University of Zurich^UZH

Modelling Bicycle Ridership Using Crowdsourced Data in the Urban Area of Zurich

GEO 511 Master's Thesis

Author
Thomas Özvegyi
17-714-536

Supervised by
Prof. Dr. Ross Purves
Dr. Ralph Straumann (ralph.straumann@ebp.ch)
Arnim Wagner (arnim.wagner@vd.zh.ch)

Faculty representative
Prof. Dr. Ross Purves

26.08.2023
Department of Geography, University of Zurich

Abstract

Background: Over the past years, great efforts are being taken to improve infrastructure for bicycling by the Canton and the City of Zurich. However, as in most places, data on bicycle ridership is still sparse and relies on counting stations at specific points in the network. On the other hand, crowdsourced data from Strava provides a continuous spatial map of ridership. The user bias of Strava data can be mitigated by including geographic variables as well as official counts to model ridership.

Goals: This thesis explores how accurate bicycle ridership can be predicted using Strava counts and geographic variables with official counts as ground truth. The goal is to get a spatially continuous map of bicycle ridership at street segment level for the urban area of Zurich. Secondly, the aim is to find out which geographic covariates are the best local predictors for bicycle ridership.

Methods: The most significant geographic variables were selected using a LASSO Regression. This work distinguishes between in-sample and out-of-sample estimations of ridership by using two distinct sampling strategies. Three Generalised Linear Mixed Models were fitted: One for each sampling strategy using cross-validated 80-20 train-test splits and one GLMM using all counting stations as training data.

Results: Results show very good in-sample accuracies and only moderate and highly unstable out-of-sample estimates. The models perform worse at stations with high seasonality, where primarily leisure riders are present. Significant predictors besides the Strava counts were the socio-economic Swiss Neighbourhood Index, distance to points of interest, exposure to accidents and winter as season.

Conclusion: Ridership is predictable using Strava and geographic data. For better performance, mainly on unseen data, it is suggested to either add more variables concerning seasonality or perform separate models for leisure-oriented paths.

Key words: Bicycle ridership, Strava, active transportation, crowdsourced data

Acknowledgements

This thesis would not have been possible without the support of several people:

- Special thanks to my supervisor, Prof. Dr. Ross Purves, who guided me from start to finish of this thesis. Thanks for all advice and your constant and encouraging reminders that I'm on track, when I already worried about a following step.

- Dr. Ralph Straumann of EBP, who helped me with his practical perspective and his expertise in rather complex topics such as network matching and modelling.

- Arnim Wagner and fellow employees of the Office of Mobility and Transport (AFM) of the Canton of Zurich. Thanks to Arnim's efforts, I got access to the Strava Data which was essential for this work. The collaboration with the AFM motivated me additionally to produce relevant results.

- My family. Thanks for your continuous emotional support and proofreading this work.

Contents

List of Figures

List of Tables and Equations

Abbreviations

AADB	Average Annual Daily Bicyclist
AIC	Akaike Information Criterion
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
MAE	Mean Absolute Error (see 4.2.3 for Formula)
MAPE	Mean Absolute Percentage Error (see 4.2.3 for Formula)
LASSO	Least Absolute Shrinkage and Selection Operator
OGD	Open Government Data
OSM	Open Street Map
POI	Points of Interest (see 3.2.2)
RMSE	Root Mean Squared Error (see 4.2.3 for Formula)
VIF	Variance Inflation Factor

1 Introduction

1.1 Motivation

Mobility is one of the sectors which account for most CO2 emissions in Central European Countries. Furthermore, it's the only sector that has increased its greenhouse emissions compared to 1990, instead of a decline (Eurostat, 2022). One highly valued attempt to make traffic more sustainable is the promotion of bicycling as mode of transport (Larsen et al., 2013). In recent years, authorities like the City of Zurich have got the strong political duty for the promotion of cycling (Kanton Zürich, 2020; Stadt Zürich, 2020), shifting towards emission free and space efficient modes of mobility. Thereby, they focus on improving conditions for bicycling in everyday life, for example for commuting or running errands (Kanton Zürich, 2023; Stadt Zürich, 2021a).

However, environmental benefits are not the only reason for the promotion of bicycling: In the past years, several studies have shown the positives of active transportation in general (Mueller et al., 2015). For instance, Celis-Morales et al. (2017) associate cycle commuting with lower risk for cardio-vascular disease, cancer and all-cause mortality, which shows that bicycling also provides benefits for health.

To support decision makers in their aim to promote cycling, insights in the actual usage of infrastructure is key, both for existing bike paths or lanes and the building of new ones. Up to today, most of the data about ridership is only punctual and thus represents the ridership volume at specific streets in the network (Graser et al., 2021; Livingston et al., 2021). This is also the case in Zurich, Switzerland, where both the city and the canton operate automated counters at specific locations (Kanton Zürich, 2022; Stadt Zürich, 2022). These counters deliver temporally rich data, as they work automatically around the clock. Nevertheless, they are a limited data resource, as they lack spatial insights into the ridership in a holistic, network-based manner.

In contrast to punctual count data, crowdsourced GPS data from tracking apps deliver a continuous flow of data points. In the 2010s, tracking apps like Strava have grown constantly - in 2020, Strava numbered 50 millions of users out of 195 countries tracking and sharing their activities online (Strava, 2020a). Soon, research has started to see the potential of this generated data. First studies emerged which tried to use the tracking data to predict ridership (Jestico et al., 2016) or assess infrastructure (Heesch & Langdon, 2016). At the same time, there also critical remarks about user submitted data and studies emerged which assess the accuracy of crowdsourced geographic features (e.g. Jackson et al., 2013).

While there are pitfalls of crowdsourced data at both the part of the user and the providers, the enormous potential of those vast amounts of data stands out. In 2019, Roy et al. presented a study to predict bicycle ridership, including a range of geographic variables that correlate to bicycling. The use of a range of other variables is their way to statistically mitigate the pitfalls of crowdsourced data.

The motivation behind this thesis is to explore the potential of bicycle ridership prediction using crowdsourced data in Zurich, Switzerland. The use of a model combining official counts as ground truth and crowdsourced data along with geographic variables as predictors is promising, as it helps to compensate for the weakness of each input. A spatially continuous model of ridership would improve the foundations for decision making in the enhancement of bicycle infrastructure by the authorities.

1.2 Research Questions, Hypotheses, Goals

The goal of this thesis is to contribute to the state of the art concerning bicycle ridership. Both the methods and results should enhance the understanding of the use of crowdsourced data for ridership prediction in Switzerland and provide a first step to close the research gap described in Section 2.4.

In context of the research gap and existing studies, the following research questions are to be answered in this study:

- RQ 1: How well is bicycle ridership in Zurich's Urban Area predictable from Strava and geographic data?
- RQ 2: Which variables are the best local predictors for ridership?

As for hypotheses, the subsequent assumptions will be questioned throughout this thesis:

H1: It is possible to predict ridership as precise as in related studies, meaning that for 80% of segments, daily counts of ±25% of riders can be predicted.

H2: There are differences in variable importance compared to related studies in the US – for instance, in Zurich's Urban Area, socio-economic variables might not be significant predictors for ridership.

As contained in the first hypothesis, a concrete goal is to predict monthly counts within the same ranges of accuracy as in related studies from the US. As for time units of the prediction, this study aims for predicting monthly counts, but also provide outputs for the annual average daily bicyclists (AADB) and average daily bicyclists in the period of April-October.

2 State of the Art

In this chapter, the background and existing research is presented to contextualise the topics of this thesis. First, the use of crowdsourced bicycle ridership data at the example of Strava is assessed in Section 2.1. Exemplary studies that already predicted bicycle ridership using Strava data are described in Section 2.2, followed by an overview of the current state of bicycling and its associated data in Zurich in Section 2.3. Lastly, the research gaps are summarised in Section 2.4.

2.1 Crowdsourced Bicycle Ridership Data at the Example of Strava

As mentioned in the motivation, crowdsourced data is a phenomenon that emerged over the past decade. In research, crowdsourced data is also named a type of emerging data, in contrast to traditional data which comes from manual counts or automated counts (see Figure 1).

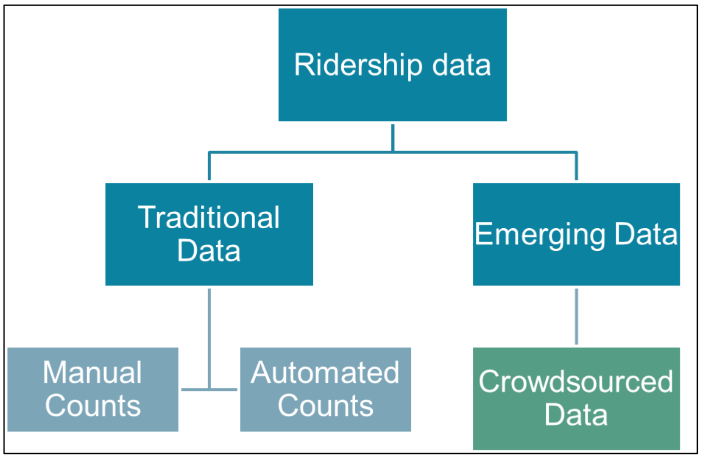


Figure 1: Types of Ridership Data

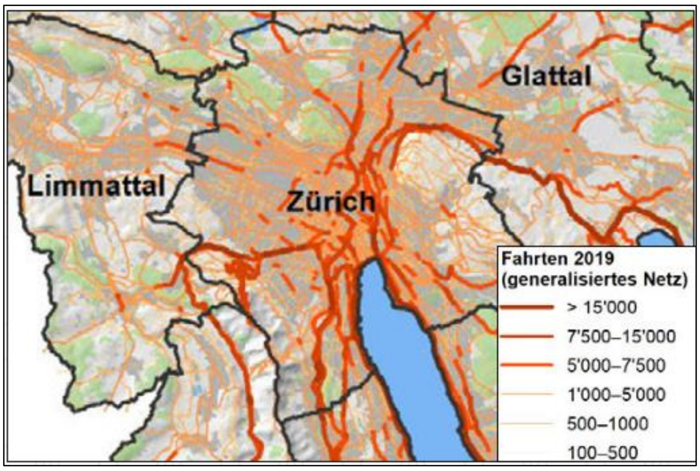
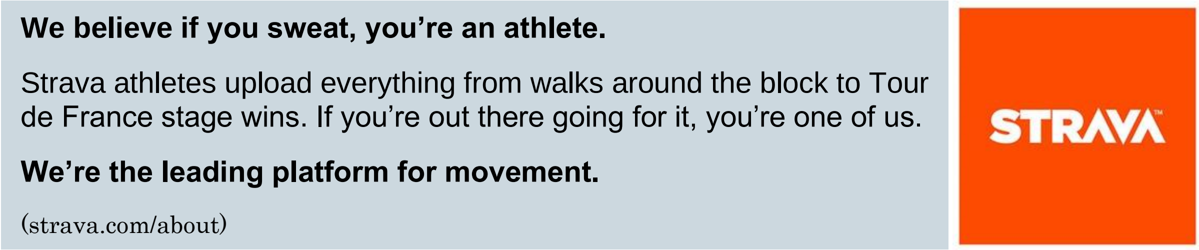


Figure 2: Continuous Map of Strava Bicycle Trips in Zurich in 2019 (EBP, 2020)

The vast advantage of crowdsourced data is the spatial coverage. Contrary to measures at specific points in a network, crowdsourced data is continuous (see Figure 2).

In the case of Strava, users track themselves to share their activities online, in what could be called a social network for athletes. Thereby, all users are motivated to contribute with encouraging messages (see Figure 3). The base functions of Strava are free to use, whereas the membership for CHF 12 a month (as of July 2023) unlocks further tools and options.



We believe if you sweat, you're an athlete.

Strava athletes upload everything from walks around the block to Tour de France stage wins. If you're out there going for it, you're one of us.

We're the leading platform for movement.

(strava.com/about)

Figure 3: Exemplary Slogans and Logo of Strava

The app can be used on smartphones but can also be synced with smart wearables of different brands.

Strava has its own data platform called Strava Metro. Through this platform, Strava provides the data of its users in an anonymised way. Public authorities, or anyone who works for them, can apply and get access to Strava Metro. The aim of Strava Metro is to help authorities make data-driven decisions and thus improve infrastructure (Strava, 2020b).

What are the things to consider when working with crowdsourced data? Among the concerns is the data acquisition setting, the app design and data characteristics (Tironi & Valderrama, 2017). These factors may change considerably over time, as companies like Strava follow their own interests, or incentives for users change. Research needs high data quality and users must not be nudged to choose certain options. For privately owned data, such characteristics are impossible to control. However, the most critical thing about crowdsourced ridership data is the missing representativeness. The sample of crowdsourced data does not represent the population. Often, few users contribute an over proportionally amount of data. Moreover, the contribution is heavily biased in terms of gender and age, mostly towards male and younger persons (Garber et al., 2019).

Besides the population, the bias towards recreational riders is an inherent problem in crowdsourced data (Garber et al., 2019; Roy et al., 2019). Recreational riders may choose other routes and have other needs than the average bicyclist. On the part of providers, the binning of usage data to boost privacy of users affects the data quality (Raturi et al., 2021). Concerning the Strava data in this study, this is discussed in Chapter 3.2.1.

Summing up, there are concerns about data quality (acquisition settings for users, binning by providers) and inherent biases towards certain groups of people, because the data is not representative. These aspects show that the usage of crowdsourced data must be carefully assessed.

2.2 Research: Crowdsourced Data to Predict Ridership

In literature, several researchers have linked crowdsourced data to official counts, to both mitigate the downsides and sum up the advantages of both methods. Over the past decade several studies emerged which focused on the estimation of ridership.

While the common goal is the same, the studies differ in the observation time considered, number of counting stations, geographic variables or type of accuracy specified. Jestico et al. (2016) provide one of the earliest studies. They used manually computed cycling counts in peak traffic hours of Victoria BC, Canada. The manual counts were done on several days in January, May, July and October, resulting in a separate ridership map for each of those months. They had an average model error of 38% over all iterations of the cross-validation, where they randomly selected 90% train and 10% test samples of the stations.

One of the two most important related studies for this thesis is written by Roy et al. (2019). As the title “Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists” says, they focus on the methodology to make the insights of crowdsourced data inclusive for all riders. They state that the inclusion of a variety of other variables accounts for the bias in crowdsourced data and makes the results representative for all riders (Roy et al., 2019, p. 15). Their study is conducted in Maricopa County, Arizona, USA. Interestingly, they use 44 counters spread in the whole County to train their data, whereas they validate their model using 60 manually computed counts during peak hours, in the City of Tempe, inside their perimeter. While the climate is arid, they do account for seasonality using April, May, October and November as sampling months. Data is extrapolated to arrive at a range of results for the AADB. According to them, their model can predict ±25% of riders for 80.3% of segments. This relative figure is written in their discussion (p.15), whereas otherwise, they refer to absolute numbers of riders in their accuracies, which makes findings hard to interpret. They provide out-of-sample accuracies, as they tested the model with another type of count data, that was manually conduced in the City of Tempe and also extrapolated from peak hours to whole day figures. The in-sample Poisson model fit in the train set resulted in an R^2 of 0.64. Even though they provide precise accuracies they refer to categorical maps as the most accurate output for prediction.

The second reference paper is by Nelson et al. (2021). They propose a generalised model for different cities in the US and Canada, involving various kinds of networks, bike cultures and climates. However, they also compute city-specific models, where another set of geographic variables is selected. As one would expect, this mostly leads to better results. In terms of numerical results, only in-sample accuracies were produced. In the city specific models, the accuracy, respectively the R^2 value varied between 0.76 and 0.92. Due to the diverse setting, they formulate four general recommendations for ridership models using crowdsourced data:

- Four variables to always use (number of Strava riders, % of Strava riders commuting, bicycle crash density and median household income)
- Further possibly relevant variables (e.g. slope, distance to residential area)
- The importance of the official counts regarding temporal and spatial characteristics
- Not to oversell results – categorical maps are the most reliable output they name, equally as Roy et al. (2019).

And what about Europe? Interestingly, despite some countries where bicycling has a very high share in daily mobility, there are only few studies with the same research goals. Livingston et al. (2021) present a study done with data from Glasgow, Scotland in which they also aim to fill gaps left by related studies concerning the characteristics of counters and making out-of-sample predictions. In contrast to other studies, they do not consider geographical variables and acknowledge the bias of Strava users as a limitation. They examined the differences in correlation between smaller and larger time aggregations and tried out-of-sample predictions testing a variety of models. The best model fit resulted using a negative binomial model. Finally, their models can predict order of magnitudes throughout the whole city, but they state that for applications that require more precision the signal to noise ratio is not appropriate. Livingston et al. (2021) still think that a precise prediction could be obtained, if more variables would be included and the number of Strava riders is not too low.

2.3 Bicycling in Zurich

The authorities of the Canton and the City of Zurich are investing a lot to improve infrastructure for bicyclists, as the political forces demand (Kanton Zürich, 2020; Stadt Zürich, 2020). A major part of an improvement of the situation should come from priority routes for bicycles. While the terms (Veloschnellrouten, respectively Velovorzugsrouten) and the corresponding definitions are slightly different between Canton and City, the aim is the same – there should be a network of routes where bicyclists of all ages and moving velocities are safe to ride. The City of Zurich talks about different types of bicycle riders concerning the frequency of rides in their strategy paper (Stadt Zürich, 2021b). They state that the future network aims to improve conditions for all groups. For the Canton of Zurich, thanks to the existence of safe and undisrupted routes, more commuters should be encouraged to refer to bicycles as mode of transport (Kanton Zürich, 2021). Even though plans are there, the implementation is taking a lot of time. Objections from local people, a few political parties or business owners delay ongoing projects, whereas the most controversial subject seems to be parking space that is removed to make space for the bicycle routes (Brun, 2022).

Authorities of City and Canton still rely on the punctual nature of automated counters, surveys and estimations to assess bicycle ridership on their infrastructure. While for motorised individual traffic and public transport, there is a detailed transport demand model operated by the Canton of Zurich, bicycles are only roughly captured as of today. In the future, cycling should be represented more completely (Amt für Mobilität, personal communication, 05.12.2022). In the meantime, other ways are explored to get a better understanding about ridership: In 2020, the Canton of Zurich explored the "Suitability of Strava Data For Questions Of Bike Traffic" (EBP, 2020). The descriptive study analyses patterns of Strava usage in the whole Canton and shows the correlation between counters and Strava counts in the year 2019. This thesis builds on that former study – it follows one of the further research possibilities named by the authors. The City of Zurich is exploring another approach: They asked for people who voluntarily share their daily movement patterns with the authorities, with the incentive that the data would improve the planning of adequate infrastructure (Stadt Zuerich, 2022). These projects all share the same goal: Using generated data of our digitalised world to improve the understanding of how people move and behave, in our case in bicycling.

As for research, there are various studies concerning bicycling around Zurich (e.g. Büchel et al., 2022; Meister et al., 2022; Menghini et al., 2010). However, the studies are mostly focused on route choice or safety assessments. The spatial ridership prediction has not been covered yet. Ridership has often been used in assessments, for example in many studies that covered the changes in mobility during the Covid-19 pandemic (e.g. Lustenberger et al., 2021).

2.4 Research Gap

The state of the art unveiled several research gaps:

- Whether for Zurich, nor Switzerland there is research about the prediction of bicycle ridership that involves the use of crowdsourced data.
- The same applies to geographic covariates, which have only been named in related studies from overseas.
- Some existing studies only presented in-sample accuracies, while predictions of unseen data could not be tested. Moreover, the difference between in- and out-of-sample estimates has not been addressed yet.

With the research questions and goals stated in Section 1.2, this thesis tries to close those research gaps and contribute to a better understanding of bicycle ridership in a spatially continuous manner.

3 Case Study and Data

This thesis involves a high amount of data, which is partly used in a generalised linear mixed model. Besides the ridership data of Strava and the official counts provided by the City, respectively the Canton of Zurich, a wide variety of geographical co-variates is tested. An overview over the study area is given in Section 3.1. Section 3.2 reflects on the data used in this thesis: The implementations of ridership data and geographic data are described and assessed.

3.1 Study Area

The study area is involving Zurich, Switzerland and a big part of the urban area around the city. It was drawn so that a variety of official bicycling counting stations operated by the City or the Canton of Zurich could be included. In Figure 4, an overview about the study area is given: All finally used counting stations, the convex hull of the final perimeter and the spatial division into urban and not urban area are shown. This division is deduced from the five action areas in the development plan of the Canton (Raumordnungkonzept ROK, kantonaler Richtplan). The reason for focusing on urban area is that the authorities focus on improving conditions for bicyclists in everyday life and not in a leisure context. Therefore, it was decided to focus on the City and the urban Limmattal and Glattal nearby.

The in- or exclusion of counting stations will be explained more thoroughly during this thesis. Of the stations used in the end, most are inside or at the border of urban area. The convex hull of the perimeter ("Study Area" in Figure 4) is deduced from the rectangular export of Strava Metro. This area has been shrunk a bit further due to time limits in the network matching.

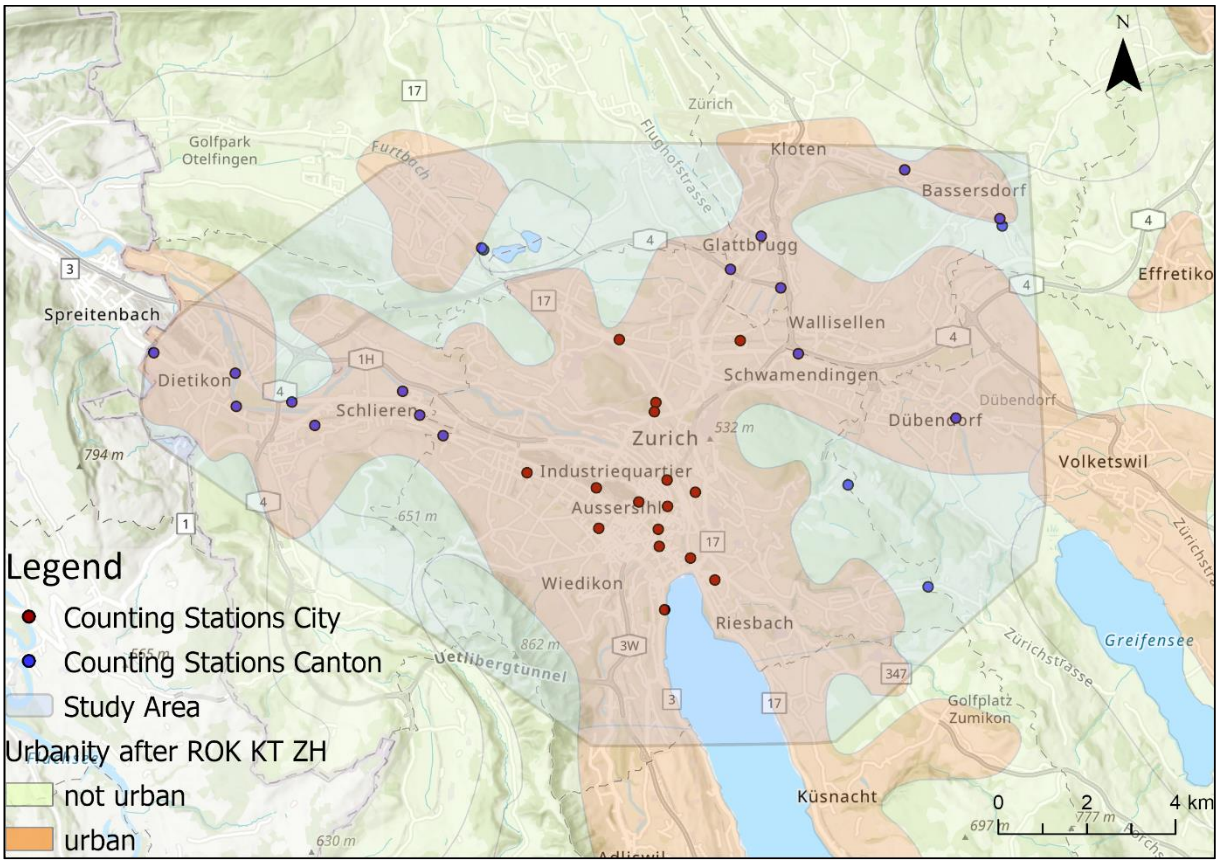


Figure 4: Study area: Counting Stations in Zurich's Urban Area (own map using ArcGIS Pro)

3.2 Data

3.2.1 Ridership Data

The main data for this analysis is bicycle ridership data. In Table 1, an overview over the six ridership variables is given, while in Figure 4, the counting stations are shown on a map. Furthermore, a table containing details of all counting stations and their considered measurements can be found in the Appendix (Letter B).

