimates of the importance of each feature to predict the target variable. Results are shown in Figure 4.14. Regarding the model that for each predictor also include the average values within a 3 km threshold, the Randomized Search algorithm selected the following best parameters: `n_estimators=1366`, `min_samples_split=2`, `min_samples_leaf=4`, `max_features='sqrt'`, `max_depth=80` and `bootstrap=False`. The model's accuracy resulted in 0.812.

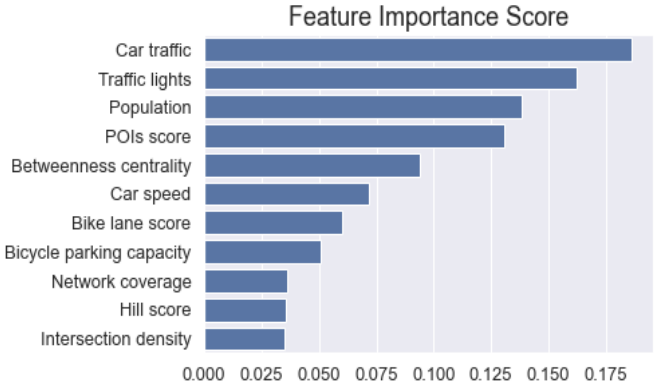

Figure 4.14. Feature importance in Random Forest (11 predictors)

Machine Learning Algorithm	Accuracy
LASSO	0.798
Decision Tree (11 predictors)	0.787
Decision Tree (22 predictors)	0.791
Random Forest (11 predictors)	0.805
Random Forest (22 predictors)	0.812

Table 4.2. Bicycle accidents prediction results

Table 4.2 presents the results obtained with the algorithms discussed above.

4.3.2. Guidances

Considering the regression and classification models' predictions, some insights and suggestions to guide policy makers and urban planners can be presented. The ideal scenario would be improving the quality of bicycle infrastructures and streets everywhere. However, addressing areas that have higher priority would be a good starting

point. Considering bicycle flow and bicycle accidents data, areas with a higher priority are defined as those units that are characterised by a bike flow above the mean and where the proportion of accidents is higher than a certain threshold. First of all, since bike accidents data refer to the entire year and bike flow regards only the month of June, the yearly bike flow was computed. In the exploratory analysis, it can be seen that June is one of the months with the highest count of bicycles according to the 26 bike stations in Berlin. Therefore, multiplying the bike flow data by 12 months would result in an overestimation of the yearly bike flow. To overcome this issue, since bike count data coming from the counting stations include a time-stamp variable, the proportion of bike count was computed for each month. June resulted in 10% of bikes. Therefore, to obtain a more precise yearly estimate, bike flow data were multiplied by 10. As illustrated before, accidents are correlated with bicycle count as more bikes inevitably increase the risk of accidents. Therefore, in order to obtain a standardised estimate of accidents, their amount was divided by the yearly bike flow. Since the amount of accidents is extremely low compared to the yearly bicycle flow, the resulting distribution consists of estimates that are extremely close to 0. Finally, units with a proportion of accidents that is higher than the median of the proportion of accidents' distribution excluding units with value equal to 0, were selected. To summarise, the units that need to be addressed first are here defined as the ones that have a bike flow higher than 5246.98 and a proportion of accidents greater than 0.0000209. Now, it is interesting to check whether the data predicted by the regression and classification models give us similar results. Concerning accidents, since the prediction outcome is binary, units with predicted class equal to *1*, that means presence of accidents, will be considered in combination with predicted bicycle flow above the mean, which was computed only considering train data. In particular, predicted bike flow comes from the GWR model, while predicted presence of accidents from the Random Forest model. Figure 4.15 presents a visualisation of the results.