Table 1: Ridership Data used for this Study

	Variable	Description	Source
a)	Official Counts Canton of Zurich	20 locations, aggregated monthly, cross-sectional data	Canton of Zurich
b)	Official Counts City of Zurich	16 locations, aggregated to monthly, cross-sectional data	City of Zurich[1] (OGD)

[1] Available at https://data.stadt-zuerich.ch/dataset/ted_taz_verkehrszaehlungen_werte_fussgaenger_velo (Assessed on the 16/08/2023)

c)	Strava Counts on edge-level	Monthly counts on street segment level;	Strava Metro
d)	Strava % of commute	Additional predictor variable. Significant predictor in the study by Nelson et al. (2021). See Chapter 3.2.1.	Strava Metro
e)	Season of Count	Additional predictor variable. Derived from months: Winter: 12-02 Summer: 06-08 Spring: 03-05 Autumn: 09-11	Strava Metro
f)	Year of Count	Additional predictor variable. 2021 or 2022.	Strava Metro

a) and b): Official Counts from Counting Stations by the City or the Canton

The counts recorded by counting stations were imported from .csv respectively .xlsx files into R. While the counting stations of the City of Zurich were all provided in one file, the Canton provided one Excel file per counting station. Therefore, all stations had to be merged first. The data was then filtered for the years 2021 and 2022. As the raw data was quarter-hourly or hourly data, the lubridate package was used to aggregate the counts to monthly data.

A link to the Strava data was necessary in order to use the counts in the analysis. The matching of the counting stations to a unique street segment of the network was done manually. As described in Section 4.1, the network, with very few exceptions, consists of one segment per cross-section, independent of the number of lanes or bicycle infrastructure. In some cases however, the authorities use two or more counting stations to capture the traffic on one cross-section. For these stations, the values were added to obtain the total ridership per cross-section.

For the use of official counts as ground truth in the regression models, the data quality is very important. Whereas the Strava data was available all year round, there were a lot of gaps in the official counts. In some cases, the stations were newly installed, in others there was a technical problem that led to missing data. For the monthly counts used in the regression, only months that had >= 25 days of recorded counts were used.

Another uncertainty is the counting itself. The two authorities work with similar methods of measurement: An induction loop is inserted on the pavement (see Figure 5).

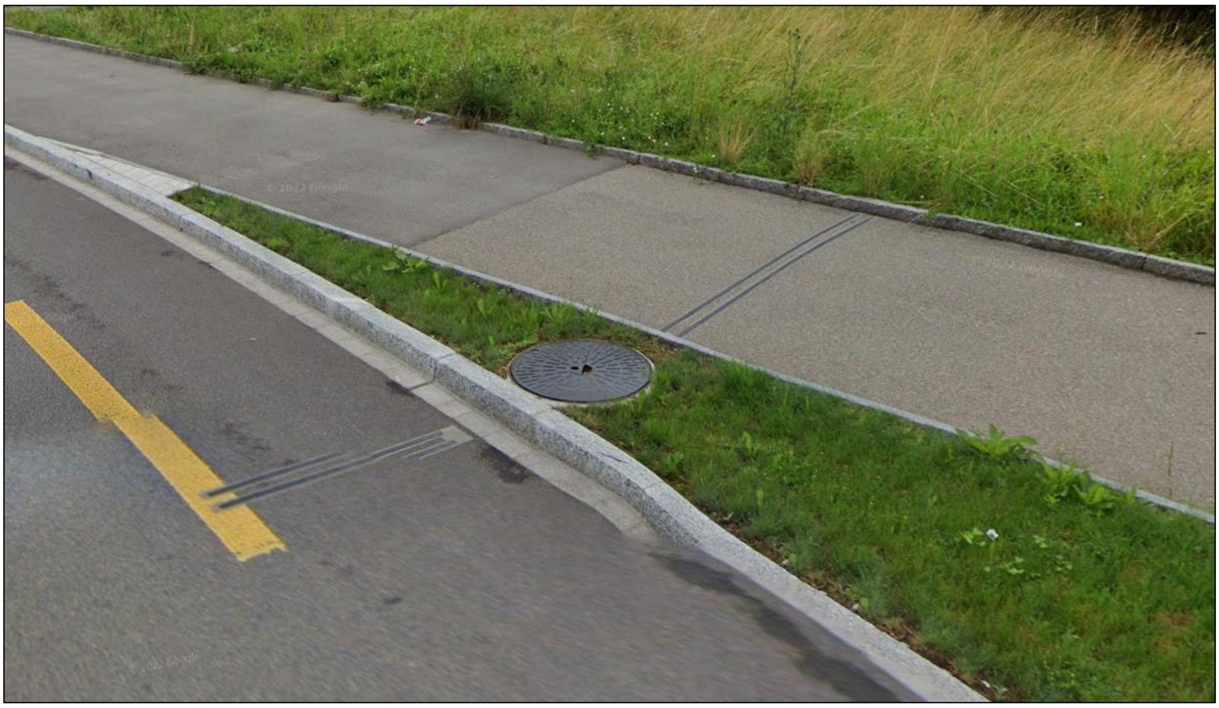


Figure 5: Example of an Induction Loop in Glattbrugg (Screenshot of Google Street View on www.maps.google.com, accessed on 16/08/2023)

The induction loop is connected with a little computer, which sends the counts to a server each day. The counting works as every bicycle containing metal is emitting an electromagnetic signal when it passes by, which is recognised by the induction loops. Further classifications of the signal filter out motorcycles or even bikes that are only pushed. The measurement with induction loops has the disadvantage that bicycles made purely out of carbon are not counted. Further uncertainties can be double counts or missed bicycles, depending on how the loops are placed and on the path somebody takes when riding over them.

The Canton and the City of Zurich use the same technology, but not the same model of counters. The Canton directly uses the counts of its stations, while the City works with correction factors for most stations, which are factors that correct the stations' count according to manual counts conducted. For some stations these factors are considerably higher than for others. This study has used those factors and multiplied all stations counts to correct them.

c) – f) Strava Counts: Characteristics and Things to Consider

The GPS data from Strava trips (*c)* in Table 1) is matched to OpenStreetMap (OSM) features. All OSM features that contain counts appear as geometries in the export of Strava Metro. OSM contains a lot of diverse features in their maps – besides the main lanes of streets, features like tramways, sidewalks or bike lanes are separate features. Unfortunately, the GPS data often gets matched to such features. A part of it might be rightfully, however the result is not very practicable for analysis of ridership as Figure 6 exemplary shows: The crossing around the tram station ETH/Universitätsstrasse has a vast number of edges and links between them.

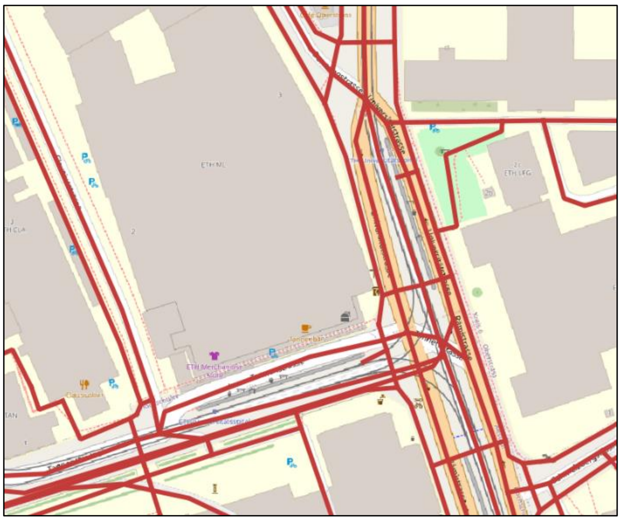


Figure 6: Example of Multiple Strava Edges at Universitätsstrasse, Zurich (own screenshot QGIS)

It would be little useful to predict ridership on this network. For this reason, the Strava counts on this OSM network have been matched to a more suitable network. This is explained in Section 4.1.

In the Strava data, there is a lot of additional information besides the bare count. The commute label (*d)* in Table 1) lets users mark all trips that have a non-leisure purpose. As not all users make use of that option, Strava uses certain definitions to decide whether a trip is marked as leisure or commute. The tagged commute trips thereby serve as validation for the commute trips that were identified using diverse models (Sunde, 2019).

In principle, commutes or other trips with a non-leisure purpose are the trips that this study is interested in. However, the study area itself might partly account for that as it is mostly urban area and, even if transport planning is focusing on promoting cycling for non-leisure trips, for the usage of infrastructure it actually does not matter what the purpose of the user is. To use the commute label without losing data, this work adopted the method already used by Nelson et al. (2021) and used the % of commute trips as an additional variable available for the regression analysis.

The month and the year (*e)* and *f)* in Table 1) of the respective counts are further additional attributes that are used in the modelling process. The aggregation of the months to seasons proved to be beneficial for the performance of the model and was therefore maintained for all models.

As mentioned, there is a strong bias regarding gender and age in the Strava data. 88.7% of Strava users around Zurich were between 18 and 54 years old in 2022 (shown in the Strava Metro Dashboard). The gender distribution is not directly visible. It can be assumed that its similar to the study by EBP (2020, p.13), where male users accounted for 80% of total users.

Another point to keep in mind is the binning of data by Strava, executed as privacy protection measure. For every edge and hour, 0-3 riders are rounded down to 0, while 4 and all other numbers of riders are rounded up to the nearest multiple of 5. This means an information loss, especially in low-ridership edges. When comparing data before and after the binning, Raturi et al. (2021) even found that in extreme cases, the opposite conclusions can be drawn. In respect to this study, the use of monthly data instead of hourly or daily counts, and no selection on commute or leisure trips are measures that mitigate this information loss. Concerning the last point, Raturi et al. (2021) remark that the binning treats leisure and commuting unequally, as commuting mostly takes place in peak hours (less omitted riders because of more concentrated flows), while leisure activities are dispersed temporally and spatially.

3.2.2 Geographic Data

As proposed by Roy et al. (2019), geographic covariates are included to account for the missing representativeness in the Strava data. Table 3 shows all geographic variables, their unit, granularity and references. If the data is Open Government Data (OGD), the link is provided All variables were prepared and computed using R (R Core Team, 2023) or QGIS (QGIS.org, 2023).. In the Appendix (Letter C), histograms of all variables can be found.

a) Exposure to Accidents: The correlation to accidents involving bicycles is straightforward: The more accidents, the more bicycles probably have passed this segment of the network. In the exemplary research by Nelson et al. (2021), bicycle crash density belongs to the most significant covariates to ridership across different cities. In a wider context, safety and street conditions are a major factor of people's choice for active transportation modes (Hankey et al., 2012). There is OGD of all accidents represented by one point each. As street segments in the used network may be short, a 40 meters buffer was applied. Each segment is assigned an absolute value of accidents that happened since 2011. Accidents may be counted numerous times as segments are treated independently. The range of values was 1 – 35 for stations and 1 – 60 for all streets.

b) Speed Limit of Street: The speed limit of streets is one characteristic of the built environment, which has effects on bicycling (Sallis et al., 2013). Many cyclists prefer off street or physically separated paths (Winters & Teschke, 2010). The speed limit inside the city borders was provided by the City of Zurich, whereas in the other areas, OpenStreetMap (OSM) data was used. A spatial join was performed with the network of this study. Both this method and the use of OSM data is a source of impreciseness. The different speed limits were classified numerically as visible in Table 2. This was seen as appropriate, as there are already steps of speed limits, and these three steps were considered as the important ones.

Table 2:Implementation of maximal Speed on Street Values

Value Range of Speed Limit	Implementation
Unknown (NA)	0
≤ 30km/h	1
30km/h < x ≤ 50km/h	2
>50km/h	3

A limitation of this variable is that with the network of this thesis, it does not matter if there is a separated bike lane, as all segments are treated the same (see Section 4.1). All classes from 0-3 feature both in stations and all street segments.

Table 3: Geographic Variables included in this Thesis

Category		Variable	Source	Unit	Granularity	Reference
Safety and Design	a)	Exposure to accidents	KTZH[2] (OGD)	Number of accidents	Street Segment with 40m Buffer	Nelson et al. (2021)
	b)	Speed limit of street	City[3] (OGD), OSM	1 Category	Street	Sallis et al. (2013)
Land Use	c)	Distance to green space	BFS Arealstatistik[4] (OGD)	Distance in [m]	Street Segment (euclidean distance to nearest ha-raster)	Roy et al. (2019) Sallis et al. (2013)
	d)	Distance to residential area	BFS Arealstatistik (OGD)	Distance in [m]	Street Segment (euclidean distance to nearest ha-raster)	Roy et al. (2019) Sallis et al. (2013)
	e)	Distance to POI	KT ZH	Distance in [m]	Street Segment Assignment of midpoint to cell	
	f)	Mixed land use in area	BFS Arealstatistik (OGD)	1 Category	Street Segment Assignment of midpoint to cell	Winters et al. (2010) Saelens et al. (2003)
Demographics	g)	Population density	BFS[5] statpop (OGD)	1 Person	500x500m Cell Grid resample from hectares Assignment of midpoint to cell	Nehme et al. (2016) Winters et al. (2010)
	h)	% of 80+ years old inhabitants	BFS statpop (OGD)	1% of inhabitants	500x500m Cell Grid resample from hectares Assignment of midpoint to cell	Sallis et al. (2013) Nehme et al. (2016)
Socio-economic	i)	Swiss Neighbourhood Index	BFS	1 Decile	500x500m Cell Grid Assignment of midpoint to cell	Nelson et al. (2021) Roy et al. (2019)
Topography	j)	Slope of the street	Swisstopo[6] (OGD)	1 Category	Street Segment split every 120m. Mean of slopes per Segment.	Winters et al. (2010) Broach et al. (2012) Hood et al. (2013)

[2] Kanton Zürich. Available at https://opendata.swiss/dataset/polizeilich-registrierte-verkehrsunfalle-im-kanton-zurich-seit-2011

[3] City of Zurich. Available at https://data.stadt-zuerich.ch/dataset/geo_signalisierte_geschwindigkeiten

[4] Bundesamt für Statistik. Available at https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/geostat/geodaten-bundesstatistik/boden-nutzung-bedeckung-eignung/arealstatistik-schweiz.assetdetail.25885691.html

[5] Bundesamt für Statistik. Available at https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/erhebungen/statpop.html

[6] Swisstopo. Available at https://www.swisstopo.admin.ch/de/geodata/height/alti3d.html

All accessed on the 16/08/2023

c) and d) Distance to green space or residential area: Both variables were found to be significant covariates by Roy et al. (2019). There is also evidence that proximity to green space and residential areas correlate positively with the choice of active transportation as mode of transport (Sallis et al., 2012). Both variables were computed as shortest distance to the nearest raster cell of interest. Of the Arealstatistik data set, for green space, AS_18_27 = 10 was used and for residential area AS_18_27 = 2. The variables were implemented by taking the shortest absolute distance in meters of the street segment to the closest POI of the named dataset. For the stations data set, the range was 0m – 653m for green space and 0m – 467m for residential area. In all streets it was 0m – 1486m for green space and 0m – 1202m for residential area.

e) Distance to other POI: Points of Interest (POI) are a land use class that induces traffic. For this reason, the Office of Mobility and Transport of the Canton of Zurich maintains a data set of facilities that greatly induce traffic, including shopping centres, concert halls or large schools. The original name of the data set is «stark verkehrserzeugende Einrichtungen». The variable was implemented by taking the absolute distance in meters from the midpoint of any street segment to the closest POI of the named dataset. The range was 105m – 2628m for stations and 15m – 4370m for all street segments.

f) Mixed Land Use in Area: Mixed land use is positively associated with the choice of active transportation (Saelens et al., 2003; Winters et al., 2010). Yang et al. (2019) on the other hand found weak correlations of bicycling to mixed land use. To assess it in this study, mixed land use was implemented using the Arealstatistik (AS_18_27) classes 2 (residential), 3 (public buildings) and 5 (not specified buildings, among others commercial area). For each cell, the number of different classes (0-3) in a 500m*500m neighbourhood was calculated using QGIS, which leads to a range of values from 0-3. Each street segment belongs to the cell of its midpoint.

g) Population Density: More people living on the same space is associated with more bicycle traffic, as Nehme et al. (2016) and Winters et al. (2010) have confirmed. Both studies focused on bicycling as mode of transport and not in a leisure context. In this study, hectare cells of the Arealstatistik have been resampled to 500x500m cells. For each cell, the total population was then divided through 2.5km^2. Each street segment belongs to the cell where its midpoint lies. The values are the absolute numbers of people living in those 2.5km^2 and ranged from 3 – 5941 for stations and 0 – 5941 if all street segments are considered.

h) Percentage of Population older than 80 years: Conceptually, a common concept and goal is to make bicycling a safe mean of transport for everyone from 8 to 80 years old. People older than 80 years are generally not considered, as they are often not fit and/or comfortable enough anymore to ride a bicycle. While related studies by Nelson et al. (2021) and Roy et al. (2019) considered median age and percentage of veterans, this study refers to this indicator to consider age. As for population density, the hectares were resampled to 500x500m cells. Each street segment belongs to the cell where its midpoint lies. The values are the percentage of people older than 80 years in those 2.5km^2 and ranged from 0 – 18.75% for stations and 0% - 75% for all street segments.

i) Swiss Neighbourhood Index (SNI): As socio-economic measure, the Swiss Neighbourhood Index (Panczak et al., 2023) was implemented. The SNI is provided by the BFS and combines income, all-cause mortality and housing parameters. For this thesis, the version SwissSEP 3 (decile values) was used. The inclusion of this variable is inspired by Roy et al (2019) and Nelson et al. (2021), as median household income was one of the most significant covariates in both studies. Therefore, the significance of this variable can be compared to the culturally different societies in the US. The SNI provides values for all single residential buildings in Switzerland, whereas a neighbourhood is always formed with the nearest 50 households. Here, the values were aggregated to a 500x500m raster. For each cell, the median SNI decile value was calculated and used. Due to the use of decile values, the range of values was 0 – 10 for stations and all streets alike.

j) Slope: Slope can be an important factor for riders to choose their route, while most riders tend to avoid steep slopes (Broach et al., 2012; Hood et al., 2013). However there are also studies where such correlations remained weak (Yang et al., 2019). As bicycling is a kind of active transportation, slope is a common factor to include. In context of this study using Strava data, the anticipated effect may not stand equally, as the leisure-oriented users of Strava possibly do not fear steep slopes - the contrary could even be true. The slope variable was included by using the qprof[7] plugin in QGIS. For each 120m of a given street segment, a slope was calculated. For each segment, the mean of the slopes was then averaged. For the use in regression models, the slope values were then categorised numerically from 0-3 as follows:

Table 4: Implementation of Slope Values

Value Range of Slope [°]	Implementation
0	0
0.1 - 2	1
2.1 - 4	2
>4	3

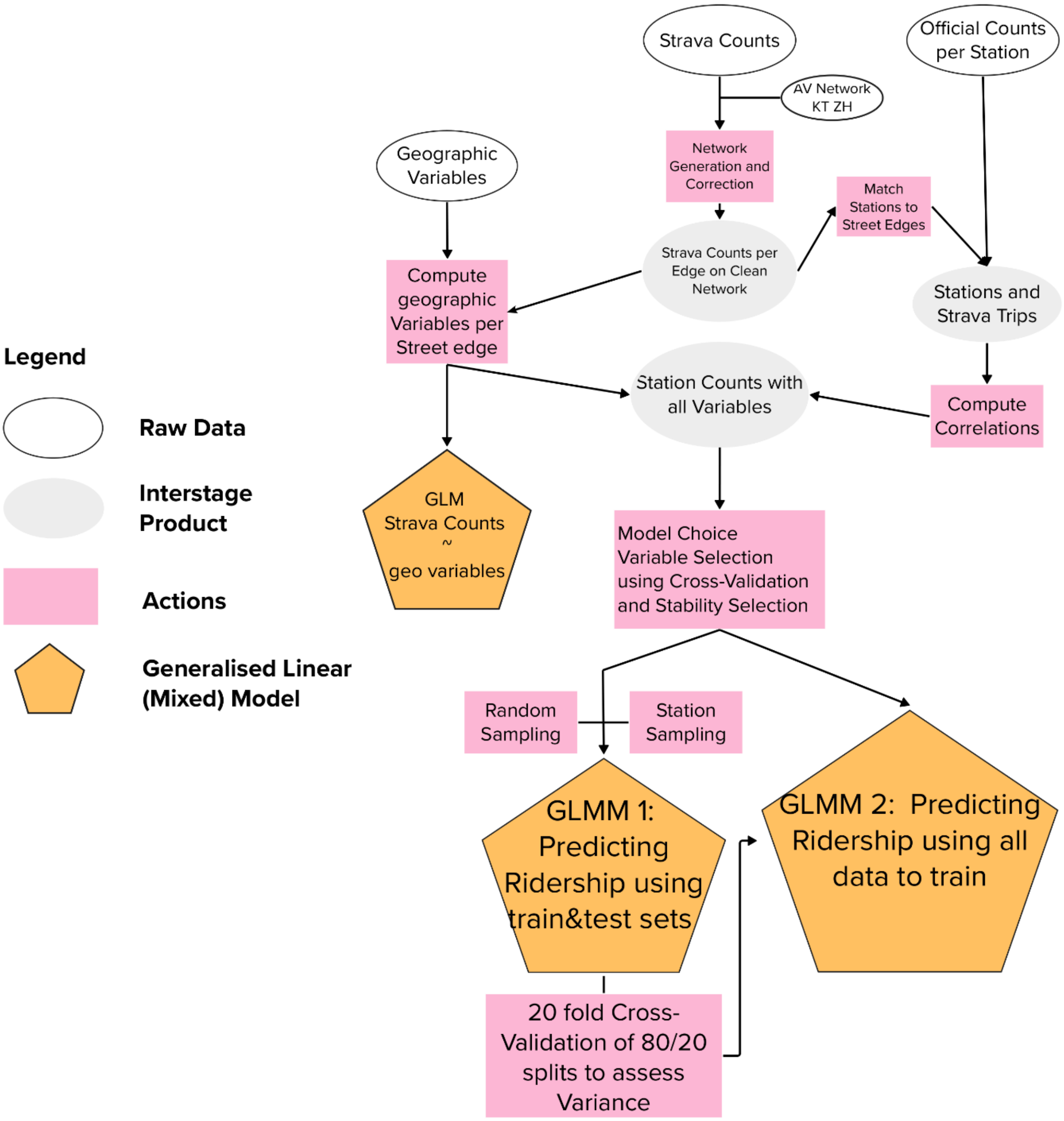
All classes appear in the stations and subsequently in all streets, however only a station that is located on a bridge has class 3, so the slope is that steep only for a very short distance.

The reason for the classification of slopes is that small differences might not affect decisions by riders. Following Broach et al. (2012), slopes were classified in two-degree steps. In that study, the route choice of riders was significantly affected by slopes over 2°.

[7] Available at https://github.com/mauroalberti/qProf (Accessed on the 16/08/2023)

4 Methods

From aggregating bicycle counts to predicting ridership, a variety of steps were necessary to tackle challenges and to arrive at the goals of this study. In Figure 7, a flowchart of the methods of this thesis is provided. In the following sections, the composition of the street network and the use of generalised linear (mixed) models will be described more closely.



Strava Counts
Official Counts per Station
AV Network KT ZH
Network Generation and Correction
Geographic Variables
Match Stations to Street Edges
Strava Counts per Edge on Clean Network
Compute geographic Variables per Street edge
Stations and Strava Trips
Legend
Station Counts with all Variables
Compute Correlations
Raw Data
Interstage Product
GLM Strava Counts ~ geo variables
Model Choice Variable Selection using Cross-Validation and Stability Selection
Actions
Generalised Linear (Mixed) Model
Random Sampling
Station Sampling
GLMM 1: Predicting Ridership using train&test sets
GLMM 2: Predicting Ridership using all data to train
20 fold Cross-Validation of 80/20 splits to assess Variance

Figure 7: Workflow of Methods of this Thesis

For the whole analysis, R (R Core Team, 2023) was used if no other software is mentioned. The respective packages are noted in each step, used versions and references can be found in the bibliography.

As an additional overview, all output types are given in Table 5 with a short explanation.

Table 5: List of Outputs of this Thesis

Overview over Outputs of this Thesis			
Product Type	Description	Purpose	Where to find
Shapefile	Generalised Network with Strava Counts	Prerequisite, Output	Supplementary Material
Table	Model Summary of Strava Counts explained through Geographic Variables	Explore Data	Results 5.2.1
Plot	Proportion of Captured Strava Trips at Counting Stations	Understand and/or filter Counting Stations	Results 5.2.2
Plot	Correlation between Strava Trips and Station Trips at Counting Stations		Results 5.2.2
2 Tables	GLMM 1: Summary of cross-validated Results Random Sampling vs Station Sampling	RQ 1: Assess the different Sampling Types	Results 5.2.3
2 Plots	GLMM 1: Variation of cross-validated Results Random Sampling vs Station Sampling		Results 5.2.3
2 Plots	Prediction Accuracy for x % of Segments of average and best performing out-of-sample GLMM	RQ 1: Assess Out-of-sample performance	Results 5.2.3
2 Tables	Jaccard Index of average and best performing out-of-sample GLMMs		Results 5.2.3
Table	Model Summary of final GLMM 2 using all Stations as Training Data	RQ 1: Assess In-Sample performance	Results 5.2.3
3 Plots	In-Sample Accuracy of Final GLMM 2		
1 Plot	In-Sample Accuracy for x % of Segments of GLMM 2		
Map	Median in-sample Accuracy and median Trips per Station		
Table	Importance of Variables in GLMM 2	RQ 2: Assess Importance of Variables	Results 5.2.3
4 Maps	Average Daily Bicyclists April-October in Limmattal, City of Zurich, Glattal and the entire Study Area	Output using GLMM 2	Appendix D
4 Maps	Average Annual Daily Bicyclists in Limmattal, City of Zurich, Glattal and the entire Study Area		

4.1 Preparing a Suitable Network with Strava Counts

As described already in Subsection 3.2.1, the raw data obtained from Strava Metro is not yet a suitable network, as numerous edges might be given for each street and some of them are paths that do not even exist. So as a pre-processing step, the editing of a clear network, that has the desired characteristics, was necessary. Besides R, QGIS was used to visualise the network and for the correction process described in Section 4.1.2. Figures 8 and 9 show the Strava counts in Zurich, Universitätsstrasse, before and after the processes described in this Section.

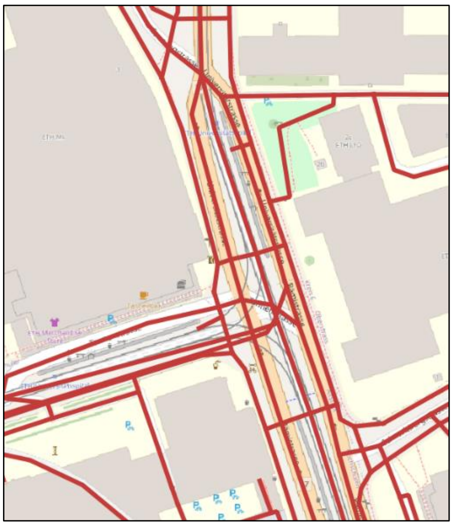


Figure 8: Strava Edges with Counts before Editing (own Screenshot QGIS)

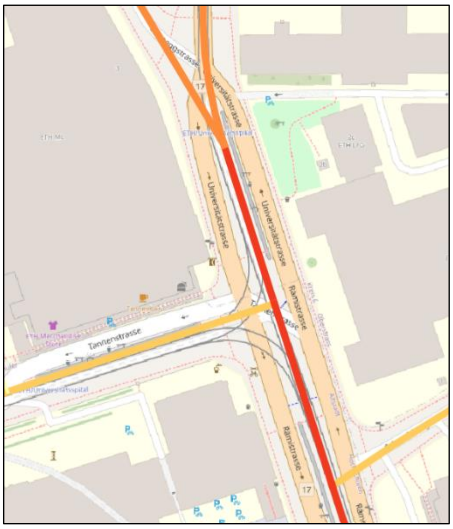


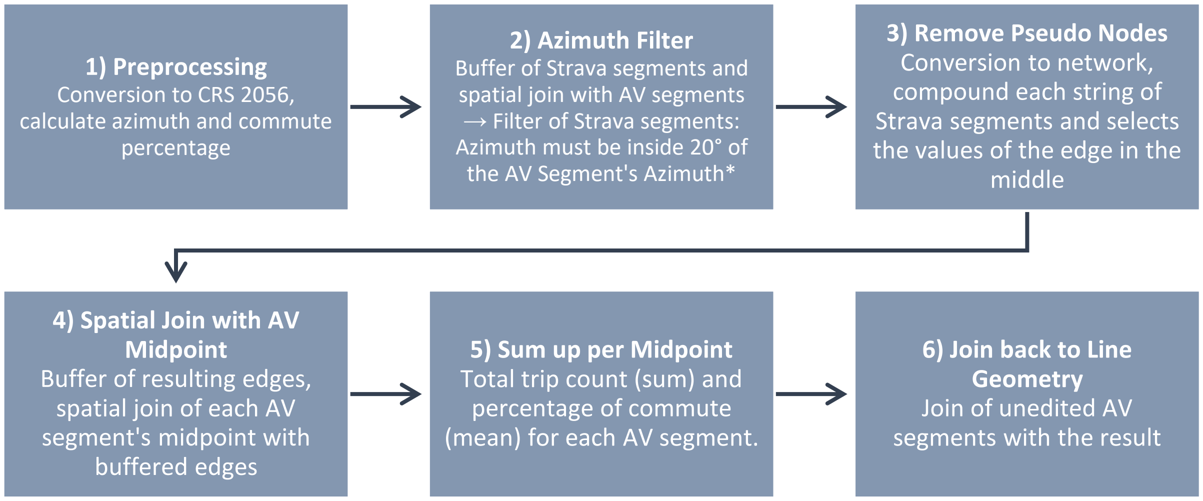
Figure 9: Strava Counts on Edges of the AV Network after Editing (own Screenshot QGIS)

4.1.1 Network Matching

The final product should be a simple network consisting of undirected street edges, as in Figure 9. Different networks were considered for the base for the final product: The swissTLM^{3d} vector data by swisstopo[8], the official measurement (Amtliche Vermessung, AV) network by the Canton of Zurich and OSM. Even though the Strava counts were already mapped on OSM features, this option was discarded first. OpenStreetMap is often not topologically correct and the coverage is very inconsistent, which could lead to issues. Between the TLM and the AV network, a closer comparison was done, where the differences between the two networks were analysed: The TLM network is more detailed than the AV network. For wider streets, it often gives more than just one edge. As the final product is

[8] Available at https://www.swisstopo.admin.ch/de/geodata/landscape/tlm3d.html (Accessed on 16/08/2023)

aimed to be a simple, undirected network with one value per street, the AV network is more suitable and was selected.



1) Preprocessing
Conversion to CRS 2056, calculate azimuth and commute percentage

2) Azimuth Filter
Buffer of Strava segments and spatial join with AV segments → Filter of Strava segments: Azimuth must be inside 20° of the AV Segment's Azimuth*

3) Remove Pseudo Nodes
Conversion to network, compound each string of Strava segments and selects the values of the edge in the middle

4) Spatial Join with AV Midpoint
Buffer of resulting edges, spatial join of each AV segment's midpoint with buffered edges

5) Sum up per Midpoint
Total trip count (sum) and percentage of commute (mean) for each AV segment.

6) Join back to Line Geometry
Join of unedited AV segments with the result

Figure 10: Flowchart of Network Matching Process
** Inverted values were also considered: Δ Azimuth ±180° ±360° <= 20°*

The matching process is visible in Figure 10. A combination of filters, buffers and spatial joins was used to map the Strava counts to the AV network as precise as possible.

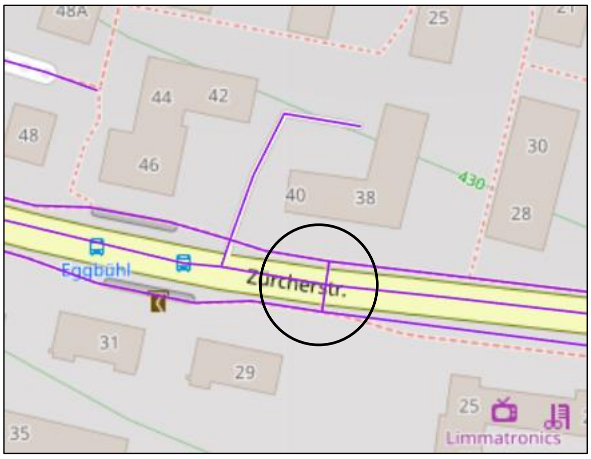


Figure 11: Example of Strava Edges (own screenshot RStudio)

Step 2 and 3 were necessary because of the already described multiple geometries of the Strava edges – there are several parallel lines that represent one street (see Figure 8 or 11 in purple). At times, these lines are then connected between each other as well, as this is the case inside the black circle in Figure 11. This leads to so called pseudo-nodes, which have only two incident edges. With the azimuth filter of Step 2), the vertical connections between the parallel lines are filtered out, so that only the parallel Strava edges remain. Step 3) makes sure that all pseudo nodes are filtered out using the to_spatial_smooth function of the sfnetworks package, so that all edges are as long as possible. In Step 4), the resulting Strava edges are buffered with a 12m rectangular buffer.

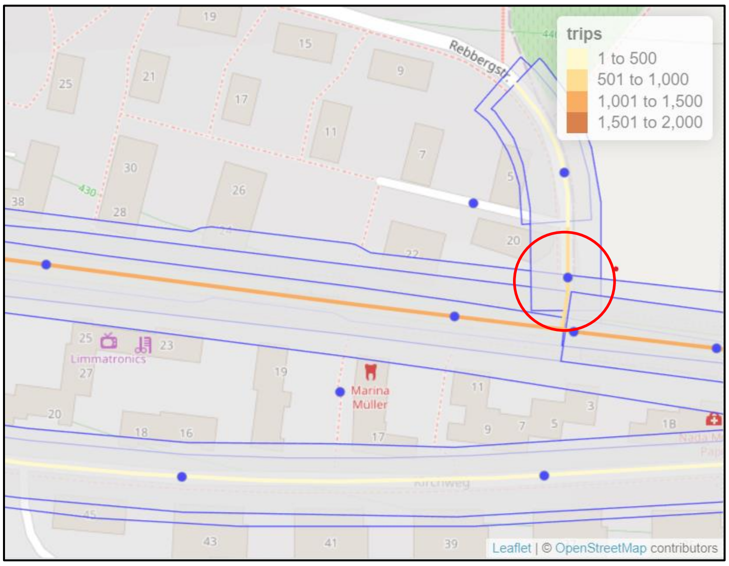


Figure 12: Example of Buffers, Midpoints and Final Network (own screenshot, RStudio)

This should make sure that all edges of a street are captured in the spatial join with the AV segment’s midpoints. In Figure 12, it is visible that the grey polygons with a blue border overlap. At the location of the midpoints (in blue), the trips from all polygons are added which corresponds to Step 5). In Step 6), the attributes of the points are rejoined to the line geometries, so that the Strava trip counts can be plotted as in Figure 9 or 12.

This resulted in a data frame containing all edges where at least for one month, Strava counts were at the minimum of 3-5 people (see Section 3.2.1., Strava Counts). One case, where the accuracy of this method is strongly limited, is visible inside the red circle of Figure 12: Whenever a smaller street at an intersection has a short AV segment, the midpoint may lies inside a polygon of the bigger street. Therefore, the trips are added, and the segment gets an inflated trip count. For this reason, the following Section describes how such cases were corrected.

4.1.2 Network Correction

In a final step, manual work has been carried out to correct such mismatching and delete unwanted features such as highways and bike paths in forests, where only leisure activity is to be expected. Despite the extensive automatic filtering, manual editing was needed as it is mostly the case for network matchings. Following the methodology noted in Figure 10, an efficient approach for manual editing was needed that ensured reproducibility. In Figure 13, a workflow of the manual editing is visible. In Step (1), a new column was created to store reference object ids. Per default, the object id of each edge is copied to this column. The network was then visually searched and sampled for unrealistic leaps in trip counts. If an error has been spotted, the immediate surroundings of the edge (whenever possible) were searched for fitting values. Hereby, the web tool of Strava Metro was used to check the true Strava counts. If a correct value is found, the object id of the edge with the right count is inserted in the reference object id column (Step (2)). This was the extensive manual part of the correction. In Step (3), a loop assigned the corrected trip counts and commute percentage to every edge, using the values of the reference edge.

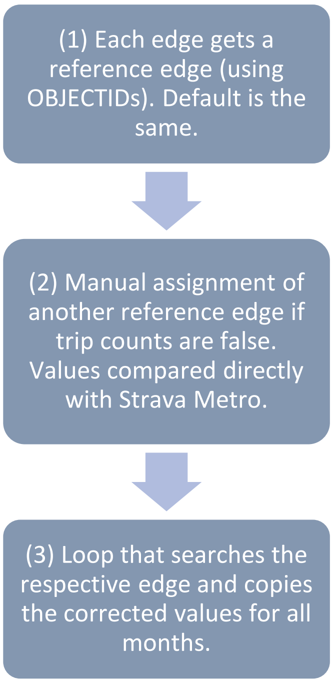

(1) Each edge gets a reference edge (using OBJECTIDs). Default is the same.
(2) Manual assignment of another reference edge if trip counts are false. Values compared directly with Strava Metro.
(3) Loop that searches the respective edge and copies the corrected values for all months.

Figure 13: Flowchart of 3 Steps in the Network Correction Process

As for the commute percentage, the decision was taken to inherit also the commute percentage of the assigned reference edge. Even though this is an obvious source of error, it was considered more logical than keeping the existent values, as in many corrected edges, trip counts were wrongly multiplied or summed up from crossing edges, which seemed as a larger and more uncertain source of error.

For editing, the month of September 2022 was chosen to be determining. That means that, if an edge for some reason (e.g. no Strava counts) did not exist in September 2022, it will not be in the network. Moreover, the correction and assignment of reference edges is based on September 2022 values. The reason for this choice is that, in transport planning, September is widely used as a reference month. For bicycling, it is not equally common, however it can be argued that

September represents both Summer and colder temperatures to a good extent. Additionally, there are no large holidays in this month, which fits well in the aim of this work to not focus on leisure activities.

In Section 5.1, results that emerged in the editing process are given. Certainly, this method of editing has its pitfalls, that are again assessed in Section 6.3. Nevertheless, for most corrections, the method is accurate, because the right values were in an edge nearby. As already mentioned, repetitive need for correction was needed in many intersections, where Strava counts of crossing streets were wrongly assigned. Another common error were network edges of squares that lay close enough to streets and thus were assigned with values (see Figure 14). These were deleted, respectively assigned with value 0.

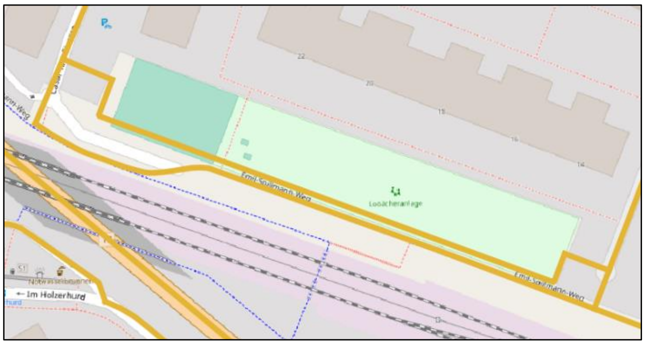


Figure 14: Wrongly assigned Edge Value at a Square in Affoltern (own Screenshot QGIS)

The full methodology including code can be found on Github (see Appendix A).

4.2 Fitting Generalised Linear (mixed) Models

Model Choice: To arrive at the goals of this study, generalised linear models are used to predict bicycle counts. Having counts as response variable, a Poisson or negative binomial model should be fitted. For models using the Poisson distribution var[y] = μ is assumed – if var[y] is significantly greater than μ, while the most explanatory model is used, overdispersion is present. Alternatively, if the residual deviance of a Poisson model is much larger than the residual degrees of freedom, it is another sign of overdispersion. One possibility to deal with overdispersion is the use of negative binomial models (Dunn & Smyth, 2018).

A general assumption of generalised linear models (GLMs) is the independency of the observations. This prerequisite is not met, as this study deals with repeated measurements over time. For this reason, the ridership prediction (Chapter 4.2.2) was computed by fitting a Generalised Linear Mixed Model (GLMM). GLMMs include random effects that account for the variability inside a cluster of repeated measurements(Breslow & Clayton, 1993; Dean & Nielsen, 2007).

Transformation of Variables: Mathematical models like GLMs must be fed by numeric inputs. The coding of variables can thus be an important step of the study design. As justified in Subsection 3.2.2, the variables speed limit on street and slope were classified from 0-3. The variable mixed land use has also classes from 0-3, as three is the maximal value of mixture of a given raster cell. For seasons (aggregated from months) and years, a coding for categorical variables had to be implemented. A common method is called one-hot encoding or dummy variable encoding. A variable with k categories is converted to a factor. Then, for a factor with *k* levels, commonly *k - 1* binary "dummy" variables are created. The drop of one level is done to avoid multicollinearity, as the absence (i.e. a 0) for k-1 levels already expresses the existence (i.e. 1) of the dropped level (Dunn & Smyth, 2018, p. 10). For interpreting outputs of GLMs, the estimates of all levels are relative to their reference level, which is the dropped level.

Another transformation was applied to the Strava Counts. In the main models, the Strava counts served as the most important predictor. After Dunn&Smyth (2018, p.121), transformations can be applied to any or all covariates. As interpretability degrades, it was decided to only transform the Strava counts and no geographic variable, so that RQ2 is not affected. The logarithmic transformation led to an improvement of the model fit and was therefore maintained. In Figure 15, the histograms of the Strava counts at stations can be seen – without transformation, the Strava counts are strongly right skewed, while the log-transformed counts are close to normally distributed.

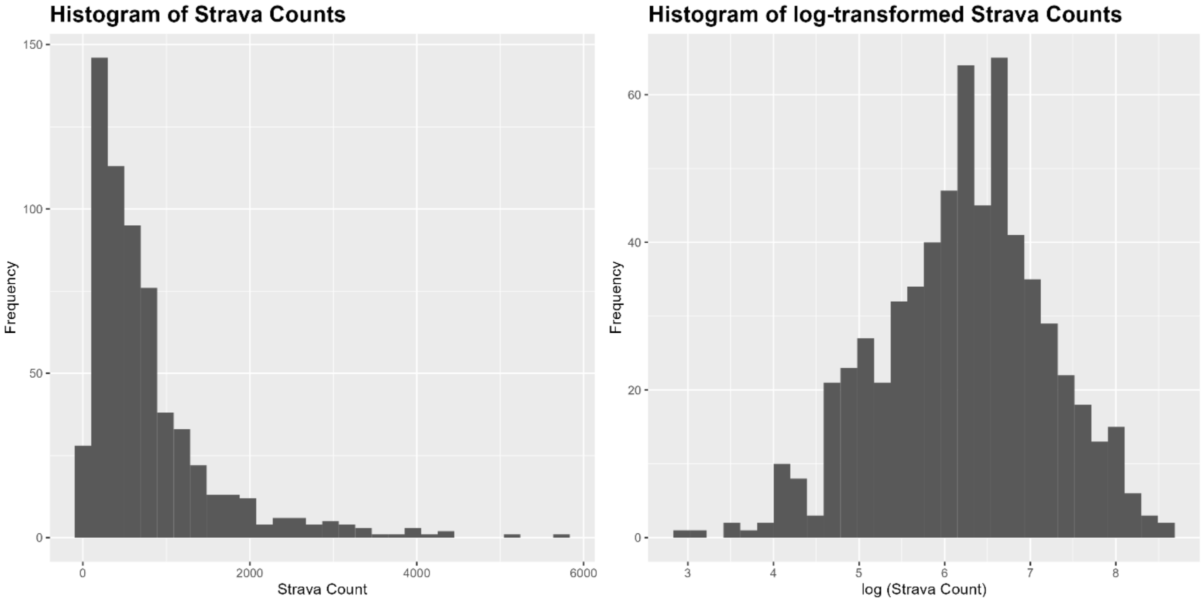


Figure 15: Histograms of Strava Counts and the Logarithm of Strava Counts

4.2.1 Explain Strava Counts using Geographical Covariates

The first model explored the relationship between the Strava counts and the geographical variables. The aim of this step was purely exploratory and has no effect on the further steps of processing towards the prediction.
The filtered Strava edge counts represent the dependent variable. The independent variables are all geographical variables, the commute percentage and the season and year of each count. For this negative binomial GLM, the function glm.nb of the MASS package was used.

In theory, there were also repeated measurements in this model – each street edge has 24 monthly values. However, the inclusion of the random effect resulted in a non-converging model. For this reason, a regular GLM with negative binomial distribution was fitted.

As a last preparation step, the Strava counts were filtered so that only edges that have more than 2000 trips in 24 months remain. This step was introduced after the first iterations of this model, as it had proved to boost the performance. Smaller counts are less robust and are often influenced by few users and vulnerable to effects of the binning procedure in Strava (see Section 3.2.1).

4.2.2 Main Models: Preparation of Ridership Prediction

For the ridership prediction, the data of the counting stations is the response variable which serves as ground truth. Following preparation steps and decisions were necessary to find the best possible model with the available data:

Correlation between Counting Stations and Strava: The analysis of the relationship between the counting stations and the crowdsourced data of Strava was the first step towards a descriptive model of ridership. First, the proportion of captured trips by Strava was plotted. Secondly, the Pearson correlation between the station trips and the Strava trips was examined. As the Pearson correlation assumes normally distributed variables, both the station and the Strava trips were log-transformed for the correlation analysis.

The results of the correlation analysis served to filter out some counting stations that are not useful for a predictive model. According to the results, 5 stations were filtered out.

At this point, the final dataset involves all stations and their available observation months (3-24) of official counts, the Strava count (available for all 24 months) and all other variables.

Sampling Strategies: Before the variable selection and the prediction could take place, the sampling strategy had to be defined. As a common figure, a decision for an 80% train and 20% test data set was taken. Following that, there are two variants to split the data in a train and test set:

1) Random Sampling: Take an 80% random sample of the whole dataset for the train set and the rest of observations for the test set. In the test set, there probably will not be stations that the model has not been trained on.
2) Station Sampling: Take 80% of stations, but all their available months for the train set and the remaining stations for the test set. The test set has completely unseen stations.

As both strategies give different insights, both were executed and described in the further process of prediction modelling. The random sampling led to figures for in-sample accuracy (even if it is not entirely in-sample), while the station sampling represents out-of-sample accuracy. In terms of related work to those strategies, Nelson et al. (2021) and Jestico et al. (2016) computed in-sample accuracies, as observations of the same locations were used in both the training and testing set. while Roy et al. (2019) determined out of sample accuracies.

Variable Selection using Cross-Validation and LASSO:

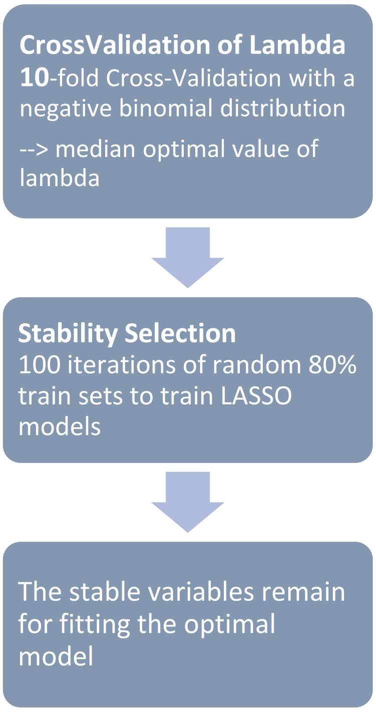

CrossValidation of Lambda
10-fold Cross-Validation with a negative binomial distribution
--> median optimal value of lambda
Stability Selection
100 iterations of random 80% train sets to train LASSO models
The stable variables remain for fitting the optimal model

Figure 16: Flowchart of Variable Selection Process

A wide range of variables was computed to contribute to the prediction models. However, some variables might be more predictive than others, and the inclusion of not correlated variables could even worsen the models. Hence, as in related studies by Roy et al. (2019) and Nelson et al. (2021), a variable selection process was conducted. For both sampling methods, the same procedure was applied (see Figure 16). The cross-validation was carried out using the cv.glmregNB function of the mpath package. This function computes the optimal lambda which maximizes the log-likelihood. In other words, the lambda which provides the model with the best fit for the data. In the whole variable selection process, the random effect was discarded for simplicity, also because, to the author's knowledge, no existing R function combines a negative binomial cross-validation with random effects.