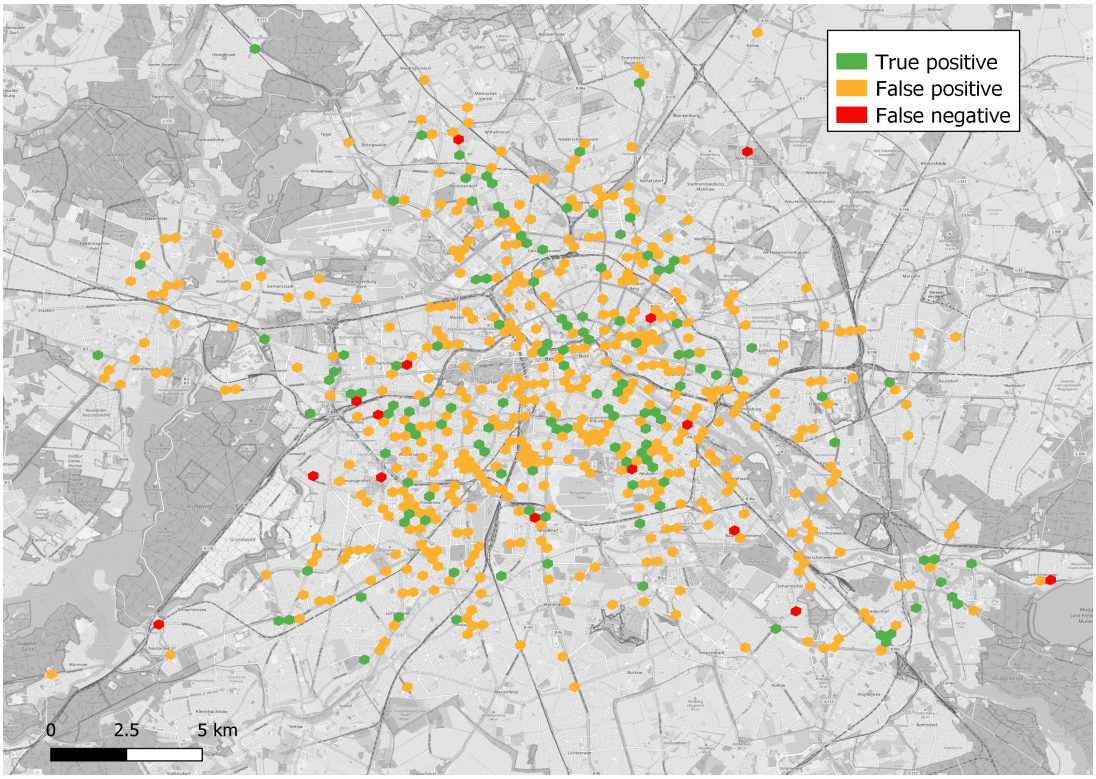

Figure 4.15. Actual and predicted units with high priority in terms of infrastructural improvements

True positive: units selected by both available data and models' predictions. False positive: units selected only by models' predictions. False negative: units selected only by available data

For simplicity, hexagons selected using available bicycle flow and accidents data will be called "true hexagons", while the ones selected using prediction models will be referred to as "predicted hexagons". According to the results obtained, true hexagons are 143, while predicted ones are 588 in total. A higher number of predicted hexagons was expected. In fact, unlike true hexagons, when selecting predicted ones, no threshold for accidents was considered. Indeed, all units with predicted presence of accidents were selected. Furthermore, data show that almost all (90%) true hexagons were detected by the models as well. Moreover, among the predicted hexagons, around 78% are false positive, 22% are true positive and 2% are false negative. In particular, precision, that is the number of true positive divided by the total number of elements labelled as positive, is equal to 22%, while recall, which represents the number of true positive divided by the total number of elements that actually belong to the positive class, is equal to 90%.

Chapter 5

Discussion

5.1. Discussion of the results

5.1.1. Models

5.1.1.1. Bicycle flow prediction

Regarding bicycle flow, LASSO and ridge regression give very similar results in terms of performance and variable selection. According to both models, the hill score resulted in a non-significant factor to predict bicycle flow. This is an expected outcome for Berlin, as the city is characterised by a rather flat terrain Figure 5.1.

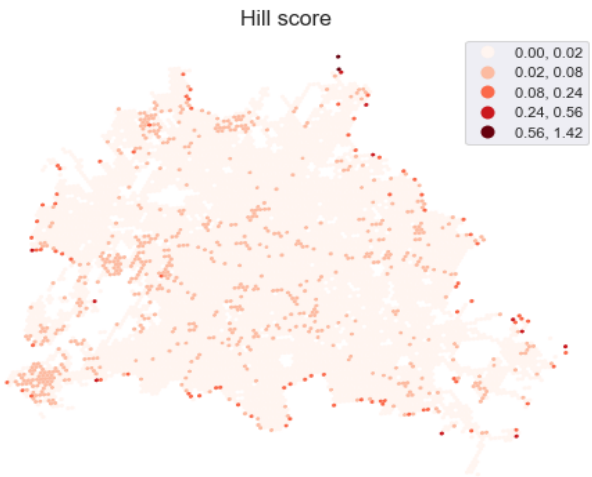

Figure 5.1. Spatial distribution of hill score

On the one hand, the traffic lights variable is given the highest positive coefficient, followed by car traffic, POIs score, bike lane score, betweenness centrality, bicycle parking capacity, parks and intersection density. On the other hand, car speed, population and network coverage negatively affect bicycle flow. First of all, these results show that areas with a high bike flow are also characterised by more

traffic lights, which could imply big and dangerous intersections. Secondly, a high car traffic with cars travelling at a low speed is observed. This could be related to the evidence that, as Figure 4.6 shows, a higher bicycle flow is detected in the city centre, meaning more crowded areas in terms of pedestrians, bicycles, motorised vehicles and POIs, where car speed is limited due to the traffic. Thirdly, more POIs in areas characterised by a higher bike flow is in line with the previous explanation. Moreover, a higher bike lane score and more bike parking slots are observed in the most popular areas for bikes. Concerning population density, a negative coefficient may be due to the fact that a greater volume of bikes is observed in areas other than residential. However, this is an hypothesis that should be verified through the data.

In terms of performance, GWR significantly outperforms LASSO and ridge regression. The main reason is that this model takes into account the spatial structure of the data and as a consequence it increases the R^2 and reduces the RMSE. Concerning positive and negative influence of the predictors on the target variable, the coefficients' average is in line with LASSO and ridge regression results.

To conclude, it is possible to state that according to these analyses, the most relevant drivers of bicycle flow and thus accessibility are traffic lights, car traffic and POIs score.

5.1.1.2. Bicycle accidents classification

According to LASSO logistic regression, parks resulted in a non-significant factor when predicting bicycle accidents. This can be motivated by the fact that parks are distributed quite homogeneously throughout the city as can be seen in Figure 5.2.