In Step 2, 100 iterations of fitting a LASSO (Least Absolute Shrinkage and Selection Operator) model were executed using the glmregNB function of the same package. This is an assessment of stability of the model using our data. Stability selection methods are commonly used to assess the performance of models like the LASSO (e.g. Meinshausen & Bühlmann, 2010). As lambda value, the median lambda of the 10-fold cross-validation was used. The LASSO is, as the name reveals, a selection operator which provide coefficients of significance to keep the most predictive variables in the model, while discarding others. It shrinks less important variables to zero by imposing a penalty on the absolute magnitude of regression coefficients (Tibshirani, 1996). In each iteration, a variable that did not emerge as beneficent for the model was shrunk to zero. The stability of the variables was then assessed based on the performance over 100 iterations to arrive at the final set of variables for both sampling methods.

A further measure to get the best possible model was the check for multicollinearity between the selected predictors. Together with model summaries, the Variance Inflation Factor (VIF) was computed, using the "multicollinearity" function of the performance package.

4.2.3 Prediction of Ridership using GLMMs

The final set of variables could now be fitted into a GLMM for both sampling methods. For the final models, the glmmTMB function was used. The random effect is implemented by the term 1 | OBJECTID (see Equation 1). The OBJECTID refers to the unique identifier of the network. As every OBJECTID has only one corresponding station, it could be used to account for the repeated measurements.

Equation 1: Final GLMM

$$\begin{aligned} glmmTMB(trips\_station \sim trips\_strava\ +\ crr\_cmm\ +\ accidents +\ vmax \\ +\ dist\_10\_green\ +\ dist\_2\_resid\ +\ dist\_sve\ +\ mix\_value \\ +\ PopDens\ +\ u80\_perc + swiss\_sep\_D\ +\ slope\ +\ season\_winter \\ +\ (1|OBJECTID), data\ =\ train\_data\_v2, family\ =\ "nbinom2") \end{aligned}$$

The family argument "nbinom2" is one of the two implementations of the negative binomial distribution provided by the glmmTMB package. It assumes a quadratic relationship between the mean and the variance of the negative binomial distribution, with variance equal to μ(1 + μ/ϕ) (Brooks et al., 2017). nbinom1 would assume a linear relationship with variance equal to μ(1 + ϕ). As nbinom2 performed better especially for out of sample predictions, it was selected for the final models.

GLMM 1 Predicting Ridership using train and test sets

To evaluate the variation of results for both sampling strategies, a 20-fold cross validation was performed, using an 80-20 train-test set partition. For assessment in each iteration, the AIC, R-Squared of all effects and R-Squared of the fixed effects were computed. The latter two using the performance package and according to Nagakawa et al. (2017), who formulated R-squared definitions for

GLMMs. Additionally, to quantify the error rates in the predictions following measures were computed:

Table 6: Error Measures used for Prediction Accuracies

Error Measure	Formula
Mean Absolute Error (MAE):	mean(abs(actual - predicted))
Root Mean Squared Error (RMSE):	sqrt(mean((actual - predicted)^2))
Mean Absolute Percentage Error (MAPE):	mean(abs((actual - predicted) / actual)) * 100

For each measure, the mean and the standard deviation of all iterations were calculated, and the different measures were plotted to visualize the variation of accuracy for both the random sampling and the station sampling method.

For the further visualisation of out-of sample performance, two models of the 20 iterations were selected. One average-performing and one of the best-performing models, determined principally by the MAPE and the AIC.

For each model, plots showing the proportion of segments predicted with a certain accuracy, the prediction accuracy per station of the test set and the Jaccard Index for categorical maps as output were created. The Jaccard Index assesses the accuracy of classifications and helps to decide if a categorical map is a reasonable output. For all categorical maps as output, breaks for 5 classes ranging from very low to very high were created using the getJenksBreaks function of the BammTools package. To compute the breaks, the medians of station trips from all stations was used, to account for the different number of observations between stations. From initially 7 classes, the classes 5-7 were aggregated to one class for very high values, following the distribution of the station trips' median. The Jaccard Index for each class following (Labatut & Cherifi, 2012) was then calculated as follows in Equation 2:

Equation 2: Jaccard Index of Classification

$$Jaccard\ Index = \frac{Intersection}{Union} = \frac{TP}{TP+FP+FN}$$

TP=	True positive
FP=	False positive
FN=	False negative

GLMM 2: Predicting Ridership using the whole Data as Training Set

Finally, to arrive at the most expressive model, all stations could be used as training data set. For this final model, only in-sample accuracies can be calculated. The importance of geographic variables was also assessed using this model.

Categorical maps have been defined as the most reliable output by both Roy et al. (2019) and Nelson et al. (2021). Therefore, a total of 8 categorical maps were produced using ArcGisPro: Besides maps showing the whole perimeter, separate maps were produced for Limmattal/Furttal, City of Zurich and Glattal. As time units, following the goals of this thesis, for each map two version were created: One shows the AADB and the other the average daily bicyclists from April-October. For that, the monthly values were divided by 30 and averaged over 7 respectively 12 months. For each map and region separately, natural breaks (Jenks) were used to divide the values in five classes.

5 Results

5.1 Network

The unedited AV network was roughly clipped to the perimeter of this study. This step led to 31945 edges remaining. Using the network matching for September 2022, 15460 edges with assigned and summed Strava counts remained. This number represents the maximum of possible edges for each month. However, in colder months there may be a lot less than that, for example, January 2021 records 12135 edges with counts. The Strava counts were finally filtered once more – all edges with less than 2000 trips over the 24 months were left out. Due to the distribution visible in Figure 17, this led to the loss of another almost 50%, resulting in 7760 edges.

A total of 699 edges were corrected manually and assigned to the counts of another edge. This accounts for 4.5% of all edges. The corrected edges are rather high-count connections: In Figure 17, the distribution of counts (median of monthly values per edge) and the medians of corrected and uncorrected edges is visible. It shows that the median of corrected edges (302.5) is three times higher than the median of uncorrected street segments (100). Table 7 shows an overview of the counts of the network.

Table 7: Overview Network Counts

Network: Overview	
Total Edges in perimeter	31945
Total edges with counts (09/22)	15460
Total edges with >2000 trips in 24 months	7760
Manually corrected edges	699

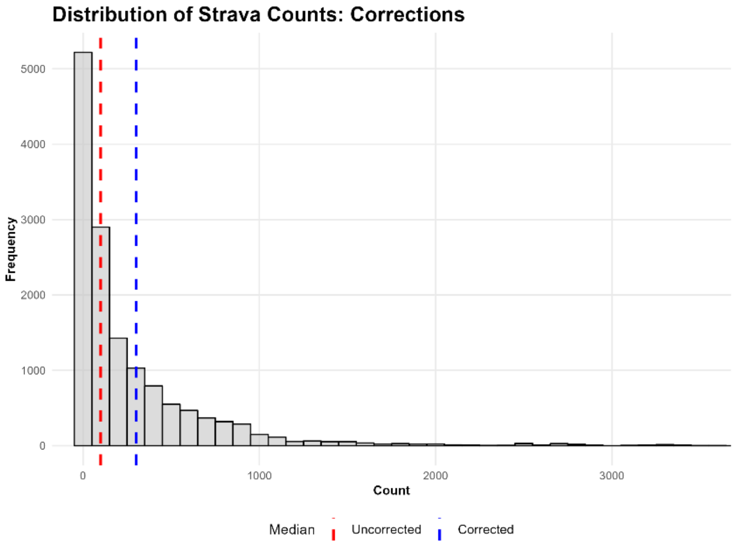


Figure 17: Distribution of Corrected Edges

5.2 Regression

5.2.1 Explain Strava Counts using Geographical Covariates

The GLM to explore the relationship between the Strava Counts and the geographic variables resulted in the following output in Table 8:

Table 8: Output GLM Strava Counts explained through Geographic Variables

Variable	Estimate	Std. Error	p-value
Intercept	5.291	1.045e-02	< 2e-16 ***
Strava % of commute	-0.001214	1.403e-04	< 2e-16 ***
Exposure to accidents	0.05715	3.879e-04	< 2e-16 ***
Speed limit on street	0.2412	2.119e-03	< 2e-16 ***
Distance to green space	-0.0002587	1.045e-05	< 2e-16 ***
Distance to residential area	0.0003888	1.619e-05	< 2e-16 ***
Distance to POI	0.0001453	2.802e-06	< 2e-16 ***
Mixed land use	-0.02359	1.553e-03	< 2e-16 ***
Population density	-0.000005345	1.855e-06	0.00396 **
% of 80+ years old inhabitants	-0.004425	2.963e-04	< 2e-16 ***
Swiss Neighbourhood Index	0.02496	7.345e-04	< 2e-16 ***
Slope of street	-0.08422	2.278e-03	< 2e-16 ***
Season: Spring	0.293	5.329e-03	< 2e-16 ***
Season: Summer	0.544	5.238e-03	< 2e-16 ***
Season: Winter	-0.6883	5.325e-03	< 2e-16 ***
Year 2022	-0.003431	3.743e-03	0.35936
AIC=2588100	Nagelkerke's R^2= 0.538	Observations=186149	
p-value: *** <0.001	** < 0.01	* <0.05	. <0.1

Almost all predictors are clearly significant, Population Density a bit less and Year 2022 stands out as the only not significant input. In the estimates, there is a larger range of values. The speed limit and the season variables have the largest estimates, even when the units of the variables are kept in mind. For the season variables, this is not surprising, as they are, together with the commute percentage, the only dynamic variables. All other variables are constant over time. However, the estimates of the season variables and the year 2022 cannot be interpreted alike the others. As they are categorical variables implemented by one-hot encoding, their estimate is always relative to the reference level (see Section 4.2). At the bottom of the table, the AIC, a value for R-squared and the number of observations is provided.

5.2.2 Main Models: Preparation of Ridership Prediction

Proportion and Correlation between Counting Stations and Strava:

Figure 18 shows the proportion of bicycle trips captured by Strava at official counting stations. The plot shows the wide variety of stations in terms of station trips and the proportion of Strava trips captured. The proportion is shown in classes that were built using the mean and the standard deviation. The triangles represent stations by the Canton, the squares stations by the City of Zurich. It shows that the counting stations of the Canton have an overall higher proportion rate of Strava trips captured – the mean of Canton stations is 0.059, whereas the mean of the city station is 0.020, resulting in an overall mean of 0.041. Generally, the variables appear to be negatively correlated as the distribution resembles an L-shape, i.e. all stations with a high number of trips have a low proportion of Strava trips captured.

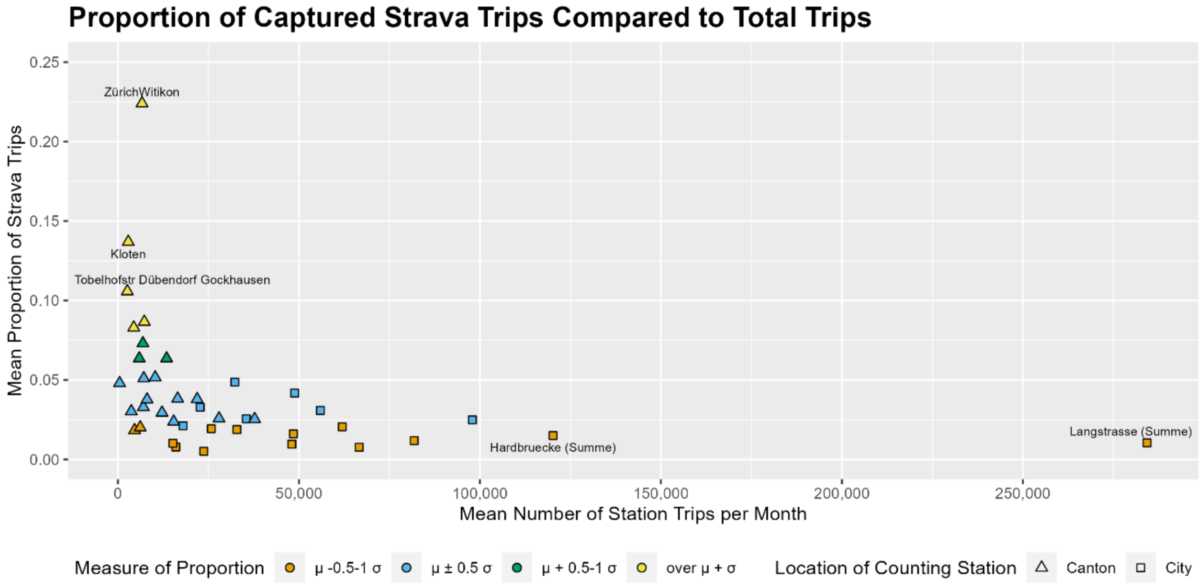


Figure 18: Proportion of Captured Strava Trips at Stations.

In Figure 19, the correlation between the counts and the Strava counts can be seen. On the x-axis, the 41 counting stations are named. The plot shows an overall very good correlation, with some stations as outliers. While most stations have a Pearson correlation of around 0.9, the outliers correlate worse. The station “Saumackerstrasse” even correlates negatively, which is the reason for the missing data point. The mean Pearson correlation is 0.86, the median 0.93 and the standard deviation 0.29. This further shows the existence of outliers, which strongly deviate the mean. Regarding the location of the stations, there is no trend visible, as both Canton and City authorities operate very well but also not strongly correlating stations. In Table 9, all discarded stations are visible - all stations that have a Pearson correlation of less or equal than 0.75 were discarded for the further analysis. The station "Glattuferweg, Opfikon" right at the threshold was discarded either way because of very small Strava counts. Finally, 36 counting stations with 679 months of measurement were used to predict ridership in the further steps of this thesis. On average, there are 18.86 months per station. The mean Pearson correlation at the used stations was 0.93.

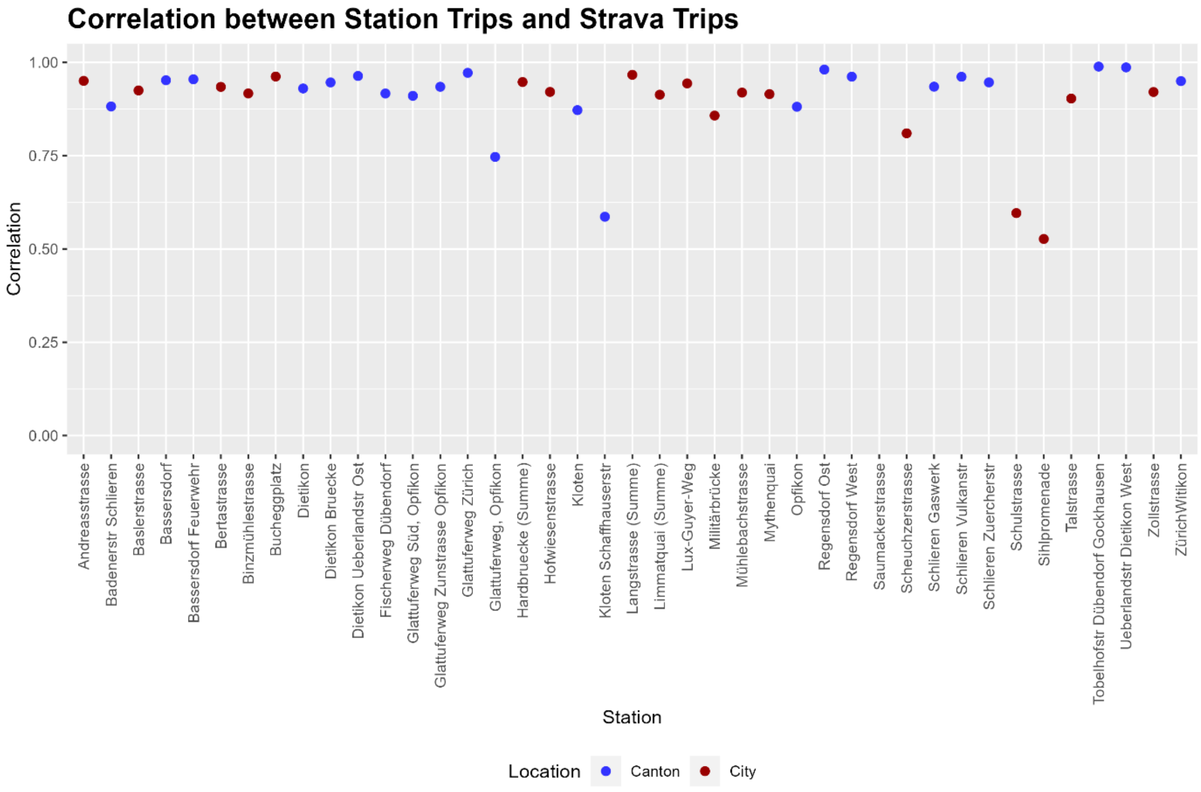


Figure 19: Correlation between Station Trips and Strava Trips at Stations

Table 9: Discarded Stations

Discarded Stations	
Station	Reason
Glattuferweg, Opfikon	Very low Strava Counts, moderate correlation (0.75)
Kloten, Schaffhauserstrasse	Low correlation (0.59)
Saumackerstrasse	Negative correlation (-0.86)
Schulstrasse	Low correlation (0.60)
Sihlpromenade	Low correlation (0.53)

Variable Selection using Cross-Validation and LASSO:

For both sampling methods, a 10-fold cross-validation was computed to obtain the optimal lambda for the respective LASSO. Table 10 shows that the optimal lambdas are very similar between the two sampling types.

Table 10: Optimal Lambdas for LASSO after Cross-Validation

Measure	Random Sampling	Station Sampling
SD	0.00920	0.01419
Mean	0.02267	0.02043
Median	0.02351	0.01720

Following that, the 100 iterations of the LASSO were executed for each sampling strategy, as a mean of stability selection. Table 11 shows for all variables the percentage of times the variable remained as useful for the model and was not shrunk to zero by the LASSO.

Results show that all variables are important in the majority of iterations. The random sampling has a bit more consistent results than the station sampling, which makes sense, as in the station sampling there can be more variations between iterations, when whole stations are in- or excluded.

Table 11: Stability Selection Results using LASSOs

Variable	Random Sampling	Station Sampling
Strava Trips	1.0	1.00
Strava % of commute	1.0	0.84
Exposure to accidents	1.0	1.00
Speed limit on street	1.0	0.98
Distance to green space	1.0	0.95
Distance to residential area	1.0	0.97
Distance to POI	1.0	1.00
Mixed land use	0.6	0.96
Population density	1.0	1.00
% of 80+ years old inhabitants	1.0	0.97
Swiss Neighbourhood Index	1.0	1.0
Slope of street	1.0	0.99
Season: Spring	0.8	0.57
Season: Summer	0.6	0.75
Season: Winter	1.0	1.00
Year 2022	0.58	0.81

Out of reasons of simplicity and easier comparison, it was decided to exclude the same variables for both sampling methods: Firstly, the year 2022 and secondly the season variables of spring and summer. Year 2022 has rather low stability values and was additionally not seen as very meaningful, as ridership is aimed to be predicted generally and not for a certain year (which represents certain conditions including weather). The season variables of Spring and Summer did not perform that well either. Nelson et al. (2021) have also only considered "Counts of Winter months", so with that stability scores, it was seen as best to exclude both Spring and Summer.

The test for multicollinearity in the GLMMs resulted in low correlation between the variables. No VIF above 4 was recorded.

After these steps, the final variables for the model predicting bicycle ridership have been selected and are shown in Table 12.

Table 12: Final Set of Variables for GLMMs

Category	Variable	Type
Ridership	Official Counts at 34 locations	Response; dependent variable
Crowdsourced Ridership	Strava Trip Count	Predictor; independent variable
	Strava % of Commute	Predictor; independent variable
	Count Collected in Winter	Predictor; independent variable (...)

Safety and Design	Exposure to accidents	Predictor; independent variable
	Speed limit of street	Predictor; independent variable
Land Use	Distance to green space	Predictor; independent variable
	Distance to residential area	Predictor; independent variable
	Distance to POI	Predictor; independent variable
	Mixed land use in area	Predictor; independent variable
Demographics	Population density	Predictor; independent variable
	% of 80+ years old inhabitants	Predictor; independent variable
Socio-economic	Swiss Neighbourhood Index	Predictor; independent variable
Topography	Slope of the street	Predictor; independent variable

5.2.3 Prediction of Ridership using GLMMs

The final GLMMs were fitted using the variables above in Table 11 and the random effect. GLMM 1 used 80% train and 20% test sets of both sampling strategies, while GLMM 2 is the most expressive model using all stations for training.

GLMM 1: 20-fold Cross-Validation of both Sampling Strategies

In Tables 13 and 14, the summarised results of the respective 20-fold cross-validation are visible. Plots in Figures 20 and 21 show the variation of results graphically. Be aware that the scales of the figures are different. Single iterations cannot be compared between the two sampling methods as the subsets are different.

Table 13: GLMM 1 - Summary of 20-fold Cross-Validation using Random Sampling

Random Sampling						
Measure	MAE	RMSE	MAPE	AIC	R2_all	R2_fixed
SD	335	877	2.59	33.96	0.00	0.01
Mean	3928	7489	15.44	10506	0.98	0.85
Median	3959	7553	14.55	10499	0.98	0.85

Table 14: GLMM 1 - Summary of 20-fold Cross-Validation using Station Sampling

Station Sampling						
Measure	MAE	RMSE	MAPE	AIC	R2_all	R2_fixed
SD	17185	36473	49.06	261.51	0.00	0.02
Mean	19138	32416	72.49	10560	0.97	0.86
Median	13512	19728	56.76	10481	0.98	0.86

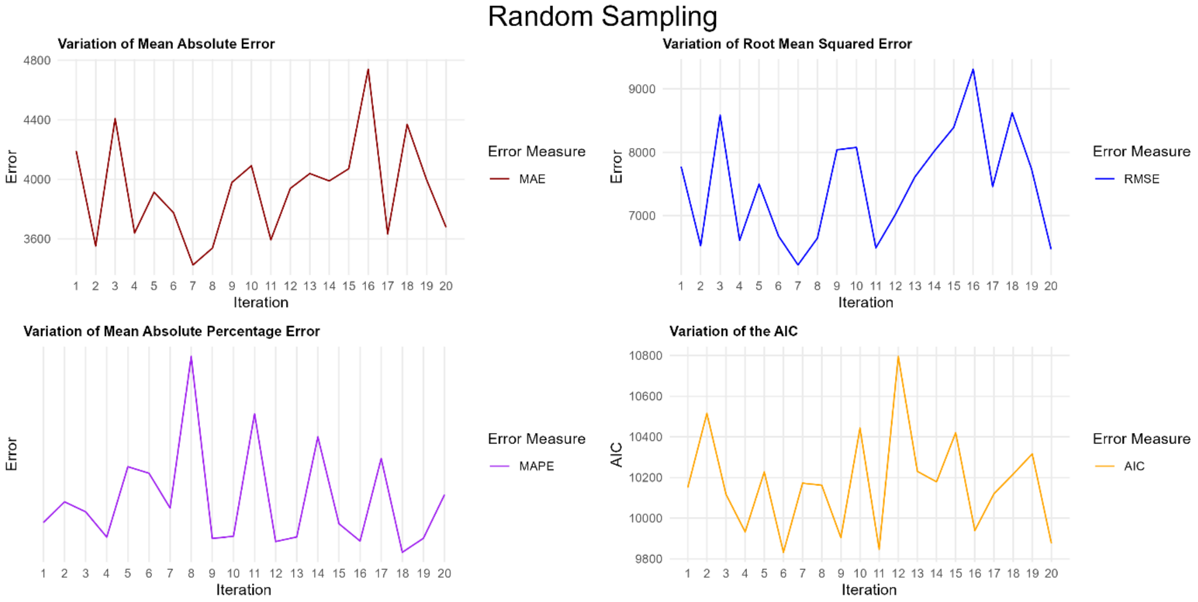


Figure 20: Plots of Variation of Error Measures in the Cross-Validation; Random Sampling GLMM 1

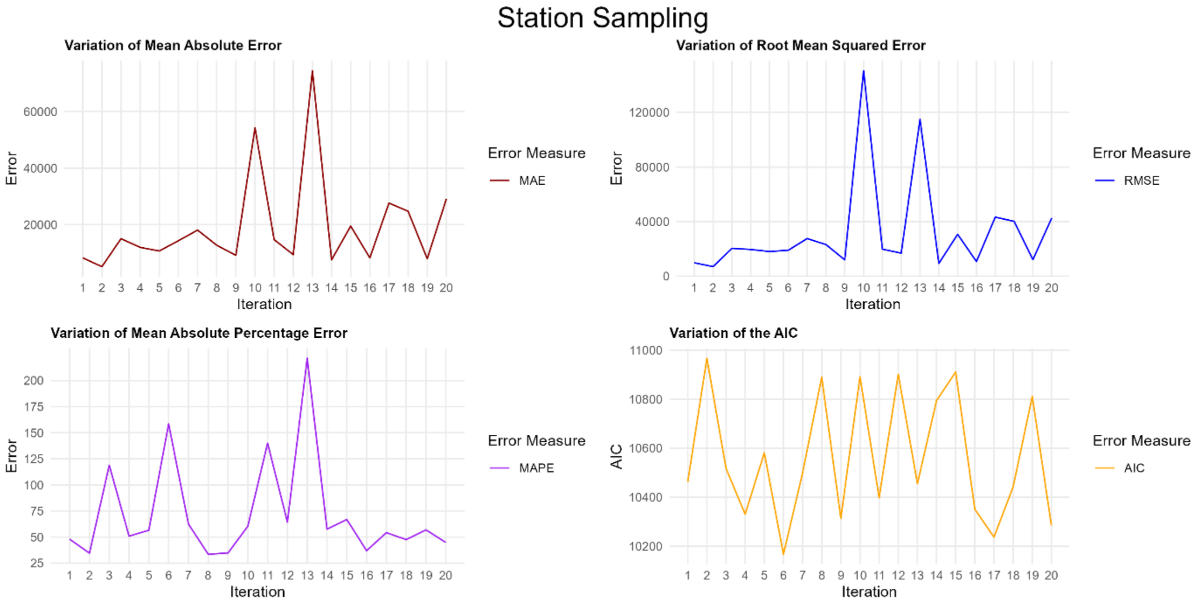


Figure 21: Plots of Variation of Error Measures in the Cross-Validation; Station Sampling GLMM 1

The results clearly differ between the two sampling strategies. For random sampling, the mean MAPE is 15.44% with a standard deviation of ±2.59%. For the station sampling, the mean MAPE is 72.49% with a standard deviation of ±49.06%, while the median MAPE is 56.76%. These values show the moderate but also very unstable results in the station sampling. The difference between mean and median further implies negative influence of outliers on the mean. Other error measures show a similar pattern, the station sampling has higher values and a much higher variability over these 20 folds. The R2 values are an exception to this, as they show stability for both sampling methods. This suggests that the model fit is very good and not dependent on the subset of data used.

The plots in Figures 20 and 21 visualise the variability of the cross-validation. For the station sampling that provides out-of-sample predictions, the results are very dependent on the iteration, as there is no stability in the predictions. Of the 20-folds, the average performing model that was selected is Model 7, while the best model is Model 9.

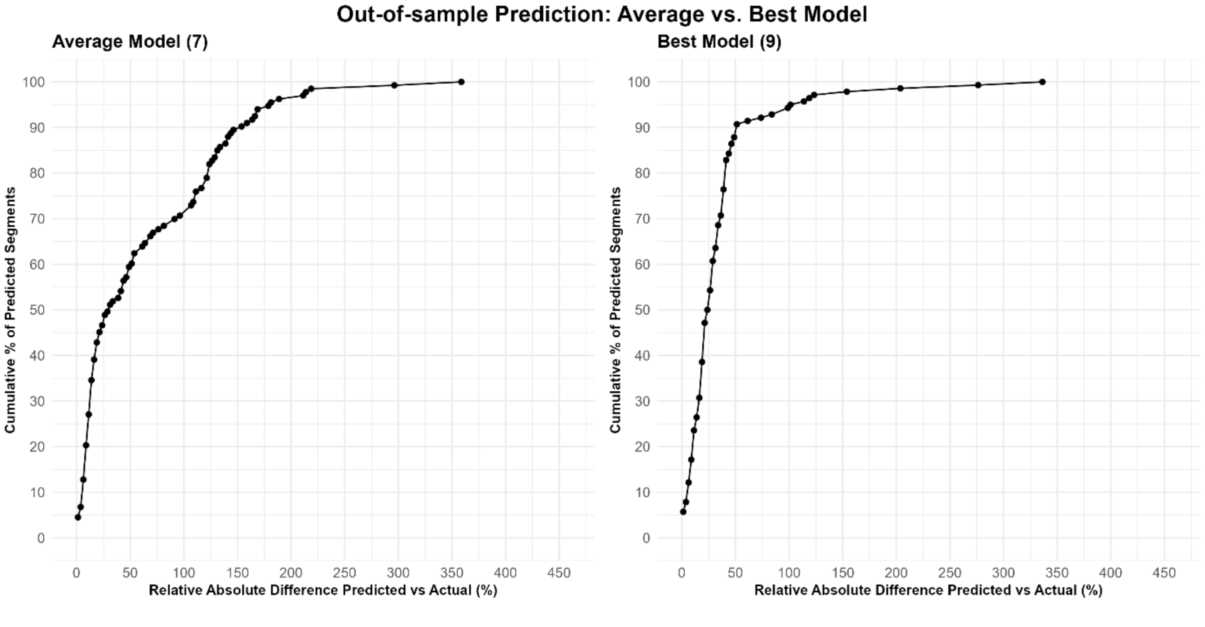


Figure 22: Out-of-Sample Prediction Accuracy proportionally to % of Segments
Observations have been classified into 2.5% classes of the relative absolute differences to plot.
Average Model: 133 observations in 56 classes
Best Model: 140 observations in 33 classes

At the left side in Figure 22, the relative prediction accuracy of the averaging performing out-of-sample prediction model is plotted against the percentage of segments where this accuracy is reached. This model, which corresponds to iteration 7 of the variation plots in Figure 21, shows that for 60% of segments, the monthly counts ±50% of riders are correctly predicted. Close to 45% of segments are predicted with ±25% of monthly riders.

In comparison to that, the best performing Model 9 is visible at the right in Figure 22. Compared to the average model, the best model predicts 90% of segments inside ±50% of riders and exactly 50% of segments inside ±25% of riders. The differences only really start for the upper half of segments on the y-axis. The number of observations and classes (see Figure 22 description) further shows that generally, the average model has less observations in more classes, showing a higher spread of accuracies. The different number of observations is due to the Station sampling (see Section 4.2.2).

Figure 23 shows the seven stations used to test the accuracy in each of the two selected out-of-sample models.: The average model has three stations that have very inaccurate predictions in general, the best model has only one.

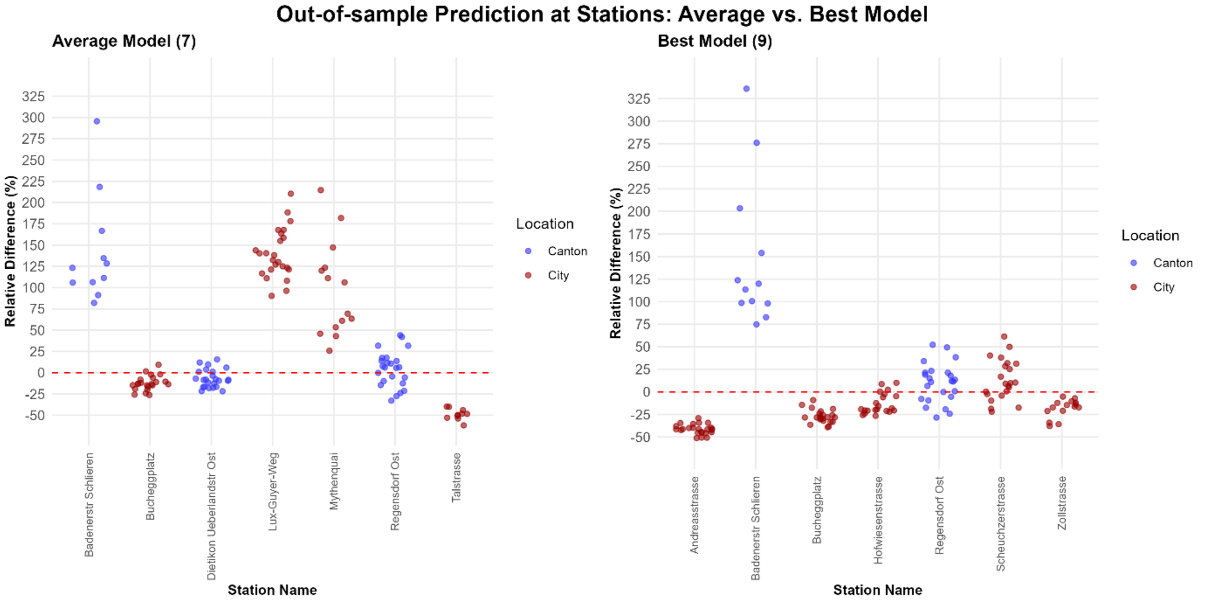


Figure 23: Out-of-sample Prediction at Stations: Average (Nr. 15) vs. Best (Nr. 16) Model
Note: 3 Observations of Station "Talstrasse" and 1 of "Badenerstrasse" of are out of bounds in the left Plot

Many stations have no high standard deviation inside their predictions, but the station has a shifted position as a whole – either over- or underestimated counts.

As for categorical maps, the Jaccard Index was calculated for each appearing class in the test set. For the classification, the natural class breaks visible in Table 15 were used. Table 16.1 and 16.2 show the results of for the respective model.

Table 15: Classes for Categorical Maps (full perimeter)

Class	Range
Very low	2725 – 12602
low	12603 – 27810
medium	27811 – 41608
high	68072 – 96681
Very high	>96680

The results in Table 16.1 and 16.2 show that the "medium" classes are predicted the worst from both models. Even in the best out-of-sample model, only approximately 1 out of 5 observations is labelled correctly in the "medium" class.

Table 16.1 and 16.2: Results Jaccard Index: Average Model (7) vs Best Model (9)

Average Model (7)					
Measure	Very low	low	medium	high	Very high
Intersection	42	8	0	11	5
Union	55	36	29	43	37
Jaccard Index	0.76	0.24	0	0.26	0.14

Best Model (9)					
Measure	Very low	low	medium	high	Very high
Intersection	18	21	9	23	9
Union	34	56	42	48	20
Jaccard Index	0.53	0.38	0.21	0.48	0.45

Between the models the average model performs better with very low counts while the best model is consistently better for all other classes. For this and all those out-of-sample comparisons, the selected stations are key and may be very influential with this small sample size.

GLMM 2: Model using all data as training set

Finally, in Table 17, the summary of the model using all data as training set is given. As no stations are used as test data, the resulting accuracy is an in-sample accuracy, where the model did not have to deal with unseen data. The respective error measures of the in-sample prediction are visible in Table 18.

The results show very good prediction accuracies. Unsurprisingly, the R^2 is again very high with 0.846 or 0.976, depending on the inclusion of the random effect. There are several significant predictors, whose effects are presented later in this section. Overall, the magnitude of the standard errors compared to the estimates means that the significance of the variables is sensitive.

Table 17: Model Summary of the GLMM using all Data as Training Set

Variable	Estimate	Std. Error	Pr(>|z|)
Intercept	5.620	0.3045	< 2e-16 ***
Strava Trips	0.6304	0.02094	< 2e-16 ***
Strava % of commute	0.001189	0.0008849	0.179155
Exposure to accidents	0.03861	0.0154	0.012153 *
Speed limit on street	-0.1247	0.0763	0.102117
Distance to green space	-0.0004298	0.000564	0.446012
Distance to residential area	0.001521	0.0008125	0.061204 .
Distance to POI	-0.0006629	0.0001419	2.96e-06 ***
Mixed land use	0.007228	0.07994	0.927948
Population density	0.0001421	0.00008715	0.103092
% of 80+ years old inhabitants	-0.01591	0.02074	0.442941
Swiss Neighbourhood Index	0.09118	0.02671	0.000641 ***
Slope of street	0.07318	0.1124	0.514943
Season: Winter	-0.2261	0.02787	5.02e-16 ***
AIC= 13099.7	Nakagawas's R^2: all effects: 0.976	Nakagawas's R^2: fixed effects: 0.846	Observations= 679
p-value: *** <0.001 ** < 0.01 * <0.05 . <0.1			

Table 18: In-Sample Prediction Accuracy Measures of GLMM using all Data

Mean Absolute Error (MAE)	3611
Root Mean Squared Error (RMSE)	6923
Mean Absolute Percentage Error (MAPE)	14.59 %

Plots in Figures 24 and 25 reinforce the impression of high accuracies: Figure 24 shows the model fit with all predicted and actual values.

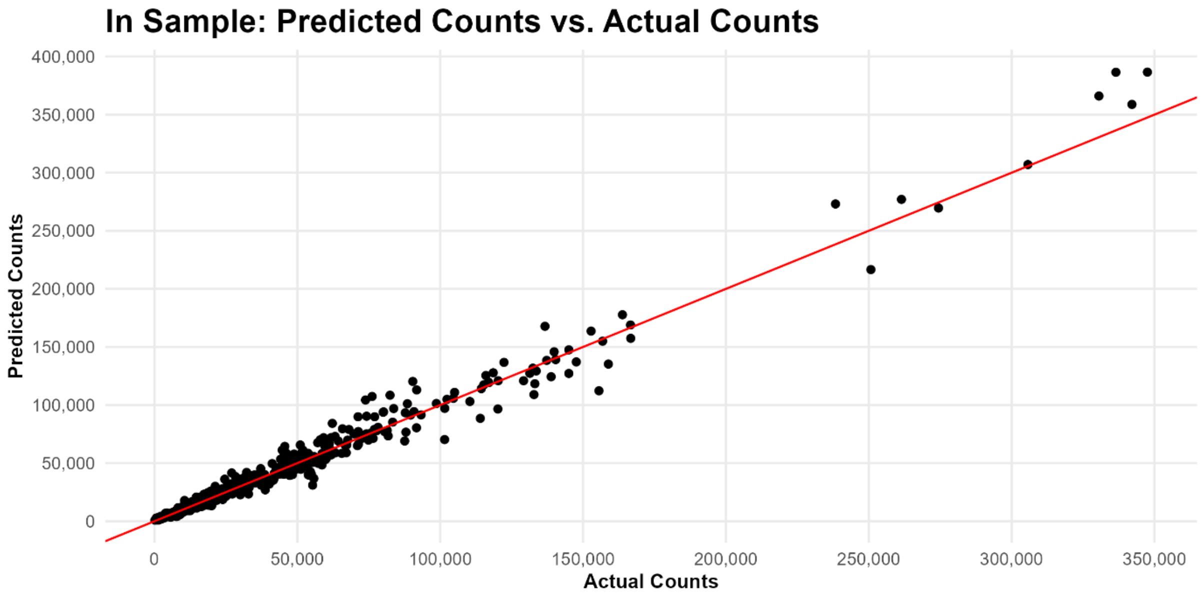


Figure 24: In-Sample Accuracy Predicted vs. Actual Values

Figure 25 shows a cumulative plot of the in-sample prediction accuracy. For 87% of the segments, ±25% of ridership and for 50% of segments, ±10% of ridership is predicted. Towards the low accuracies, there are few outliers, as for 97% of segments ±50% of riders are called correctly. 5 observations have a relative absolute percentage error of over 100%.

Figure 26 and 27 show the distribution of predicted observations after station, location and season.

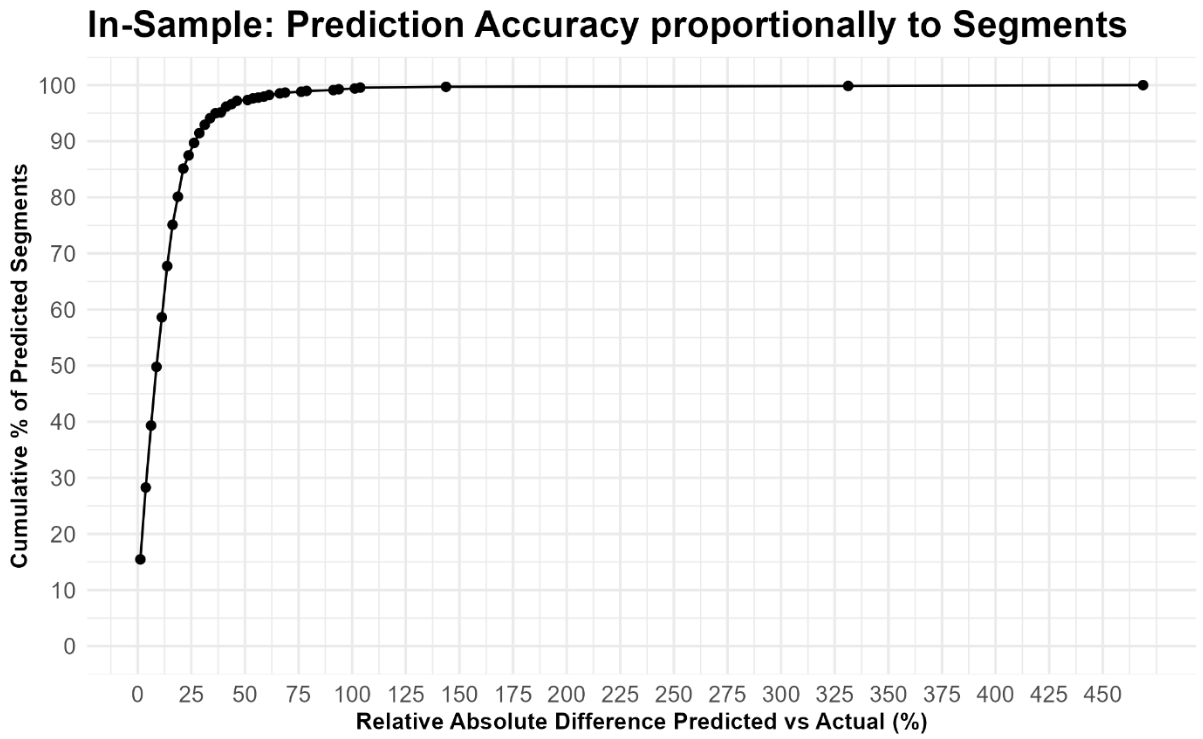


Figure 25: In-Sample Prediction Accuracy proportionally to % of Segments

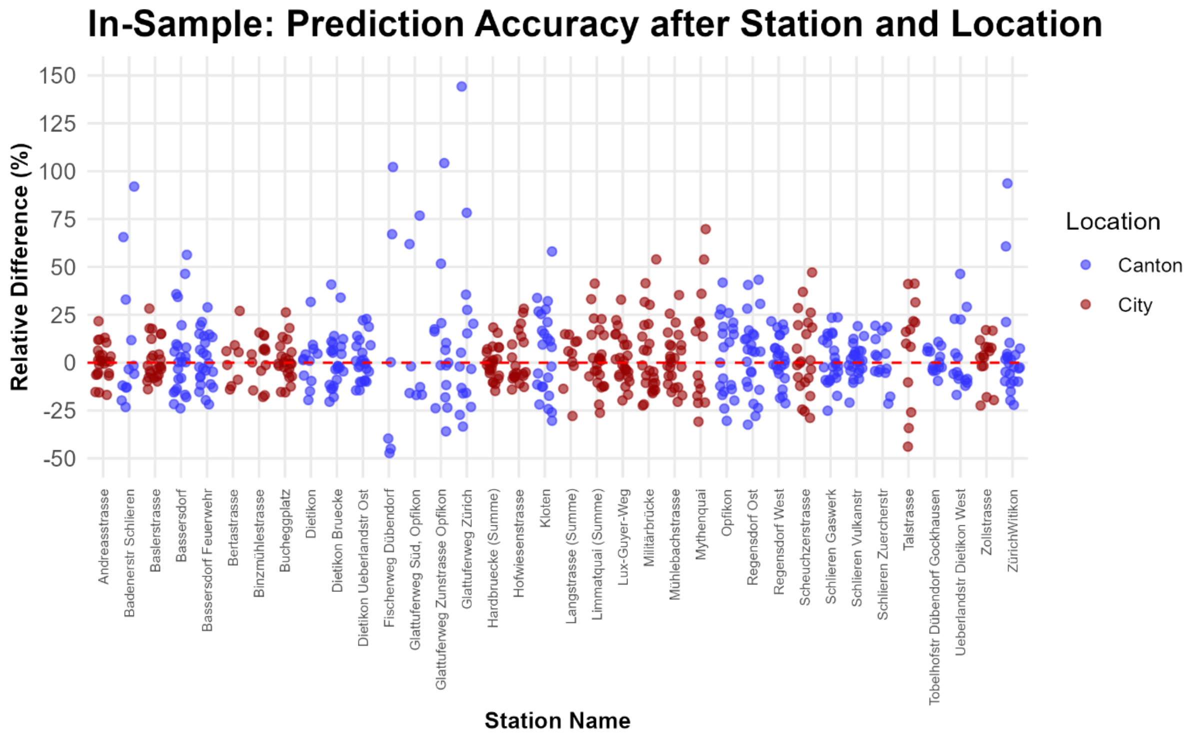


Figure 26: In-Sample Accuracy after Station and Location
Note: Two observations of station "Fischerweg Dübendorf" are outside the bounds (y= 330 and y= 468)

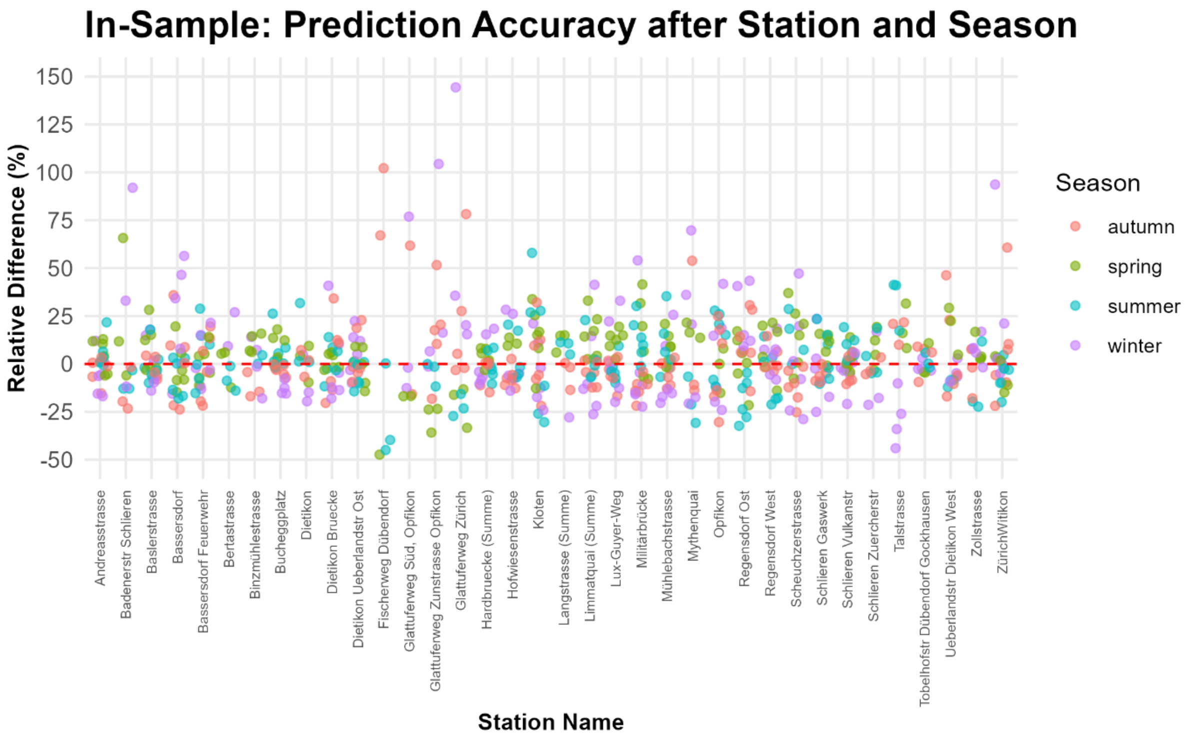


Figure 27: In-Sample Accuracy after Station and Season
Note: Two observations of station "Fischerweg Dübendorf" are outside the bounds (y= 335 and y= 452)

Figure 26 shows that of the stations with higher variation in predictions, almost all are operated by the Canton. The bad predictions show a similar pattern, of 19 observations that are worse than ±50% relative difference, 16 are Canton stations and 3 are City stations. In general, there are fewer strongly underestimated than overestimated counts.

Figures 28.1 and 28.2 show the distribution of seasons in accuracies lower than 25% respectively 50%.