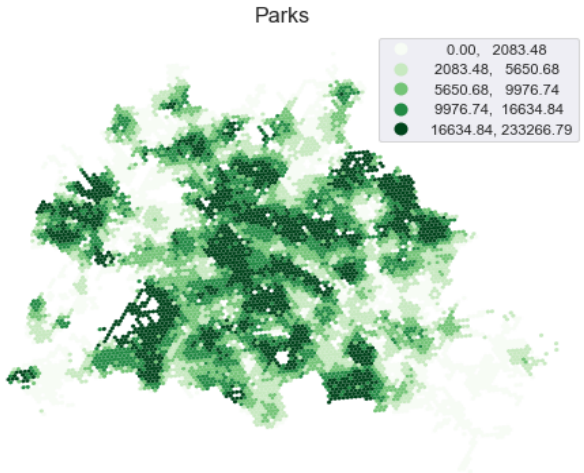

Figure 5.2. Spatial distribution of parks (quantiles)

All the other predictors have a positive influence on accidents except for hill score and network coverage, which negatively affect the target variable. However, as discussed before, hill score does not play an important role in Berlin due to the flat surface of the city. Surprisingly, bike lane score is positively associated with the presence of accidents. A plausible interpretation could be that as already discussed, the great majority of accidents happen at crossings and intersections. In these zones, protected bike lanes lose their benefits and cyclists are in close contact with motorised vehicles. In other words, if the accident happens at an intersection, it does not make a difference if right after or before there are well separated bike lanes.

Random Forest is the model that obtained the highest accuracy score when predicting the target variable. According to LASSO, Decision Trees and Random Forest, the main drivers of bicycle accidents and hence safety are car traffic, traffic lights, population density and POIs. Except for population density, these are the same factors that have the highest influence on bike flow. Therefore, it is possible to state that accessibility and safety are related. Indeed, bicycle flow and bicycle accidents show a Pearson's correlation of $r = 0.66$. These analyses demonstrate that areas characterised by high bicycle flow are also characterised by a higher number of cars, traffic lights and POIs, which are the same factors that contribute to increased accidents and hence make streets less safe.

5.1.2. Guidances

The purpose of the analysis was to detect those spatial units that need to be targeted first by urban planning measures as they are characterised by a high volume of bikes and accidents. Hence, these areas are characterised by high accessibility and low safety. Looking at the results presented in Figure 4.15, it can be seen that false positive units, meaning hexagons selected only by the models, dominate the scene. As discussed before, this result can find a partial explanation in the fact that no threshold was established on predicted accidents. Having a high number of false positive units is preferred rather than observing a high number of false negative ones. In other words, we are more interested in maximising recall (recall = 90%), which indicates how many of the truly positive hexagons are retrieved as positive, rather than maximising precision (precision = 22%), which measures how many of the hexagons labelled as positive are correct. Indeed, we would like the models to detect all most accessible and most unsafe areas to be checked in the first place and targeted by policies in the second place. However, it must be said that a too high number of false positive units can become problematic as well. Indeed, time and

resources would be spent on checking areas that do not have high priority. Additional criteria can be established in order to reduce the number of units to be checked. For instance, if new measures aiming at improving the quantity and quality of bike infrastructures are going to be planned and implemented, it would be of interest to check those units that have high accessibility, low safety and lack bike lanes. Among the so-called "predicted hexagons" ($n = 588$), the ones in which the sum of bike paths and bike lanes is lower than shared lanes can be highlighted ($n = 140$). In other words, these units are characterised by high accessibility, low safety and have more streets where bikes and cars share the same lane rather than bike lanes.

5.2. Policies and urban planning measures

The analyses presented in this work have the potential to suggest urban planning measures and policies to be implemented in order to promote cycling and make it a safer activity. First of all, the descriptive analysis underlines that the majority of accidents happen at crossings and intersections and involve mostly cars as a second vehicle. The prediction models confirm these findings as car traffic and traffic lights are considered two of the most relevant factors responsible for the presence of accidents. Taking this into consideration, in order to increase safety for cyclists, the quality of intersections should be improved and car traffic should be reduced. Regarding intersection quality, new urban design measures need to be implemented to build the so-called protected intersections. The urban planner and designer Nick Falbo presents with a video four measures inspired from dutch intersections: corner refuge islands, forward stop bars for cyclists, setback bicycle crossings and bicycle friendly signals (Figure 5.3). Corner refuge islands physically separate cyclists who turn right as well as cyclists who are waiting at a red signal from the rest of the traffic flow. Forward stop bars allow cyclists to stop at a waiting area further ahead in the intersection with respect to motorised vehicles when crossing a street. In this way cyclists acquire visibility and are faster in crossing the street when the light turns green. Setback crossings provide space and time for motorised and non-motorised vehicles to avoid potential conflicts. Finally, bicycle friendly signals allow cyclists to understand how and when they can proceed without any collision risks with other vehicles [57]. Moreover, other design strategies to improve intersections' quality could be building roundabouts and protected cycling trails around them and where possible, underpasses and overpasses to separate cyclists and pedestrians from motorised vehicles.