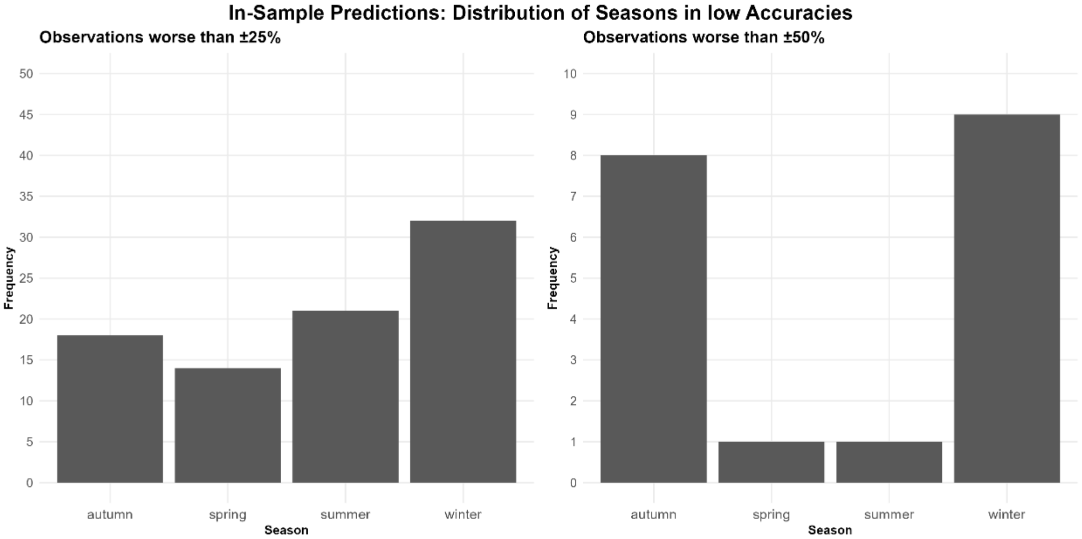


Figure 28.1 and 28.2: Distribution of Seasons in inaccurate In-Sample Predictions

Looking at the seasons of observations worse than ±50%, it appears that those too high predicted counts were mostly autumn and winter observations. For all observations worse than ±25%, the distribution is more even, whereas Winter has still the most observations of all seasons.

In Figure 29 a map shows the median relative difference for each station plotted on a map. While the relative accuracy is visualised by colours, the size of the circles also classifies the median station trip count. There are no clear trends visible. Geographically, it appears that stations in the Limmattal in the left are more accurate than in the Glattal towards the right upper corner. However, many stations in the Glattal are at riversides of the Glatt, compared to the stations in Limmattal, which are mostly at busy streets.

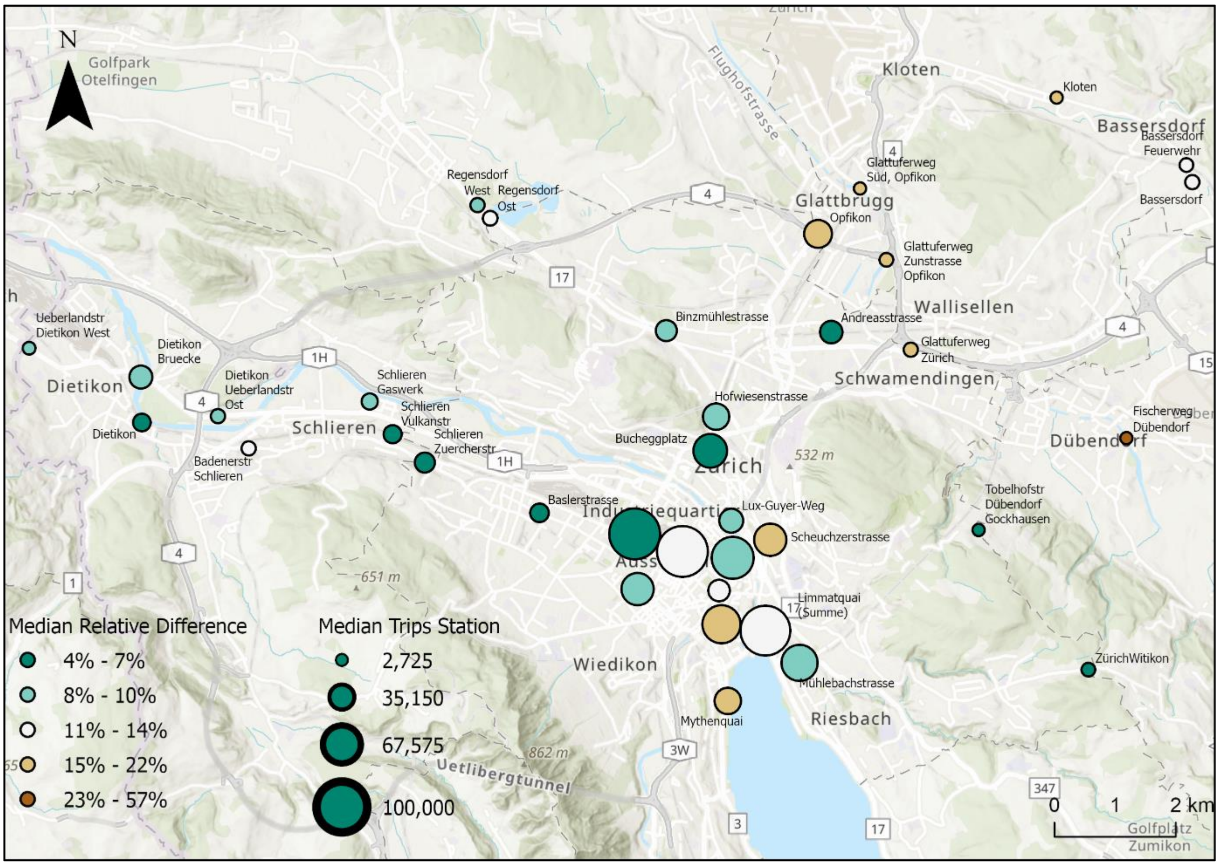


Figure 29: In-Sample: Median Relative Difference at Counting Stations

GLMM 2: Importance of Variables

Table 19: Effect of Geographic Covariates on Ridership

Variable	Unit	Estimate and p-value significance	exp(Estimate)	% Change in Ridership for 1 Unit increase
Intercept	1 bicycle rider	5.620 ***	239.3684	N/A
Strava Trips	log(1 Strava rider)	0.6304 ***	1.880	88% increase
Strava % of commute	1 % of Strava trips that are commuting	0.001189	1.001	0.1% increase
Exposure to accidents	1 accident	0.03861 *	1.039	3.9% increase
Speed limit on Street	1 street category	-0.1247	0.882	11.8% decrease
Distance to green space	100m	-0.0004298	0.957	4.3% decrease
Distance to residential area	100m	0.001521 .	1.164	16.4% increase
Distance to POI	100m	-0.0006629 ***	0.935	6.5% decrease
Mixed land use	1 more different land use class	0.007228	1.007	0.7% increase
Population density	100 Persons	0.0001421	1.014	1.4% increase
% of 80+ years old inhabitants	1 % of 80+ years olds	-0.01591	0.984	-1.6% decrease
Swiss Neighbourhood Index	1 decile of the index	0.09118 ***	1.095	9.5% increase
Slope of street	1 slope category	0.07318	1.075	7.5% increase
Season: Winter	Binary. Compared to reference season autumn	-0.2261 ***	0.797	20.7% decrease
AIC= 12301.1	Nakagawas's R^2: all effects: 0.974	Nakagawas's R^2: fixed effects: 0.858	Observations= 631	
p-value: *** <0.001 ** < 0.01 * <0.05 . <0.1				

Table 19 shows the effect of each covariate on bicycle ridership. Note that the Strava trips and the winter season cannot be compared to all other variables, because of their respective implementation.

5 Predictors were found to be significant for predicting bicycle ridership in Zurich's Urban Area: Number of Strava Trips (***), the exposure to accidents (*), the distance to POI(***), the Swiss Neighbourhood Index(***) and if the Count has been recorded in Winter(***). The intercept is also strongly significant with p < 0.001. The value of the intercept shows that, if all predictors are zero-values and the season is autumn (reference level), a street segment has a base line ridership of 239 riders.

Regarding the effects, the table can be interpreted as follows: For 100m more distance to a POI, there is a 6.9% decrease in bicycle ridership. For the Strava trips, the unit increase is on the log-scale. Due to the log-scale, the one-unit increase appears larger, which results in the high percentage of 88%. The seasons are interpreted as follows: if an observation was recorded in Winter, bicycle ridership decreases by 20.7% compared to autumn.

GLMM 2: Categorical Maps

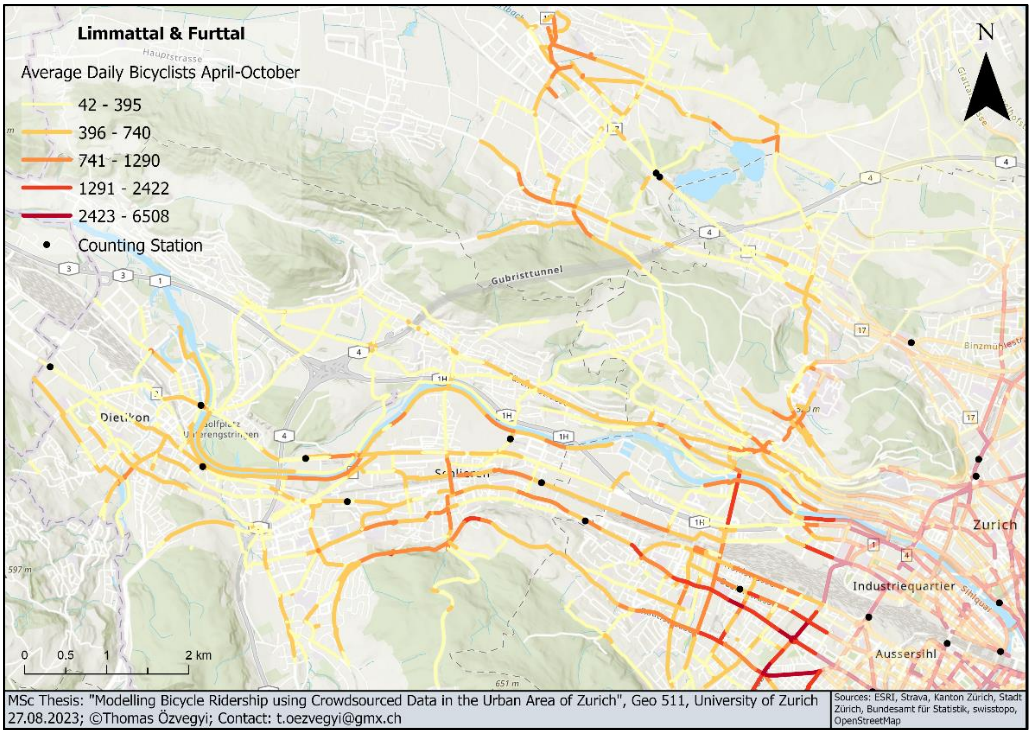

Limmattal & Furttal
Average Daily Bicyclists April-October
42 - 395
396 - 740
741 - 1290
1291 - 2422
2423 - 6508
Counting Station
N
Gubristtunnel
Dietikon
Schlieren
Zurich
Industriequartier
Aussersihl
0
0.5
1
2 km
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch
Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap

Figure 30: Example of a Categorical Map as Output of GLMM 2

Finally, Figure 30 shows an example of a categorical map for one of three regions of the perimeter. All other maps and this map in original size can be found in the Appendix D.

6 Discussion

In this thesis, various steps were taken to predict bicycle ridership in the urban area of Zurich, Switzerland. The crowdsourced data collected by Strava had to be matched to a complete but simple street network. To mitigate the user bias in contribution to Strava, a wide range of geographic variables were included. Ridership data from 36 counting stations of the authorities of the Canton and the City of Zurich served as ground truth for the GLMM, which aimed to predict ridership for at least 7760 street segments, where a sufficient number (>2000 in 24 months) of Strava trips was available.

The prediction of ridership led to diverse results. Depending on the sampling strategy, both in-sample or out-of-sample accuracies have been computed. To answer the research questions, a general assessment of the prediction will be done in Section 6.1., whereas the importance of geographic variables is covered in 6.2. The limitations of this thesis will be summarised in Section 6.3.

Apart from the ridership prediction using GLMMs, the use of crowdsourced data by Strava more generally can be assessed. For the proportion of captured trips by Strava, the results are as one might expect: In the city there are higher numbers of bicycle trips, but lower proportions of trips are recorded by Strava users. Strava is used more for leisure trips than for commuting and those leisure trips might rather be in suburban areas than in the core city. Another explanation would be that those who track their commutes in Strava are people who ride longer distances and move not only inside city borders. All the highest proportion values are recorded at stations of the Canton, however only few stations stand out. The proportions of the counting stations of the Canton can also be compared to the figures in the study by EBP (2019), who analysed all stations in the Canton of Zurich: They recorded a mean proportion rate of 6.6%, which is close to the mean of canton stations in this thesis (5.9%). The stations that lie in more urban areas also performed worse in their report.

The correlation is generally similar in the sense that most stations perform well with a Pearson correlation of around 0.9. The mean in the study by EBP (2019) is higher (0.94) than in this thesis (0.86 before removal of outlier stations, and 0.93

of considered stations). However, the results could be even closer, as in their study, many stations which performed above average lie outside of the perimeter of this thesis. Looking at stations that are covered in both studies, most results are replicated in this thesis: The correlation of many stations has remained high. An exception is the counting station Kloten, Schaffhauserstrasse (Nr. 1819), which was above average in the study by EBP but was the worst performing canton station in this study with only around 0.59 Pearson correlation. A change in the measurement or other street conditions may have caused the significant decrease of trip counts in 2022.

EBP only considered the year 2019 in their analysis and for some stations only one or a few months were available. This thesis included years 2021 and 2022 and had on average 18.86 months per station, but this may still not be enough to make valid statements in all cases. Proportion and/or correlation values are sensible to interruptions like construction works, as there are only 3-24 observations per station.

6.1 Prediction of Ridership

Different studies have shown examples of ridership prediction using Strava data as Section 2.3 has shown. One aspect that has not been examined in those studies is the difference between in-sample and out-of-sample accuracies. Nelson et al. (2021) referred to in-sample accuracies due to the lack of ground-truth data. Jestico et al. (2016) computed accuracies which also count towards in-sample accuracies, as they used what is called random sampling in this thesis. Roy et al. (2019) on the other hand presented out-of-sample accuracies. All studies include precise accuracies, however they are not always easy to interpret due to the absence of relative accuracies (Roy et al.,2019 and Nelson et al.,2021) or because of different time intervals, as they strictly used the annual daily average (AADB). In this thesis, both types of accuracies are covered. This thesis is also the first work in this study field that used a mixed model with a random effect to account for repeated measurements. In the following chapters, results of both types of accuracies are discussed and contextualised with the respective state-of-the-art study: First, the out-of-sample estimates of GLMM 1 in Section 6.1.1 and secondly the in-sample accuracies of the final GLMM 2, in Section 6.1.2. The research question regarding prediction (RQ 1) is adressed in Section 6.1.3.

6.1.1 Out-of-Sample Accuracies

For out-of-sample predictions, there is no single figure of accuracy. The high variability during the 20-fold cross-validation is a clear sign that more stations, or more similar conditions at stations, would be needed to arrive at more stable predictions. The far more stable results of the random sampling strategy prove this point. The model fit on the other hand is constant, which shows that the model itself is well suited for all kind of station subsets.

The average MAPE over 20 folds is 72% with a median of 56.76% and a standard deviation of 49%. Selecting a fairly average and the best model, ±25% of riders can be predicted for about 45% of segments (average model), respectively 50% of segments for the best iteration. The difference is higher looking at the accuracy where ±50% relative difference is still reached: the average model does that for 60%, the best model already for 90% of segments. This could mean that certain characteristics, i.e. certain stations are more consistently predicted over 20 iterations than others. Roy et al. (2019) state ±25% for 80.3% of segments, although it is difficult to match this statement with the results in their paper. Even if the monthly values of this thesis are converted and averaged to the AADB figure, the comparison with their study is not insightful, as they reported absolute values of differences between predicted and observed counts.

To better understand the results one can further refer to the plots showing the predictions for each station of the respective test set. It shows that many stations do not have great variance between their observations but are consistently over- or underestimated (biased). This suggests that the model is not capable of identifying the characteristics of an unseen street edge so well that precise predictions are possible. Even when the resulting accuracies are classified for a categorical map with 5 classes, the Jaccard Index showed bad performances in the attribution to the right class. Nevertheless, predictions for some single stations are very good, but the accuracy of the test set as a whole is clearly affected by single stations.

Overall, the difficulty of out-of-sample estimates of bicycle ridership are illustrated well by the results of this study. The different conditions in terms of street category and environment and the differences in seasonality are not sufficiently captured by the GLMM. Moreover, results of different out-of-sample iterations are not easy to interpret and compare. It does not only matter which stations are in the test set but also on which stations the model has been trained on. This is illustrated nicely as the station “Badenerstr. Schlieren” is in the test set of both selected iterations, yet the vertical positions of observations are not exactly equal. In this case, the best iteration has slightly higher variation and worse predictions. Hence the in- or exclusion of the 6 other stations make a difference in the accuracy of any given station.

6.1.2 In-Sample Accuracies

GLMM 2 is the final and most expressive model of this thesis. All 36 stations were used to train the model and predict ridership for the whole perimeter. Consequently, only in-sample accuracies could be obtained. The MAPE of 14.59% is a very good result, which is slightly better compared to the mean MAPE of the random sampling in GLMM 1 (15.44%). This is the same for all error measures. Interestingly, the median MAPE (14.55%) over 20 iterations of the random sampling even outperforms the MAPE of GLMM 2. This may indicate that some outlier observations still have a considerable effect on the prediction accuracies and if in many iterations of the random sampling, those are not in the test set, the accuracy is considerably better.

The plots confirm the presence of outliers: The station “Badenerstrasse Schlieren” (Nr. 421) has some bad predictions again, but is overall much more precise than in the out-of-sample models. Many of the outliers belong to stations located at the riverside of the Glatt. Looking at all observation with a relative difference above 50%, a certain pattern catches one’s eye: Of those 19 observations, most are Canton stations (16 vs. 3). Moreover, a majority are located next to a river or the lake of Zurich and the seasons are not equally distributed (9 winter, 8 autumn, 1 spring and 1 summer observations). This suggests that the model does not predict well at places where the seasonality may be very strong and trips are more likely to be of a leisure purpose as, in the study area, path at riversides are often meandering and not paved.

Nevertheless, the GLMM 2 outperforms in terms of accuracy related work by Jestico et al. (2016) or Nelson et al. (2021). For the R^2 as measure of the in-sample model fit, both the random sampling of GLMM 1 and GLMM 2 resulted in an R^2 of 0.98. Nelson et al. (2021) had 0.92 at their best performing city model. Relative Prediction Accuracy is only comparable to the study by Jestico et al. Their stated average model error of 38% is clearly higher than both in-sample MAPEs in this study. The prediction inside ±25% of the actual value for 87% of segments and inside ±10% for 50% of segments are not comparable to related studies, but clearly indicate the very good performance of the model if all characteristics of station are in the training data.

6.1.3 Synthesis and General Assessment

RQ 1 asked "how well is bicycle ridership in Zurich's Urban Area predictable from Strava and geographic data?". In H1, it was suspected that a similar prediction accuracy could be obtained to Roy et al. (2019).

Overall, it can be said that bicycle ridership is predictable from Strava and geographic data. As Sections 6.1 and 6.2 showed, the great difference lies in the type of prediction that is intended. Good out-of-sample predictions would be the ultimate goal: A model that correctly predicts ridership at random locations on the base of Strava riders and a set of expressive geographic variables. Results suggest that for out-of-sample estimates, only up to 50% of segments can be predicted in a range of ±25% of riders and the median MAPE over 20 iterations is 57%. Even for the classification into categories to display on a map, the results show too little veracity with the assigned classes. The 20-fold cross-validation has shown high variability in accuracy which indicates that the GLMM does not perform equally well for all subsets of 29 stations as training set and more stations would likely be needed. The same GLMM performed significantly better if it already had a knowledge of stations. When observations were sampled randomly, a MAPE of 15.44 % was reached. A notable disparity between in- and out-of-sample estimates is also that for out-of-sample estimates, whole stations are shifted to under- or overestimations, while for in-sample predictions, the model seems to correctly identify a base line for each station. The great difference between these two sampling methods is one of the notable outcomes of this thesis, as in literature, this issue has generally been ignored.

As best-outcome model of this thesis, the GLMM 2 has been run using all stations as training data. The results show similar accuracies to random sampling in GLMM 1. The prediction accuracies plotted per station reveal some insights into where the model has difficulties even at in-sample predictions, where train and test data are the same. It appears that stations at watersides in colder months account for most of the bad in-sample predictions. Compared to warmer months, these stations may have a lack of leisure riders, which leads to the wrong estimates. The partition of leisure and functional riders (including activities such as commuting or running errands) is a point which should be more rigorously assessed for future models. It may even make sense to compute separate models.

In the process towards the final results, two rather non-urban stations (421 Kloten, 919 Zürich Witikon) with a high proportion of Strava riders were at first not considered. When they were added for reasons of consistency, the out-of-sample results improved. However, the standard deviation in the results also surged considerably - not the mean, but only the median MAPE got significantly better. For in-sample accuracies, the results remained the same or got slightly worse. This was a further sign that the addition of stations helps, but a clear focus on the characteristics of streets and station would possibly improve the model even more. Certainly, for a model like in this thesis, more stations would still be needed to get better out-of-sample estimates and therefore make the model spatially more predictive.

Using the GLMM 2, a set of categorical maps for the whole and three areas of the perimeter has been produced (see Appendix D). All 36 stations have been used as training data, which means that there is no estimate of the out-of-sample accuracy. The classification in categories is a further simplification. For the best out-of-sample model, the veracity of this classification was still moderate. Nevertheless, the accuracy in the final model is probably better, as a 20-fold Cross-Validation using 90-10 train-test splits has resulted in a median MAPE of 49%, compared to the median MAPE of 57% for 80-20 splits. 31 stations were used as training data for those 90%, thus it can be seen as intermediate step between the 80-20 splits and the final GLMM 2. The variability is still large, as the mean MAPE of 68% with a standard deviation of 48% suggests. Visually, it is evident that the counts in the final maps show some jumps between neighbouring edges. This happens because segments are treated independently and spatial autocorrelation is not considered.

Certainly, predictions of the final models can be used as reference for other models predicting bicycle ridership. For the upcoming years, the model and predictions could again be used provided that all necessary data is made available. If authorities of the Canton and/or the City of Zurich intend to adopt the approach of this thesis with one model, more counting stations are recommended, but also a sort of classification of each station regarding the type of riders it captures. Additional measurements would provide valuable ground-truth to assess the accuracy of the GLMM 2 of this thesis. Alternatively, two separate models are suggested: One for leisure-oriented routes and one for all other streets.

The research question RQ 1 cannot be answered providing a single estimate. Certainly, bicycle ridership can be predicted well from Strava and geographic data, but it cannot yet be used to get accurate predictions at random locations. Assuming that the accuracy stated by Roy et al. (2019) is correct, H 1 must be rejected for out-of-sample estimates. For in-sample accuracy, this thesis presents even better results than related studies.

Key Messages

- This thesis assesses the difference between in- and out-of-sample predictions using different sampling strategies.
- Out-of-sample predictions are unstable and depend on the subsets of data used in the 80-20 train-test split. ±25% of riders can be predicted for a maximum of 50% of segments for the best iteration.
- In-sample accuracy of ±25% of riders is reached for 87% of segments.
- Outliers affect the performance of the model. It appears that the strong seasonality of some stations is not captured well. (...)

- The predictions using all stations as training data can be used as reference for similar studies and provide a continuous map of bicycle ridership in Zurich's Urban Area

6.2 Importance of Variables

The geographic covariates were necessary to account for the bias in the Strava data. Due to the absence of similar work in Switzerland or even Central Europe, it was completely unknown which variables besides the Strava trips would be the best predictors for bicycle ridership in Zurich's urban area. RQ 2 aimed to answer that question, while a hypothesis (H2) of the author was that socio-economic variables, that proved to be significant in related studies in the US, might not be that relevant here in Switzerland. Results show that the most significant local predictors are Strava trips, Distance to POI, the Swiss Neighbourhood Index (Panczak et al., 2023) and Winter as season. Also significant ($p<0.05$) was the exposure to accidents.

As the socio-economic variable Swiss Neighbourhood Index belongs to the most important covariates, the hypothesis H2 must be rejected. It must be noted however, that the significant variable in the studies by Nelson et al. (2021) and Roy et al. (2019) was the median household income. The Swiss Neighbourhood Index considers other socio-economic variables as well, such as all-cause mortality and housing parameters. A 9.5% increase in ridership for a one-decile higher socio-economic index is a considerable result. Future work could focus on the association between those variables to validate this effect.

Looking at the other variables in order of the model summary, the Strava trips were the most significant predictor, as was expected. Due to the logarithmic transformations, the magnitude of the effect is not comparable to related studies.

The Strava % of commute was not significant for bicycle ridership in this perimeter. This contrasts to Nelson et al. (2021), where it was significant for all cities. Reasons could be the uncertainties in the network matching, but also differences of usage in Strava users or in the commuting patterns in Switzerland.

Crash density is relevant as already Nelson et al. (2021) suggested. Considering the range of values, the 3.9% increase seems to be a realistic result. If a safe, separate and convenient bicycle network would be available, people might favour routes that have seen less accidents. As long as this perfect state is not met, the number of accidents is a solid covariate if a long enough period of time is considered.

The speed limit was not significant, which contrasts to findings by Roy et al. (2019). A reason could be that many stations are on streets with at least 50km/h speed limit (Category 2). Also the implementation in this thesis is prone to errors, due to the use of incomplete OSM data and a simple matching to the network.

Distance to green space was no significant predictor in this study, in contrast to Roy et al. (2019) or in some cities in the study by Nelson et al. (2021). The decision to consider only one class of the Arealstatistik raster data for this variable may be an issue here. However, the addition of other classes including for example forests led to a higher AIC, worse in-sample accuracies and no change in significance. Therefore, the old version was maintained.

The distance to residential area is also not significant. The correlation would have been positive, which means that a higher distance is associated with more ridership. This would have been a clear contrast to findings in related studies (e.g. Roy et al.,2019). The reasons for this probably lie in the distribution of values visible in the histogram (see Appendix C).

The distance to POI was strongly significant and the model records a 5% decrease in ridership for each 100m more distance to a POI. This is a reasonable result and the association is as expected. For future work, similar implementations are suggested to account for amenities inducing traffic on a daily basis.

Mixed land use was clearly not significant in the model. Research is divided whether mixed land use correlates with active transportation. Concerning this thesis, the overall mixed land use patterns and the inclusion of diverse counting stations may have caused that result.

Both population density and the percentage of 80+ year olds are not significant. Probably, the correlation to bicycle ridership exists, but the effect is just not significant enough.

The slope of the street segment was no significant predictor for ridership. The use of Strava data has certainly influenced this result. It could also be that small slopes are not a great hindrance for riders in cities or that the implementation using three classes from 0-3 has smoothed the effect of this variable.

The result for winter as season is as expected: Compared to autumn, on average 20.7% less ridership is recorded in this perimeter. For Ottawa, Nelson et al. (2021) recorded a strong decrease of 52%, when all other months served as reference, again showing the importance of seasonality.

As was already discussed above in Subsection 6.1.3, seasonality is one of the factors that make the modelling of bicycle ridership so difficult. The significance of winter as a season is confirmed equally in the assessment of predictions and in the model output itself. The methodology to capture the seasonal variability is one of the major choices to make. This thesis provides an example of an approach, using Strava counts that include the variability and, adding to that, the winter season as variable.

To conclude, the Strava trips, the exposure to accidents, the distance to POI, the Swiss Neighbourhood Index and winter as season are significant predictors for bicycle ridership. As in several studies before (e.g. EBP, 2020; Jestico et al.,2016), Strava trips proved to be strong correlates to ridership and follow the same patterns of seasonality. The significance of accidents is a confirmation that accidents are generally related to volumes of ridership, like Nelson et al. (2021) suggested. The significance of the Swiss Neighbourhood Index means that bicycle ridership correlates with socio-economic parameters. The lower the socio-economic status of a neighbourhood, the less bicycle ridership is expected, coinciding with studies of other countries (Roy et al, 2019; Nelson et al.,2021). The seasonality of ridership has been mentioned several times in this thesis and the significance of winter as the season of measurement confirm that weather is a main factor for ridership volumes.

As for the predictions, the addition of stations influenced the significant variables. The magnitude of the standard errors suggested that the variables might be sensitive and this was proved, as due to the addition of two stations, speed limit on street, population density and distance to residential area lost their significance. The latter was even highly significant before. This shows that single stations can also make a big difference for the covariates, especially when the station has extreme values for some variables. The use of a separate model for routes focused on leisure riders would probably lead to other significant variables.

Key Messages

- The most significant local predictors for bicycle ridership are Strava trips, Distance to POI, the Swiss Neighbourhood Index, winter season and the exposure to accidents.
- There are similarities but also contrasting results to related studies conducted in the US or Canada.
- The implementation and distribution of values of a variable in the training set must be taken into consideration when interpreting the significance and magnitude of the effect.
- Depending on their characteristics, the in-or exclusion of single stations can alter the significance of variables in a sample of 36 stations

6.3 Limitations

6.3.1 Data Availability

Network

The network posed the first challenge of this thesis. As described, the construction of the network involved various steps and uncertainties. The azimuth filter for example managed to filter out many unwanted street fragments, however, also some valid street edges got omitted, in the rare case that Strava segments had an azimuth of more than ±20° in long, winding segments, often at riversides. The further uncertainties lay in the counts. As mentioned, there were mismatches especially at crossings. Not all of them caught the eye of the manual corrector. As the correction focused on large gaps on rather popular routes, there may be more remaining mismatches on streets with low Strava ridership. Furthermore, the correction process itself involves large uncertainties as it based on the assumption that spatially close streets with similar counts in one month have similar counts in all months. The limitations of this method not only apply to the Strava counts, but also the percentage of commute. Possibly, the percentage of commute would have been a more expressive variable if it was not for the network matching process.

The use of a network with simple, undirected edges facilitates the interpretation and visualisation of counts. However, it has the disadvantage that all streets are represented alike. There is no possibility to distinguish a street with traffic-separated bike lanes, for example.

Bicycle Data

The infrastructure of bicycling is one of the most important factors for bicycle ridership (Sallis et al.,2013) particularly the improvement of infrastructure could motivate many people to use bicycles more. A major limitation of this thesis is that bicycle infrastructure is not considered as predictor. Unfortunately, the availability of suitable data is fragmentary. One issue is the existence of two different authorities in the perimeter, the Canton and the City of Zurich. Both have their own mobility and bicycle departments. Naturally, there is coordination, but from a data perspective, no joint base data is available. Datasets by the City, for example the speed limit on streets, stop at the city borders. On the other hand, data maintained by the Canton includes only streets and infrastructure maintained by them, thus even if there is data for bicycle infrastructure, only big cantonal streets running through municipalities are included. The goal of the approach of this thesis on the other hand is to have prediction for all possible streets. Another problem is that occasionally, data is not up-to-date or it only shows a desired state of the future. The latter applies to the network of fast lanes for bicycles that is being built at the moment – the routes of the finished network are not suitable for training a model that predicts ridership in the past, on the base of data from 2021 and 2022.

Overall, more and more data got available as OGD and over the last years, authorities have certainly made efforts to improve data and its availability. Nevertheless, there is still much room for improvement for new data, its availability and its maintenance.

6.3.2 Data Quality

Trip Counts

The data quality of the counts is very important, as they form the ground truth in the case of the official counts or the most significant predictor in case of the Strava counts. As described in Subsection 3.2.1, technical problems of the counting stations can lead to gaps in the data, which can complicate comparisons between stations. Other limitations are changing street conditions at the stations: For some filtered stations, there were inexplicable leaps in the counts which were so unrealistic that these months were filtered out. Some decisions on filtering can be difficult as bicycling data has patterns which seem extreme at times, especially in routes popular for leisure trips. The uncertainties in the Strava data concerning the matching to a suitable network and binning has been discussed above, respectively in Subsection 3.2.1. The binning is one reason that smaller Strava counts are not as reliable. For this reason, one counting station got discarded and the filter of edges (>2000 Strava trips in 24 months) was applied.

Geographic Variables

The sources of uncertainties in the covariates have been described in the description of each variable in Subsection 3.2.2. The implementations had influenced effects of the variable on ridership, which has been discussed in Section 6.2. It can generally be said that future work could focus more on this part of the model. Better data, more tests on the sensibility of different implementations per variable and the discussion among different authors could improve a future model of bicycle ridership.

Spatial Autocorrelation

While this study aimed to predict bicycle ridership using Strava data and a range of geographic variables within a GLMM, it is important to acknowledge that spatial autocorrelation was not considered in our analysis. Spatial autocorrelation, where nearby areas have similar values, may affect the reliability of predictions. Neglecting spatial autocorrelation could have led to biased variables or unrealistic differences in nearby predictions. Future research could account for spatial autocorrelation and enhance the approach of this thesis. A first step would be checking for spatial autocorrelation in model residuals.

7 Conclusion and Outlook

This thesis provides a continuous prediction of bicycle ridership in the urban area of Zurich. For the model of ridership, crowdsourced Strava counts and various geographic covariates of bicycling were used as predictors of a Generalised Linear Mixed Model, in which official count data of the City and the Canton of Zurich served as ground truth. Two different sampling strategies led to results for in-sample as well as out-of-sample accuracies: For 87% of segments, monthly counts can be predicted within ±25% relative difference for in-sample estimates. Out-of-sample predictions are instable and depend on the respective subsets of stations: ±25% relative difference is predicted for at most 50% of segments. A final model using all ground truth data to train the model has been used to map the ridership.

The main contribution of this work is providing a state-of-the-art workflow to predict ridership using Strava and geographic data. Compared to related studies, this thesis also includes the construction of a generalised network to match the Strava data. More importantly, this is the first study in this field which assesses the differences between in-sample and out-of-sample estimates.

7.1 Future Work

Future studies could build upon this work: To better capture the seasonality of bicycle ridership, it is suggested to either build a separate model for routes that are primarily frequented by leisure riders or add the predominant type of rides as an additional variable for each counting station. By doing so, the model may better understand the stronger seasonality of leisure rides. Furthermore, more counting stations and the consideration of a longer time period would help to improve predictions. As for the model inputs, the correlation of geographic variables could be assessed further and the incorporation of bicycle infrastructure would compensate for a limitation of this thesis.

8 Bibliography

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, *46*(10), 1730–1740. https://doi.org/10.1016/J.TRA.2012.07.005

Brun, I. (2022). *Zoff um Zürcher Velovorzugsrouten verzögert Umsetzung*. Tsüri. https://tsri.ch/zh/velovorzugsroute-zuerich-einsprachen-hoengg-parkplaetze-blaue-zonen.LUCjzUfCEXEPJJ9N

Büchel, B., Marra, A. D., & Corman, F. (2022). COVID-19 as a window of opportunity for cycling: Evidence from the first wave. *Transport Policy*, *116*, 144–156. https://doi.org/10.1016/J.TRANPOL.2021.12.003

Celis-Morales, C. A., Lyall, D. M., Welsh, P., Anderson, J., Steell, L., Guo, Y., Maldonado, R., Mackay, D. F., Pell, J. P., Sattar, N., & Gill, J. M. R. (2017). Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ (Clinical Research Ed.)*, *357*, j1456. https://doi.org/10.1136/BMJ.J1456

Dean, C. B., & Nielsen, J. D. (2007). Generalized linear mixed models: A review and some extensions. *Lifetime Data Analysis*, *13*(4), 497–512. https://doi.org/10.1007/s10985-007-9065-x

Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. http://www.springer.com/series/417

EBP on behalf of the Canton of Zurich. (2020). *Eignung von STRAVA-Daten für Fragestellungen des Veloverkehrs*. https://www.zh.ch/de/mobilitaet/veloverkehr/veloinfrastruktur/datengrundlagen.html#406702762

Eurostat. (2022, August 25). *Climate change - driving forces - Statistics Explained*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Climate_change_-_driving_forces&stable=1#Total_emissions.2C_main_breakdowns_by_source_and_general_drivers

Garber, M. D., Watkins, K. E., & Kramer, M. R. (2019). Comparing bicyclists who use smartphone apps to record rides with those who do not: Implications for representativeness and selection bias. *Journal of Transport & Health*, *15*, 100661. https://doi.org/10.1016/J.JTH.2019.100661

Graser, A., Stutz, P., & Loidl, M. (2021). Tracks vs. Counters: Towards a Systematic Analysis of Spatiotemporal Factors Influencing Correlation. *GIScience: International Conference on Geographic Information Science*.

Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., & Xu, Z. (2012). Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. *Landscape and Urban Planning*, *107*(3), 307–316. https://doi.org/10.1016/J.LANDURBPLAN.2012.06.005

Heesch, K. C., & Langdon, M. (2016). The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health Promotion Journal of Australia*, *27*(3), 222–229. https://doi.org/10.1071/HE16032

Hood, J., Sall, E., & Charlton, B. (2013). A GPS-based bicycle route choice model for San Francisco, California. *Http://Dx.Doi.Org/10.3328/TL.2011.03.01.63-75*, *3*(1), 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information 2013, Vol. 2, Pages 507-530*, *2*(2), 507–530. https://doi.org/10.3390/IJGI2020507

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97. https://doi.org/10.1016/J.JTRANGEO.2016.03.006

Kanton Zuerich. (2021). *Veloschnellrouten | Kanton Zürich*. Tsüri. https://www.zh.ch/de/mobilitaet/veloverkehr/infrastruktur/veloschnellrouten.html

Kanton Zürich. (2020, December 10). *Regierungsratsbeschluss Nr. 1195/2020 | Kanton Zürich*. https://www.zh.ch/de/politik-staat/gesetze-beschluesse/beschluesse-des-regierungsrates/rrb/regierungsratsbeschluss-1195-2020.html

Kanton Zürich. (2022). *Datengrundlagen Veloverkehr | Kanton Zürich*. https://www.zh.ch/de/mobilitaet/veloverkehr/veloinfrastruktur/datengrundlagen.html#-18063639

Kanton Zürich. (2023). *Veloinfrastruktur*. https://www.zh.ch/de/mobilitaet/veloverkehr/infrastruktur.html

Labatut, V., & Cherifi, H. (2012). Accuracy Measures for the Comparison of Classifiers. *ArXiv E-Prints*, arXiv:1207.3790. https://doi.org/10.48550/arXiv.1207.3790

Larsen, J., Patterson, Z., & El-Geneidy, A. (2013). Build It. But Where? The Use of Geographic Information Systems in Identifying Locations for New Cycling Infrastructure. *Http://Dx.Doi.Org/10.1080/15568318.2011.631098*, *7*(4), 299–317. https://doi.org/10.1080/15568318.2011.631098

Livingston, M., McArthur, D., Hong, J., & English, K. (2021). Predicting cycling volumes using crowdsourced activity data. *Environment and Planning B: Urban Analytics and City Science*, *48*(5), 1228–1244. https://doi.org/10.1177/2399808320925822/ASSET/IMAGES/LARGE/10.1177_2399808320925822-FIG2.JPEG

Lustenberger, N., Becker, F., Hintermann, B., & Axhausen, K. W. (2021). *Änderung des Verkehrsverhaltens während der COVID-19 Pandemie*. https://doi.org/10.3929/ETHZ-B-000519119

Meinshausen, N., & Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(4), 417–473. https://doi.org/10.1111/J.1467-9868.2010.00740.X

Meister, A., Axhausen, K. W., Felder, M., & Schmid, B. (2022). Route Choice Modelling for Cyclists on Dense Urban Networks. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4267767

Menghini, G., Carrasco, N., Schüssler, N., & Axhausen, K. W. (2010). Route choice of cyclists in Zurich. *Transportation Research Part A: Policy and Practice*, *44*(9), 754–765. https://doi.org/10.1016/J.TRA.2010.07.008

Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., de Nazelle, A., Dons, E., Gerike, R., Götschi, T., Int Panis, L., Kahlmeier, S., & Nieuwenhuijsen, M. (2015). Health impact assessment of active transportation: A systematic review. *Preventive Medicine*, *76*, 103–114. https://doi.org/10.1016/J.YPMED.2015.04.010

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nehme, E. K., Pérez, A., Ranjit, N., Amick, B. C., & Kohl, H. W. (2016). Sociodemographic Factors, Population Density, and Bicycling for Transportation in the United States. *Journal of Physical Activity and Health*, *13*(1), 36–43. https://doi.org/10.1123/JPAH.2014-0469

Nelson, T., Roy, A., Ferster, C., Fischer, J., Brum-Bastos, V., Laberee, K., Yu, H., & Winters, M. (2021). Generalized model for mapping bicycle ridership with crowdsourced data. *Transportation Research Part C: Emerging Technologies*, *125*, 102981. https://doi.org/10.1016/J.TRC.2021.102981

Panczak, R., Berlin, C., Voorpostel, M., Zwahlen, M., & Egger, M. (2023). The Swiss neighbourhood index of socioeconomic position: update and re-validation. *Swiss Medical Weekly*, *153*, 40028. https://doi.org/10.57187/smw.2023.40028

Raturi, V., Hong, J., McArthur, D. P., & Livingston, M. (2021). The impact of privacy protection measures on the utility of crowdsourced cycling data. *Journal of Transport Geography*, *92*, 103020. https://doi.org/10.1016/J.JTRANGEO.2021.103020

Roy, A., Nelson, T. A., Fotheringham, A. S., Winters, M., & Edu, S. F. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science 2019, Vol. 3, Page 62*, *3*(2), 62. https://doi.org/10.3390/URBANSCI3020062

Saelens, B. E., Sallis, J. F., & Frank, L. D. (2003). Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine*, *25*(2), 80–91. https://doi.org/10.1207/S15324796ABM2502_03

Sallis, J. F., Conway, T. L., Dillon, L. I., Frank, L. D., Adams, M. A., Cain, K. L., & Saelens, B. E. (2013). Environmental and demographic correlates of bicycling. *Preventive Medicine*, *57*(5), 456–460. https://doi.org/10.1016/J.YPMED.2013.06.014

Sallis, J. F., Floyd, M. F., Rodríguez, D. A., & Saelens, B. E. (2012). Role of Built Environments in Physical Activity, Obesity, and Cardiovascular Disease. *Circulation*, *125*(5), 729–737. https://doi.org/10.1161/CIRCULATIONAHA.110.969022

Stadt Zuerich. (2022). *Mit freiwilligen Datenspenden gesellschaftlichen Mehrwert schaffen*. https://www.stadt-zuerich.ch/prd/de/index/ueber_das_departement/medien/medienmitteilungen/2022/august/220823a0.html

Stadt Zürich. (2020). *Vorlage 1: Volksinitiative «Sichere Velorouten für Zürich» - Stadt Zürich*. https://www.stadt-zuerich.ch/portal/de/index/politik_u_recht/abstimmungen_u_wahlen/archiv_abstimmungen/vergangene_termine/200927/200927-1.html

Stadt Zürich. (2021a). *Velostrategie 2030 Massnahmenband*. www.stadt-zuerich.ch/velo

Stadt Zürich. (2021b). *Velostrategie 2030. Mit dem Velo sicher und einfach durch Zürich*.

Stadt Zürich. (2022). *Automatische Zählungen des Veloverkehrs - Stadt Zürich*. https://www.stadt-zuerich.ch/ted/de/index/taz/verkehr/webartikel/webartikel_velozaehlungen.html

Strava. (2020a, February 4). *Strava Milestones: 50 Million Athletes and 3 Billion Activity Uploads*. https://blog.strava.com/press/strava-milestones-50-million-athletes-and-3-billion-activity-uploads/

Strava. (2020b, September 23). *Strava Metro FAQ*. https://metro.strava.com/faq

Sunde, E. (2019, February 8). *Tracking the rise of bike commuting around the world | by Erik Sunde | Strava Metro | Medium*. https://medium.com/strava-metro/tracking-the-rise-of-bike-commuting-around-the-world-5bada94585c5

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X

Tironi, M., & Valderrama, M. (2017). Unpacking a citizen self-tracking device: Smartness and idiocy in the accumulation of cycling mobility data. *Https://Doi.Org/10.1177/0263775817744781*, *36*(2), 294–312. https://doi.org/10.1177/0263775817744781

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2010). Built environment influences on healthy transportation choices: Bicycling versus driving. *Journal of Urban Health*, *87*(6), 969–993. https://doi.org/10.1007/S11524-010-9509-6/TABLES/6

Winters, M., & Teschke, K. (2010). Route Preferences among Adults in the near Market for Bicycling: Findings of the Cycling in Cities Study. *American Journal of Health Promotion*, *25*(1), 40–47. https://doi.org/10.4278/ajhp.081006-QUAN-236

Yang, Y., Wu, X., Zhou, P., Gou, Z., & Lu, Y. (2019). Towards a cycling-friendly city: An updated review of the associations between built environment and cycling behaviors (2007–2017). *Journal of Transport and Health*, *14*. https://doi.org/10.1016/J.JTH.2019.100613

8.1 Software

Esri Inc. (2021). *ArcGIS Pro* (Version 2.9.6). Esri Inc. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

QGIS.org (2023). QGIS Geographic Information System (Version 3.22 Bialowieza). QGIS Association. http://www.qgis.org

R Core Team (2023). R (Version 4.3.0): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

8.1.1 R Packages

Name and Version

Package	Version	Citation
BAMMtools	2.1.10	Rabosky et al. (2014)
base	4.3.0	R Core Team (2023)
car	3.1.2	Fox and Weisberg (2019)
corrr	0.4.4	Kuhn, Jackson, and Cimentada (2022)
data.table	1.14.8	Dowle and Srinivasan (2023)
fastDummies	1.6.3	Kaplan (2020)
ggrepel	0.9.3	Slowikowski (2023)
glmmTMB	1.1.7	Brooks et al. (2017)
glmnet	4.1.7	Friedman, Tibshirani, and Hastie (2010); Simon et al. (2011); Tay, Narasimhan, and Hastie (2023)
grateful	0.2.3	Rodriguez-Sanchez & Jackson (2023)
gridExtra	2.3	Auguie (2017)
igraph	1.4.3	Csardi and Nepusz (2006)
jtools	2.2.1	Long (2022)
leaflet	2.1.2	Cheng, Karambelkar, and Xie (2023)
lme4	1.1.33	Bates et al. (2015)
lwgeom	0.2.13	E. Pebesma (2023)
MASS	7.3.58.4	Venables and Ripley (2002)

Package	Version	Citation
mpath	0.4.2.23	Wang, Ma, Zappitelli, et al. (2014); Wang, Ma, Wang, et al. (2014); Wang, Ma, and Wang (2015); Wang (2019); Wang (2020); Wang (2022)
nngeo	0.4.7	Dorman (2023)
osmdata	0.2.5	Mark Padgham et al. (2017)
performance	0.10.4	Lüdecke et al. (2021)
progress	1.2.2	Csárdi and FitzJohn (2019)
randomForest	4.7.1.1	Liaw and Wiener (2002)
raster	3.6.20	Hijmans (2023a)
rmarkdown	2.22	Xie, Allaire, and Grolemund (2018); Xie, Dervieux, and Riederer (2020); Allaire et al. (2023)
scales	1.2.1	Wickham and Seidel (2022)
sf	1.0.13	E. Pebesma (2018); E. Pebesma and Bivand (2023)
sfnetworks	0.6.3	van der Meer et al. (2023)
sp	1.6.1	E. J. Pebesma and Bivand (2005); Bivand, Pebesma, and Gomez-Rubio (2013)
statmod	1.5.0	Dunn and Smyth (1996); Smyth (2002); Smyth (2005b); Smyth (2005a); Hu and Smyth (2009); Phipson and Smyth (2010); Giner and Smyth (2016)
terra	1.7.29	Hijmans (2023b)
tidygraph	1.2.3	Pedersen (2023)
tidyverse	2.0.0	Wickham et al. (2019)
tmap	3.3.3	Tennekes (2018)
units	0.8.2	E. Pebesma, Mailund, and Hiebert (2016)

Bibliography

Allaire, J. J., Xie, Yihui, Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., ... & Atkins, A. (2023). *rmarkdown: Dynamic Documents for R*. Retrieved from https://github.com/rstudio/rmarkdown

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. https://asdar-book.org/

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378-400. https://doi.org/10.32614/RJ-2017-066

Cheng, J., Karambelkar, B., & Xie, Y. (2023). *leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library*. Retrieved from https://CRAN.R-project.org/package=leaflet

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695. https://igraph.org

Csárdi, G., & FitzJohn, R. (2019). *progress: Terminal Progress Bars*. Retrieved from https://CRAN.R-project.org/package=progress

Dorman, M. (2023). *nngeo: K-Nearest Neighbor Join for Spatial Data*. Retrieved from https://CRAN.R-project.org/package=nngeo

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of "data.frame"*. Retrieved from https://CRAN.R-project.org/package=data.table

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist, 5*, 236-244

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression, Third*. Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22. https://doi.org/10.18637/jss.v033.i01

Giner, G., & Smyth, G. K. (2016). statmod: Probability calculations for the inverse Gaussian distribution. *R Journal, 8*(1), 339-351.

Hijmans, R. J. (2023a). *raster: Geographic Data Analysis and Modeling*. Retrieved from https://CRAN.R-project.org/package=raster

Hijmans, R. J. (2023b). *terra: Spatial Data Analysis*. Retrieved from https://CRAN.R-project.org/package=terra

Hu, Y., & Smyth, G. K. (2009). ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of Immunological Methods, 347*(1), 70-78.