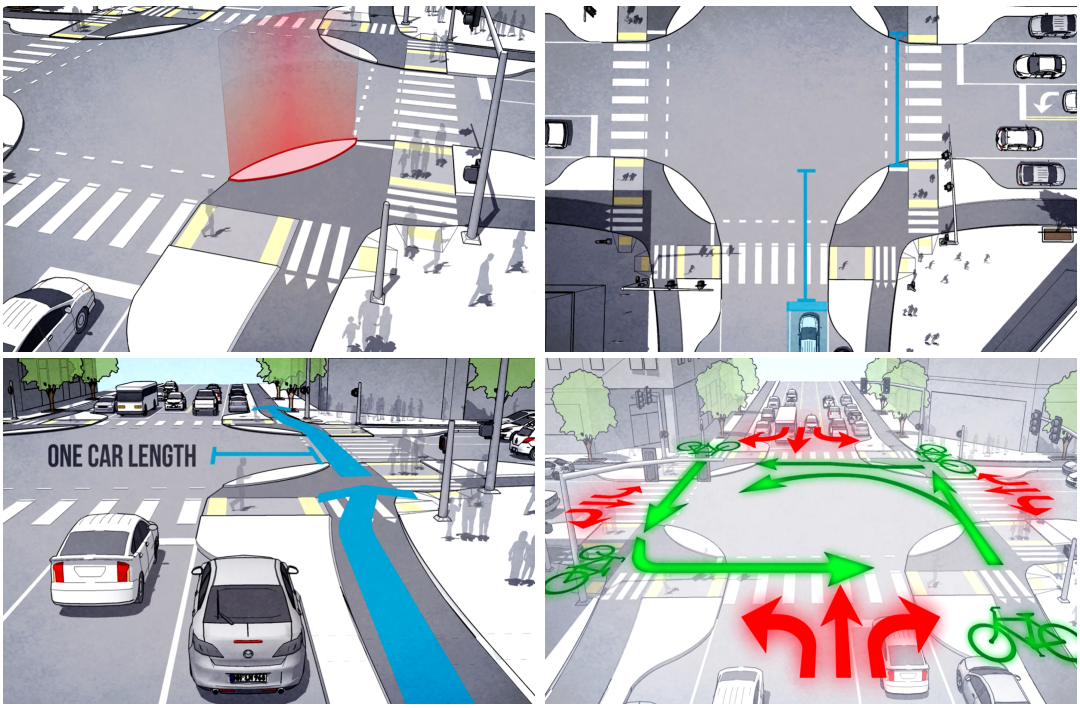

Figure 5.3. Urban design measures suggested by Nick Falbo to build protected intersections

Corner refuge islands (top-left), forward stop bars for cyclists (top-right), setback bicycle crossings (bottom-left) and bicycle friendly signals (bottom-right). The images come from Nick Falbo's video.

Concerning car traffic reduction, many initiatives proposed by institutions and citizens have been planned and are ready to be implemented in the city of Berlin. In April 2016, a group of citizens called *Changing Cities* presented Berlin's and Germany's first draft cycling law [74]. Since the end of May, they started collecting signatures and gaining support from several environment and transportation associations such as Greenpeace. 105,425 signatures were collected in only three weeks and a half, while 20,000 signatures within six months would have been enough to propose the plan to Berlin's Parliament. A referendum did not take place as the initiative was accepted. As a consequence, the Parliament approved the so-called *Berlin Mobility Act*, which came into force in July 2018. At the beginning of 2020, the Berlin Senate passed the draft law and forwarded it to the House of Representatives for further discussion and approval [5]. The plan aims at improving and extending the bike network in the city by 2030. The goal is to provide more and safer cycle paths in order to promote a healthy and environmentally friendly lifestyle in the capital of Germany. Cycling should increase from 18% of all trips made, as reported in a survey conducted in 2018, to 23% by 2023. Main objectives are: increasing pedestrian and cyclist safety, reducing traffic, improving the bike network and parking, promoting bike sharing and cargo bikes, improving the efficiency of the transport system as a

whole and making car traffic in Berlin climate-neutral by 2045 [5]. 600 million euros by 2030, which means 51 million euros annually, are to be invested in the expansion of the cycling infrastructure and in promoting cycling [73].

To promote the sustainable transition, in September 2021 a statutory ordinance established the rebuilding and development of a cycling network with a total length of 2,371 km. Of these, 550 km of new bike paths will be created on main roads and 100 km of high-speed bike connections. In addition, the plan aims at building new and more secure bike parking facilities especially around main U-bahn and S-bahn stations [5, 62].

Another interesting citizens' initiative called *Berlin Autofrei* (car-free Berlin) was officially launched in October 2021. It aims at creating the world's largest car-free urban area [2]. The idea is to drastically reduce car traffic in all streets within the "S-Bahn-Ring" trainline, which covers 88 sq km, meaning a surface that is bigger than Manhattan or equal in size to all the boroughs in London's zones 1 and 2. With some exceptions such as people with reduced mobility and public and emergency services that will still be allowed to use the car, all streets will be limited to walking, cycling and public transport. Furthermore, each person will be able to do up to twelve trips per year by car. This measure would have several benefits. First of all, it implies an equal redistribution of space for everyone who needs it. At the moment the majority of space is devoted to cars even though, according to 2014 data, only a third of journeys on Berlin streets were made by car [33]. Secondly, Berlin streets would become safer. In fact, according to 2020 road traffic accidents data provided by the Berlin city data portal, 79% of accidents involve cars, 68% of accidents by bike involve cars and 64% of pedestrian accidents involve cars. Thirdly, car-free streets would play an important role in terms of climate protection and healthy lifestyle. People would be encouraged to spend time outside walking, cycling or chilling in parks and they would breathe clean air [76]. Berlin Autofrei managed to collect 50,000 signatures since April 2021 and the Berlin Senate is at the moment considering the idea. If the city rejects it, Berlin Autofrei will try to collect 175,000 signatures, which will allow them to organise a referendum and citizens will decide [6].

Changing Cities, the organisation that promoted the successful cycling law in Berlin in 2016, is currently encouraging the creation of the so-called Kiezblocks. They are traffic-calmed neighbourhoods in which cut-through traffic has been eliminated to increase safety and allow neighbours to enjoy the streets for other activities. However, residents and public and emergency vehicles can still drive in those areas. Less cars means more space for walking, cycling, playing and doing sport. Further-

more, neighbourhoods will be safer especially for children and elderly and a sense of community will rise. Moreover, the environment would also benefit from this initiative as less people driving leads to less air pollution. In addition, since more people will spend time in their neighbourhood, local businesses would benefit too [38]. Changing Cities is helping neighbourhoods in collecting signatures and developing traffic calming plans to implement Kiezblocks. At the moment 54 neighbourhoods blocks out of 180 have started the process [37]. Bergmann Kiez is one of the first neighbourhoods that has implemented the concept of Kiezblocks. Another successful example of a car-free area is Friedrichstraße in Berlin-Mitte. A test, to evaluate the possibility to make the street a car-free zone, was conducted from the end of August 2020 to the end of October 2021. Due to the positive results, Friedrichstraße has become a car-free shopping street [18]. Furthermore, Oranienstraße in Berlin-Kreuzberg is another example of a street that will be rebuilt to become car-free by 2024. Except residents, no private motorised vehicles will be allowed. The choice is motivated by the fact that the street is considered dangerous as there is high car traffic and no separated cycle paths can be implemented due to lack of space [3].

5.3. Limitations and future research

Even though the methodology and models presented in this work could easily be used to analyse new cities, regions or countries, they first have to be adapted to the new context. As already mentioned, each geographic area presents different characteristics and hence, factors that contribute to increasing bicycle flow and bicycle accidents would have different levels of importance depending on the context. For instance, Berlin is characterised by a rather flat surface and as a consequence the so-called "hill score", which estimates the terrain's elevation, does not play an important role. Differently, I expect opposite results for Lisbon in Portugal, which is characterised by a very steep terrain that can really discourage cycling in certain areas. Therefore, before applying the prediction models, it is a good practice to adapt them to the characteristics of the geographic area under investigation. Future research could analyse in this way other German, European and non-European big cities as well as smaller towns in a comparative manner. Moreover, additional attributes such as bike sharing stations, bike repair shops, car parking spots and e-bike charging stations could be included in the analysis. Furthermore, more complex predictive models, which take into account the spatial structure of the data could be built to increase the performance. In addition, the same analysis could be proposed considering 2020 bicycle flow and bicycle accidents data. It would be interesting to see whether the movement restrictions imposed by the Covid-19 pandemic led to

significant changes. Finally, since cycling behaviour is the result of both objective measures of the built environment and subjective perceptions, future research could integrate the latter dimension into the analysis. It would be interesting to investigate through surveys which streets' configurations are perceived as the most accessible and the safest by cyclists and check whether the results are in line with the models' predictions. An interesting study concerning subjective safety was conducted in 2020 in Berlin by FixMyCity, an IT company based in Berlin, in collaboration with the Berlin newspaper *Tagesspiegel*. They built a survey that reached 21,401 respondents of which 19,109 come from Berlin. Respondents were first assigned to one category among pedestrians, cyclists and motorists based on some preliminary questions. Then, through the survey, participants were asked to rate 1,900 images depicting streets with different configurations and characteristics based on the subjective level of safety perceived. The collected data allowed to analyse how street attributes such as cycle path width, path surface, physical barriers, parking on the right side of the cycle path and so on, influence the sense of safety perceived by cyclists, pedestrians and motorists. The results show that the most relevant influencing factors when talking about street safety are: width of the cycle path, colouration of the cycle path and physical barrier alongside the flow of motor traffic. In particular, wide cycle paths, especially if there is a car parking on the right side, with a green colouration and a barrier that separates the paths from the flow of motor traffic, are proved to be perceived as the safest by respondents. Furthermore, paths along the pavement are on average rated safer as those along the road. Speed limits and traffic volumes only played a smaller role. Motorists' safety perceptions are in line with cyclists' ones, except from the fact that motorists seem to not be aware that car parks on the right side of a cycle path are perceived as dangerous by the majority of cyclists. Concerning pedestrians, they feel safe if the cycle area is well separated from the pedestrian one [17, 63]. The scenarios that people were asked to rate were realistic images, which did not correspond to existing street's configurations. In the future, more specific research could be conducted considering real scenarios in order to validate the results obtained with the prediction models presented in this work.

Chapter 6

Conclusion

Starting from the theoretical definitions of the concept of bikeability and several related studies, this research aims at investigating bikeability in Berlin through a data-driven approach. Unlike the majority of the reviewed studies, which present a one-dimensional bikeability index resulting from a weighted combination of several cycling-related indicators, in this work bikeability is explored considering its main sub-dimensions namely accessibility and safety. Of the predictors considered, 12 were built by combining various data sources, specifically bike lane score, intersection density, hill score, parks, traffic lights, POIs score, population, betweenness centrality, car traffic, car speed, bicycle parking capacity and network coverage. Bicycle flow and bicycle accidents are instead the dependent variables, that is the proxies to estimate accessibility and safety respectively. LASSO, ridge regression and Geographically Weighted Regression (GWR) were used to perform feature selection and predict bicycle flow, while LASSO, Decision Trees and Random Forest were used to select relevant features as well as perform bicycle accidents prediction. Concerning bike flow, all predictors were considered significant except for hill score, which was excluded from the analysis. The model with the best performance was GWR (RMSE $=4623.54$, $R^2=0.865$). Regarding bike accidents, parks were the only non-significant predictor, hence they were excluded from the analysis. Random Forest gave the best results in terms of performance (Accuracy $=0.805$). Results show that the main drivers of bicycle flow and thus accessibility are traffic lights (+), car traffic (+) and POIs score (+), while car traffic (+), traffic lights (+), population density (+) and POIs score (+) are the most relevant drivers of bicycle accidents and hence the factors that make areas more unsafe. Therefore, according to the analysis, accessibility and safety are related, meaning that the most popular areas with cyclists are also the least safe. Moreover, descriptive statistics show that the majority of accidents happened at crossings and intersections and involved cars too. Taking all these findings into consideration, coherent policies and urban planning measures could be implemented in order to promote cycling and make it a safer activity. In particular, areas with high bike flow and accidents or, when this information is not available, areas with high predicted bike flow and predicted presence of

accidents should be addressed first. Improving the quality of intersections together with reducing car traffic are the main measures suggested by these analyses in order to increase bikeability and promote sustainable travel in Berlin.

Bibliography

[1] 5,500 more Berliners compared to the end of 2020. https://www.statistik-berlin-brandenburg.de/031-2022.

[2] Berlin without cars – how it works. https://volksentscheid-berlin-autofrei.de/wie.php?lang=en.

[3] Berliner Bezirk will „Flaniermeile“ schaffen : Oranienstraße soll weitgehend autofrei werden - Berlin - Tagesspiegel. https://www.tagesspiegel.de/berlin/berliner-bezirk-will-flaniermeile-schaffen-oranienstrasse-soll-weitgehend-autofrei-werden/27989326.html.

[4] Berlin and its districts. https://www.statistik-berlin-brandenburg.de/248-2021.

[5] Berlin Mobility Act - Berlin.De. https://www.berlin.de/sen/uvk/en/traffic/transport-policy/berlin-mobility-act/.

[6] Berlin is planning a car-free area larger than Manhattan. https://www.fastcompany.com/90711961/berlin-is-planning-a-car-free-area-larger-than-manhattan.

[7] Berlin population density Germany 1995-2020 | Statista. https://www.statista.com/statistics/1109974/population-density-berlin-germany/.

[8] Bike Score Methodology. https://www.walkscore.com/bike-score-methodology.shtml.

[9] Thomas Blondiau, Bruno van Zeebroeck, and Holger Haubold. Economic Benefits of Increased Cycling. *Transportation Research Procedia*, 14:2306–2313, 2016.

[10] BMDV - Evaluation of crowdsourced data to improve municipal bicycle infrastructure - MOVEBIS. https://www.bmvi.de/SharedDocs/DE/Artikel/DG/mfund-projekte/verbesserung-der-fahrradinfrastruktur-movebis.html.

[11] Ugo N. Castañon and Paulo J. G. Ribeiro. Bikeability and Emerging Phenomena in Cycling: Exploratory Analysis and Review. *Sustainability*, 13(4):2394, Feb 2021.

[12] Copenhagenize. https://copenhagenizeindex.eu/about/the-index.

[13] Jeroen Johan de Hartog, Hanna Boogaard, Hans Nijland, and Gerard Hoek. Do the Health Benefits of Cycling Outweigh the Risks? *Environmental Health Perspectives*, 118(8):1109–1116, Aug 2010.

[14] Digitale LiDAR-Geländemodelle von Deutschland | Digital LiDAR-Terrain Models of Germany - Data Europa EU. https://data.europa.eu/data/datasets/dtm-germany?locale=en.

[15] Nikolaos Eliou, Athanasios Galanis, and APOSTOLOS PROIOS. Evaluation of the bikeability of a Greek city: Case study "City of Volos". 5, Jul 2009.

[16] FIS-Broker. https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart.

[17] FixMyBerlin. https://fixmyberlin.de/research/subjektive-sicherheit#umfragekonzept.

[18] «flaniermeile friedrichstraße» bleibt dauerhaft autofrei – Berlin.De. https://www.berlin.de/tourismus/infos/verkehr/nachrichten/7008014-4357821-flaniermeile-friedrichstrasse-bleibt-dau.html.

[19] Daniel Fuller and Meghan Winters. Income inequalities in Bike Score and bicycling to work in Canada. *Journal of Transport & Health*, 7:264–268, Dec 2017.

[20] Peter G. Furth, Maaza C. Mekuria, and Hilary Nixon. Network Connectivity for Low-Stress Bicycling. *Transportation Research Record: Journal of the Transportation Research Board*, 2587(1):41–49, Jan 2016.

[21] Global Spatial Autocorrelation – Geographic Data Science with Python. https://geographicdata.science/book/notebooks/06_spatial_autocorrelation.html.

[22] Thomas Götschi, Jan Garrard, and Billie Giles-Corti. Cycling as a Part of Daily Life: A Review of Health Perspectives. *Transport Reviews*, 36(1):45–71, Jan 2016.

[23] Elena Grigore, Norman Garrick, Raphael Fuhrer, and Ing. Kay W. Axhausen. Bikeability in Basel. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(6):607–617, Jun 2019.

[24] H3: Uber's Hexagonal Hierarchical Spatial Index. https://eng.uber.com/h3/.

[25] Zahra Hamidi, Rosalia Camporeale, and Leonardo Caggiani. Inequalities in access to bike-and-ride opportunities: Findings for the city of Malmö. *Transportation Research Part A: Policy and Practice*, 130:673–688, Dec 2019.

[26] Michael Hardinghaus, Simon Nieland, Marius Lehne, and Jan Weschke. More than Bike Lanes – A Multifactorial Index of Urban Bikeability. *Sustainability*, 13(21):11584, Oct 2021.

[27] HERE Urban Mobility Index. https://urbanmobilityindex.here.com/.

[28] Home - Targomo. https://www.targomo.com/.

[29] Tanya M. Horacek, E. Dede Yildirim, K. Kattelmann, O. Brown, C. Byrd-Bredbenner, S. Colby, G. Greene, S. Hoerr, T. Kidd, M. M. Koenings, J. Morrell, M. D Olfert, B. Phillips, K. Shelnutt, and A. White. Path Analysis of Campus Walkability/Bikeability and College Students' Physical Activity Attitudes, Behaviors, and Body Mass Index. *American Journal of Health Promotion*, 32(3):578–586, Mar 2018.

[30] How cities aim to be '20-minute friendly' to destinations like grocery stores and workplaces - The Washington Post. https://www.washingtonpost.com/business/2021/05/20/see-you-20-or-less-living-where-access-is-within-short-walk-or-bike-ride/.

[31] The Index - Copenhagenize. https://copenhagenizeindex.eu/the-index.

[32] Introduction | H3. https://h3geo.org/docs/.

[33] Leander Jones. Berlin's car ban campaign: 'It's about how we want to live, breathe and play'. *The Guardian*, Oct 2021.

[34] Mohamed Bayoumi Kamel, Tarek Sayed, and Alexander Bigazzi. A composite zonal index for biking attractiveness and safety. *Accident Analysis & Prevention*, 137:105439, Mar 2020.

[35] HaeLi Kang, Dong Ha Kim, and Seunghyun Yoo. Attributes of Perceived Bikeability in a Compact Urban Neighborhood Based on Qualitative Multi-Methods. *International Journal of Environmental Research and Public Health*, 16(19):3738, Oct 2019.

[36] Debra K. Kellstedt, John O. Spengler, Margaret Foster, Chanam Lee, and Jay E. Maddock. A Scoping Review of Bikeability Assessment Methods. *Journal of Community Health*, 46(1):211–224, Feb 2021.

[37] kiezblocks - Kiezblocks-Initiativen. https://www.kiezblocks.de/kiezblocks/.

[38] kiezblocks - Konzept. https://www.kiezblocks.de/konzept/.

[39] Puay Ping Koh and Yiik Diew Wong. Influence of infrastructural compatibility factors on walking and cycling route choices. *Journal of Environmental Psychology*, 36:202–213, Dec 2013.

[40] Patricia Jasmin Krenn, Pekka Oja, and Sylvia Titze. Development of a Bikeability Index to Assess the Bicycle-Friendliness of Urban Environments. *Open Journal of Civil Engineering*, 05(04):451–459, 2015.

[41] Yee Leung, Chang-Lin Mei, and Wen-Xiu Zhang. Testing for Spatial Autocorrelation among the Residuals of the Geographically Weighted Regression. *Environment and Planning A: Economy and Space*, 32(5):871–890, May 2000.

[42] Jen-Jia Lin and Yi-Hsuan Wei. Assessing area-wide bikeability: A grey analytic network process. *Transportation Research Part A: Policy and Practice*, 113:381–396, Jul 2018.

[43] Local Spatial Autocorrelation – Geographic Data Science with Python. https://geographicdata.science/book/notebooks/07_local_autocorrelation.html.

[44] Michael B. Lowry, Daniel Callister, Maureen Gresham, and Brandon Moore. Assessment of Communitywide Bikeability with Bicycle Level of Service. *Transportation Research Record: Journal of the Transportation Research Board*, 2314(1):41–48, Jan 2012.

[45] Liang Ma and Jennifer Dill. Do people's perceptions of neighborhood bikeability match "Reality"? *Journal of Transport and Land Use*, Jan 2016.

[46] Richard Manton, Henrike Rau, Frances Fahy, Jerome Sheahan, and Eoghan Clifford. Using mental mapping to unpack perceived cycling risk. *Accident Analysis & Prevention*, 88:138–149, Mar 2016.

[47] Bendik Manum, Tobias Nordström, Jorge Gil, Leonard Nilsson, and Lars Hilding Marcus. Modelling bikeability; Space syntax based measures applied in examining speeds and flows of bicycling in Gothenburg. 2017.

[48] Nathan McNeil. Bikeability and the 20-min Neighborhood: How Infrastructure and Destinations Influence Bicycle Accessibility. *Transportation Research Record: Journal of the Transportation Research Board*, 2247(1):53–63, Jan 2011.

[49] Harvey J. Miller. Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004.

[50] Momepy documentation – momepy 0.5.2 documentation. http://docs.momepy.org/en/stable/.

[51] Multiscale Geographically Weighted Regression (MGWR) – mgwr v2.1.1 Manual. https://mgwr.readthedocs.io/en/latest/index.html.

[52] Pekka Oja, Stephanie Titze, Andreas Bauman, Bas De Geus, Philipp Krenn, Bill Reger-Nash, and Timo Kohlberger. Health benefits of cycling: a systematic review: Cycling and health. *Scandinavian Journal of Medicine & Science in Sports*, 21(4):496–509, Aug 2011.

[53] OpenStreetMap. https://www.openstreetmap.org/#map=10/52.5097/13.3614.

[54] OSMnx 1.1.2 – OSMnx 1.1.2 documentation. https://osmnx.readthedocs.io/en/stable/

[55] Ottawa Cycling Level of Traffic Stress Map. https://maps.bikeottawa.ca/lts/.

[56] Anna K. Porter, Harold W. Kohl, Adriana Pérez, Belinda Reininger, Kelley Pettee Gabriel, and Deborah Salvo. Bikeability: Assessing the Objectively Measured Environment in Relation to Recreation and Transportation Bicycling. *Environment and Behavior*, 52(8):861–894, Oct 2020.

[57] Protected Intersections for Bicyclists | A new design for US streets. http://www.protectedintersection.com/.

[58] John Pucher, Ralph Buehler, David R. Bassett, and Andrew L. Dannenberg. Walking and Cycling to Health: A Comparative Analysis of City, State, and International Data. *American Journal of Public Health*, 100(10):1986–1992, Oct 2010.

[59] Brian E. Saelens, James F. Sallis, and Lawrence D. Frank. Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine*, 25(2):80–91, Apr 2003.

[60] Jonas Schmid-Querg, Andreas Keler, and Georgios Grigoropoulos. The Munich Bikeability Index: A Practical Approach for Measuring Urban Bikeability. *Sustainability*, 13(1):428, Jan 2021.

[61] Scikit-learn: machine learning in Python – scikit-learn 1.0.2 documentation. https://scikit-learn.org/stable/.

[62] Senat beschließt Radverkehrsplan: 3000 Kilometer Fahrradnetz entstehen in Berlin - Berlin - Tagesspiegel. https://www.tagesspiegel.de/berlin/senat-beschliesst-radverkehrsplan-3000-kilometer-fahrradnetz-entstehen-in-berlin/27805286.html.

[63] Solche Straßen will Berlin | Tagesspiegel. https://interaktiv.tagesspiegel.de/lab/strassencheck-ergbnisse-diese-strassen-will-berlin/.

[64] Spatial Weights – Geographic Data Science with Python. https://geographicdata.science/book/notebooks/04_spatial_weights.html.

[65] STADTRADELN - Home. https://www.stadtradeln.de/home.

[66] Statistisches Bundesamt Deutschland - GENESIS-Online: Die Datenbank des Statistischen Bundesamtes. https://www-genesis.destatis.de/genesis/online?sequenz=statistikTabellen.

[67] Straßenverkehrsunfälle nach Unfallort in Berlin 2019 | Offene Daten Berlin. https://daten.berlin.de/datensaetze/strassenverkehrsunf%C3%A4lle-nach-unfallort-berlin-2019.

[68] Heatmap Radverkehr - mCLOUD. https://www.mcloud.de/web/guest/suche/-/results/detail/3096DB7A-9EE4-4C14-B2AA-79E33A7FFF01.

[69] Sustainable Development Goals & Cycling - United Nations Western Europe. https://unric.org/en/sustainable-development-goals-cycling/.

[70] Table of Cell Areas for H3 Resolutions | H3. https://h3geo.org/docs/core-library/restable/.

[71] Travel Times API | Targomo Developers! https://www.targomo.com/developers/apis/travel_time/.

[72] About Us - The Bikeability Trust Organisation | Bikeability. https://www.bikeability.org.uk/about/.

[73] Verein | Changing Cities. https://changing-cities.org/verein/.

[74] Volksentscheid Fahrrad – in English - Volksentscheid Fahrrad. https://volksentscheid-fahrrad.de/de/english/.

[75] Welcome to the QGIS project! https://www.qgis.org/en/site/.

[76] Why car-free? https://volksentscheid-berlin-autofrei.de/warum.php?lang=en.

[77] Meghan Winters, Michael Brauer, Eleanor M Setton, and Kay Teschke. Mapping bikeability: a spatial tool to support sustainable travel. *Environment and Planning B: Planning and Design*, 40(5):865–883, 2013.

[78] Meghan Winters, Kay Teschke, Michael Brauer, and Daniel Fuller. Bike Score ®: Associations between urban bikeability and cycling behavior in 24 cities. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1):18, Dec 2016.

[79] Zählstellen und Fahrradbarometer: Fahrradverkehr in Zahlen - Berlin.De. https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/#dauer.

[80] Mohamed Anwer Zayed. Towards an index of city readiness for cycling. *International Journal of Transportation Science and Technology*, 5(3):210–225, Oct 2016.

[81] Dennis Zielstra and Alexander Zipf. Quantitative Studies on the Data Quality of OpenStreetMap in Germany. 2010.