Kaplan, J. (2020). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. Retrieved from https://CRAN.R-project.org/package=fastDummies

Kuhn, M., Jackson, S., & Cimentada, J. (2022). *corrr: Correlations in R*. Retrieved from https://CRAN.R-project.org/package=corrr

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22. https://CRAN.R-project.org/doc/Rnews/

Long, J. A. (2022). *jtools: Analysis and Presentation of Social Scientific Data*. Retrieved from https://cran.r-project.org/package=jtools

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Padgham, M., Rudis, B., Lovelace, R., & Salmon, M. (2017). Osmdata. *Journal of Open Source Software, 2*(14), 305. https://doi.org/10.21105/joss.00305

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal, 10*(1), 439-446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. (2023). *lwgeom: Bindings to Selected "liblwgeom" Functions for Simple Features*. Retrieved from https://CRAN.R-project.org/package=lwgeom

Pebesma, E., & Bivand, R. (2005). Classes and methods for spatial data in R. *R News, 5*(2), 9-13. https://CRAN.R-project.org/doc/Rnews/

Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://r-spatial.org/book/

Pebesma, E., Mailund, T., & Hiebert, J. (2016). Measurement units in R. *R Journal, 8*(2), 486-494. https://doi.org/10.32614/RJ-2016-061

Pedersen, T. L. (2023). *tidygraph: A Tidy API for Graph Manipulation*. Retrieved from https://CRAN.R-project.org/package=tidygraph

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology, 9*(1), Article 39.

Rabosky, D. L., Grundler, M. C,Anderson, C. J., Title, P. O., Shi, J. J., Brown, J. W., ... & Larson, J. G. (2014). BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution, 5*, 701-707.

Rodriguez-Sanchez, F., & Connor, J. P. (2023). _grateful: Facilitate citation of R packages_. Retrieved from https://pakillo.github.io/grateful/

Simon, N., Friedman, J., Tibshirani, R., & Hastie, T. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software, 39*(5), 1-13. https://doi.org/10.18637/jss.v039.i05

Slowikowski, K. (2023). *ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2"*. Retrieved from https://CRAN.R-project.org/package=ggrepel

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics, 11*, 836-847.

Smyth, G. K. (2005a). Numerical integration. In *Encyclopedia of Biostatistics* (pp. 3088-3095).

Smyth, G. K. (2005b). Optimization and nonlinear equations. In *Encyclopedia of Biostatistics* (pp. 3088-3095).

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software, 106*(1), 1-31. https://doi.org/10.18637/jss.v106.i01

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software, 84*(6), 1-39. https://doi.org/10.18637/jss.v084.i06 van der Meer, L., Abad, L., Gilardi, A., & Lovelace, R. (2023). *sfnetworks: Tidy Geospatial Networks*. Retrieved from https://CRAN.R-project.org/package=sfnetworks

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S, Fourth*. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Wang, Z. (2019). MM for penalized estimation. arXiv e-prints.

Wang, Z. (2020). Unified robust estimation. arXiv e-prints.

Wang, Z. (2022). *mpath: Regularized Linear Models*. Retrieved from https://CRAN.R-project.org/package=mpath

Wang, Z., Ma, S., & Wang, C. Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical Journal, 57*(5), 867-884.

Wang, Z., Ma, S., Wang, C. Y., Zappitelli, M., Parikh, C., Devarajan, P., & Parikh, C. R. (2014). EM for regularized zero inflated regression models with applications to postoperative morbidity after cardiac surgery in children. *Statistics in Medicine, 33*(29), 5192-5208.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C. Y., & Devarajan, P. (2014). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Statistical Methods in Medical Research*.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., ... & Riddell, A. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Seidel, D. (2022). *scales: Scale Functions for Visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman & Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown Cookbook*. Chapman & Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook

Appendix

A) Link to Github Repository

All necessary R-code to to conduct this thesis can be found on Github: https://github.com/toezve/geo511_modelling_bicycle_ridership.git

B) Counting Stations

City

Table 20: List of Counting Stations operated by the City of Zurich

FK_Zaehler (primary key)	Street	OBJECTID AV Network	Veloinfrastruktur	Correction Factor	Months of measurement	Remarks
Y2H19090841	Andreasstrasse	70185	Fuss/Veloweg	1.14	full	
Y2H20063173	Baslerstrasse	62169	30er Zone mit Velostreifen	1.00	full	
ECO09113506	Bertastrasse	62208	30er Zone	1.27	01/2021 – 08/2021	Until 21/09/2021

ECO09113500	Binzmühlestrasse	67710	50er Zone mit Velostreifen	1.22	01/2021 – 05/2021;
					07/2021 – 11/2021;
					01/2022 – 02/2022;
					05/2022	

Y2H20114444	Bucheggplatz Richtung Höngg, Hofwiesenstr.	100010 added	Velostreifen	1.27	03/2021 – 12/2022	Counts two diverging traffic lanes
Y2H16069943	Hardbrücke HB	68732	Fuss/Velobrücke	1.03	full	summed
Y2H16069942	Hardbrücke Seite Altstetten	64741		1.10	full	
ECO09113507	Hofwiesenstr Richtung Bucheggplatz	65149 (ganze Str)	Velostreifen Hauptstrasse	1.28	01/2021 – 02/2022; 06/2022 – 12/2022	Only one directional
Y2H21056106	Langstrasse (Fahrbahn Nord)	69248 (ganze Querschnitt)	Velostreifen Hauptstrasse	1	03/2022 – 12/2022	summed
Y2H21056105	Langstrasse (Fahrbahn Süd)	70010 Ganzer Querschnitt	Velostreifen Hauptstrasse	1		
Y2H19101255	Langstrasse (Unterführung Nord)	70010 Ganzer Querschnitt	Separierter Veloweg	0.96		
Y2H21111102	Langstrasse (Unterführung Süd)	70010 Ganzer Querschnitt	Separierter Veloweg	1		
Y2H19070585	Limmatquai -> Bellevue	67630/64037	30er Zone	1.47	full	summed

Y2H20083483	Limmatquai -> Central	64037	30er Zone	1.33	full	summed
Y2H19101198	Lux-Guyer_Weg Oberer Letten	65866	Fuss/Veloweg	1	full	
Y2G13124879	Militärbrücke	62546	Fuss/Velobrücke	0.95	full	
Y2H18106792	Mühlebachstrasse	66294	30er Zone	1.19	full	
Y2H22073807	Mythenquai	63293	Fuss/Veloweg	1.00	12/2022	
ECO09113499	Mythenquai	63293	Fuss/Veloweg	1.20	01/2021-01/2022	Until 03/03/2022
Y2H20011946	Saumackerstrasse	61733	50er Zone ohne Velostreifen	1.27	04/2021 – 05/2021; 11/2021 – 04/2022	
Y2H19111477	Scheuchzerstrasse	65622	30er Zone	1.05	01/2021 – 05/2021; 08/2021 – 12/2022	
ECO10053914	Schulstrasse	68239	30er Zone	1.43	full	

ECO09113503	Sihlpromenade	69378	Fuss/Veloweg beidseitig	1.09	01/2021 – 05/2021;
					07/2021 – 08/2021;
					11/2021;
					02/2022;
					05/2022 – 07/2022;
					09/2022	

Y2H19111476	Talstrasse	66542	50er Zone mit Velostreifen	1.35	12/2021 – 12/2022	Leap in counts in 12/2022
Y2H19070283	Zollstrasse HB	64179	Velostreifen (eng)	1.49	01/2021 – 03/2022	Until 03/2022

Canton

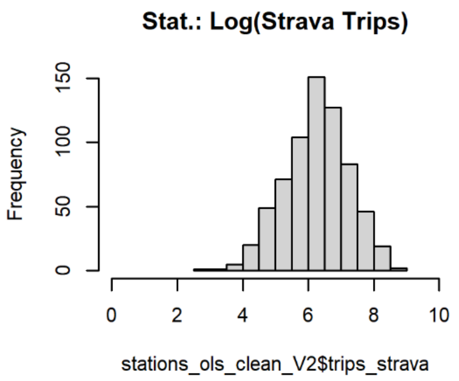
Table 21: List of Counting Stations operated by the Canton of Zurich

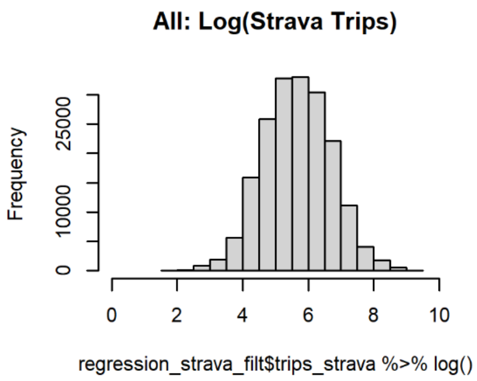
NR	Street	OBJECTID AV Network	Type Route	Veloinfrastruktur	Months of measurement	Remarks
421	Badenerstrasse Schlieren	20257	Alltag		2022	Summed
521	Badenerstrasse Schlieren	20257				
2221	Fischerweg, Dübendorf	31197		Fuss-Veloweg	05/2022 – 12/2022	

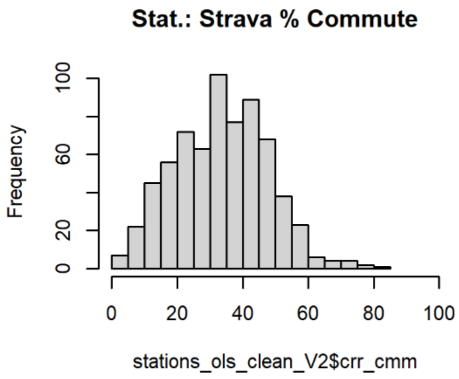
918	Gaswerkstrasse, Schlieren	20417	Hauptverbindung, Alltag	Nur LV oder landwirt.	alle	
2721	Glattuferweg Süd, Opfikon	93246	Nebenverbindung, Alltag	Fuss-Veloweg	10/2021 – 04/2022	
2521	Glattuferweg Zunstrasse Opfikon	93226	Zusätzliche Freizeitverbindung	Strasse und Trottoir	10/2021 – 12/2022	
2621	Glattuferweg Opfikon	93207		Fuss-Veloweg	11/2021 – 12/2022	
1519	Opfikon	92895	Nebenverbindung, Alltag	Hauptstrasse mit Velostreifen	01/2021 – 02/2022 06/2022 – 12/2022	Summed
5019	Schaffhauserstrasse Glattbrugg	92895				
1819	Schaffhauserstrasse, Kloten	55714	Nebenverbindung, Alltag	Hauptstrasse mit Velostreifen	alle	
619	Tobelhofstrasse Dübendorf	70609	Nebenverbindung, Alltag	Veloweg	11/2021 – 12/2022	Not urban
221	Ueberlandstrasse Dietikon West	61004	Alltag	Veloweg	10/2021 – 12/2022	
2020	Dietikon Brücke	61704	Veloschnellroute/Nebenverbindung Alltag	Veloweg auf Brücke, (einseitig Fuss/Velo)	alle	Summed
5120	Dietikon Brücke	61704				
1018	Dietikon Überlandstrasse Ost	61092	Veloschnellroute, Alltag	Fuss/Veloweg an 60er Zone	alle	
818	Vulkanstrasse,Schlieren	20538	Veloschnellroute, Alltag	Veloweg ohne MIV	alle	
716	Regensdorf West	55154	Nebenverbindung, Alltag	Fuss/Veloweg	alle	Not urban
616	Regensdorf Ost	54606	Nebenverbindung, Alltag	Fuss/Veloweg	alle	Not urban
1121	Zürcherstr Dietikon	60804	Nebenverbindung, Alltag	Velostreifen	2022	Summed
1021	Dietikon	60804				
2119	Schlieren, Zürcherstrasse	20016	Haupt&Nebenverbindung, Alltag	Velostreifen	01/2022 – 06/2021; 05/2022 – 12/2022	

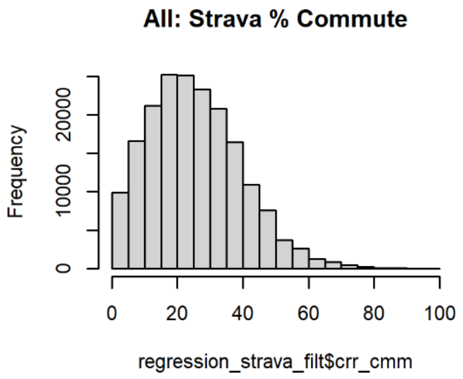
2421	Glattuferweg Zürich	70803		Fuss-Veloweg	alle	
919	Zürich Witikon	70971 Ref. edge 69693 real edge	Hauptverbindung	Veloweg	alle	Not urban
317	Bassersdorf	9486	Nebenverbindung	Fuss-Veloweg	alle	Not urban
2219	Basserdorf Feuerwehr	9597		Fuss-Veloweg	alle	
417	Kloten	9106	Nebenverbindung	Fuss-Veloweg	alle	

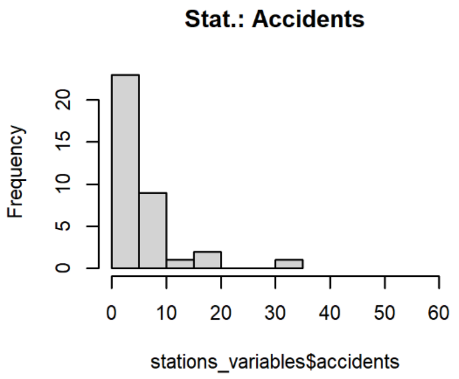
C) Histograms of all Variables

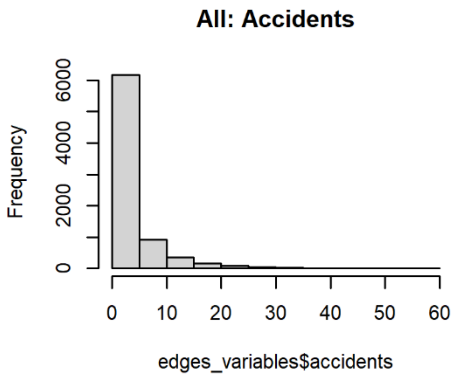


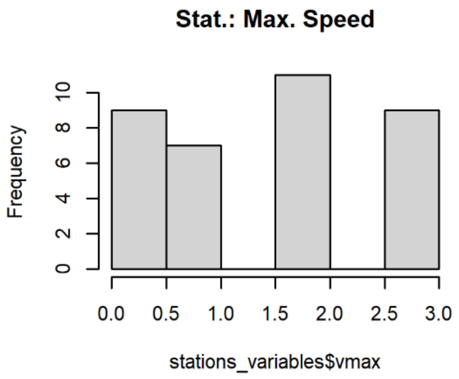


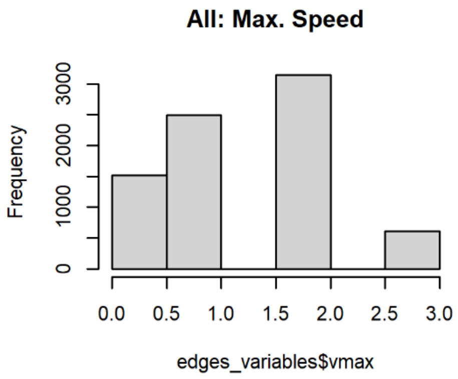


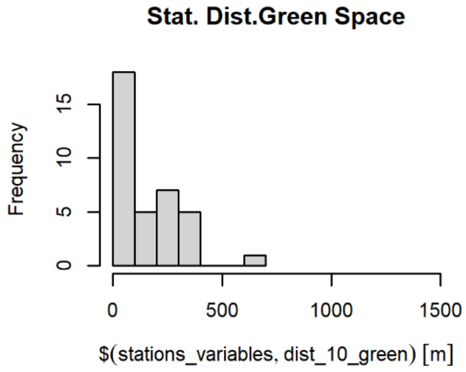


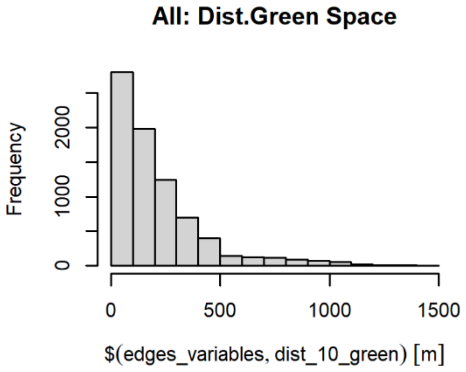


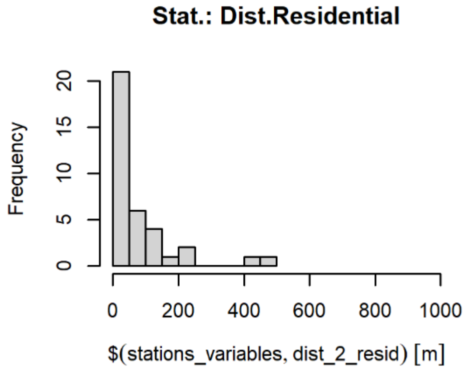


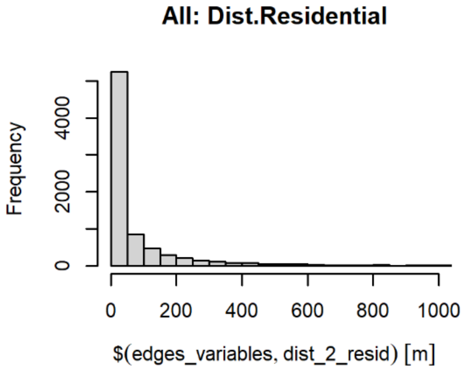


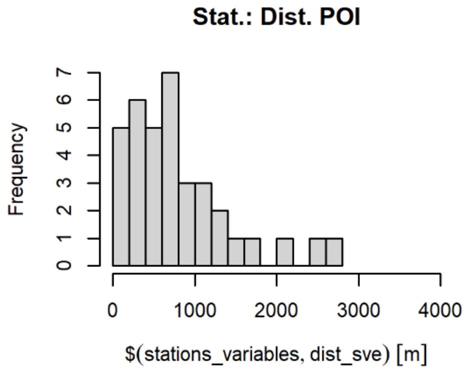


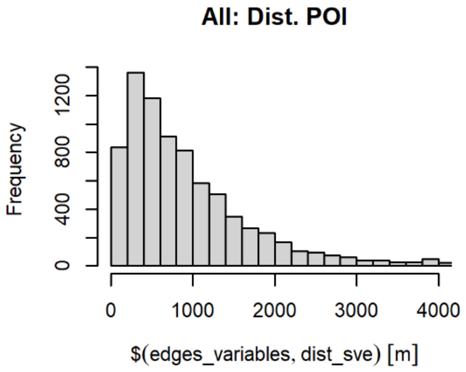

Stat. Dist.Green Space
Frequency
0
5
10
15
0
500
1000
1500
$(stations_variables, dist_10_green) [m]

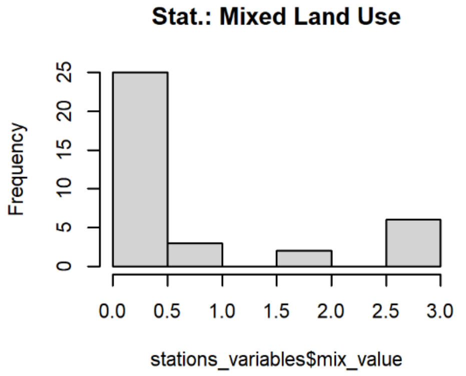

All: Dist.Green Space
Frequency
0
1000
2000
0
500
1000
1500
$(edges_variables, dist_10_green) [m]

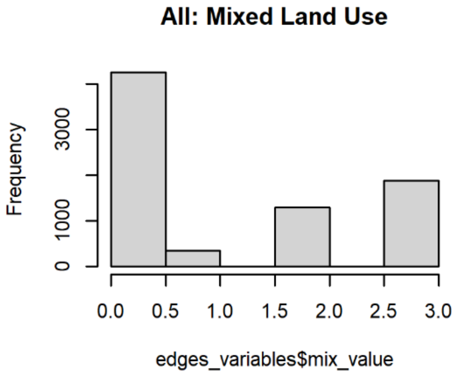

Stat.: Dist.Residential
Frequency
0
5
10
15
20
0
200
400
600
800
1000
$(stations_variables, dist_2_resid) [m]


All: Dist.Residential
Frequency
0
2000
4000
0
200
400
600
800
1000
$(edges_variables, dist_2_resid) [m]


Stat.: Dist. POI
Frequency
0
1
2
3
4
5
6
7
0
1000
2000
3000
4000
$(stations_variables, dist_sve) [m]


All: Dist. POI
Frequency
0
400
800
1200
0
1000
2000
3000
4000
$(edges_variables, dist_sve) [m]


Stat.: Mixed Land Use
Frequency
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
stations_variables$mix_value


All: Mixed Land Use
Frequency
0
1000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
edges_variables$mix_value

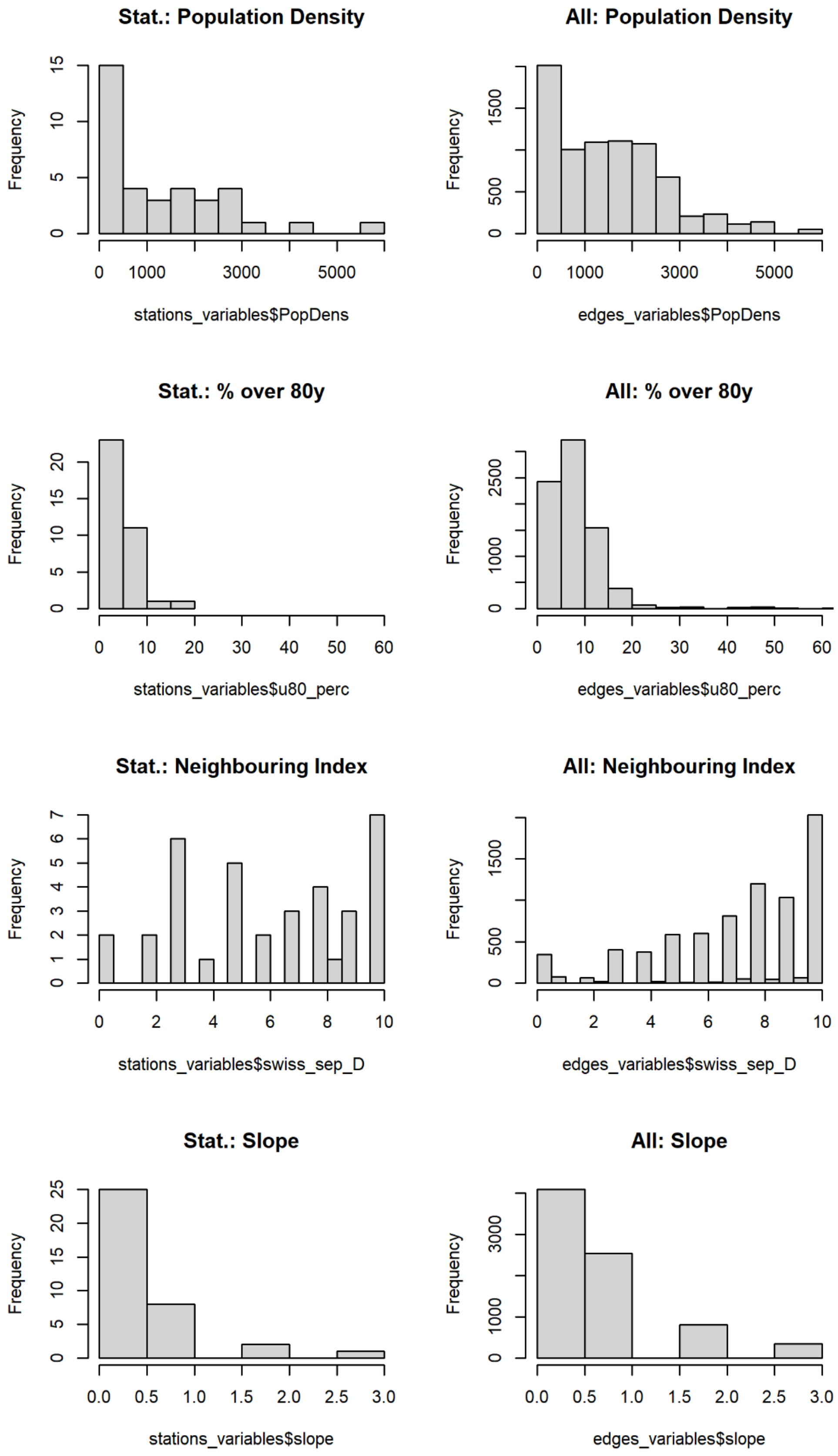
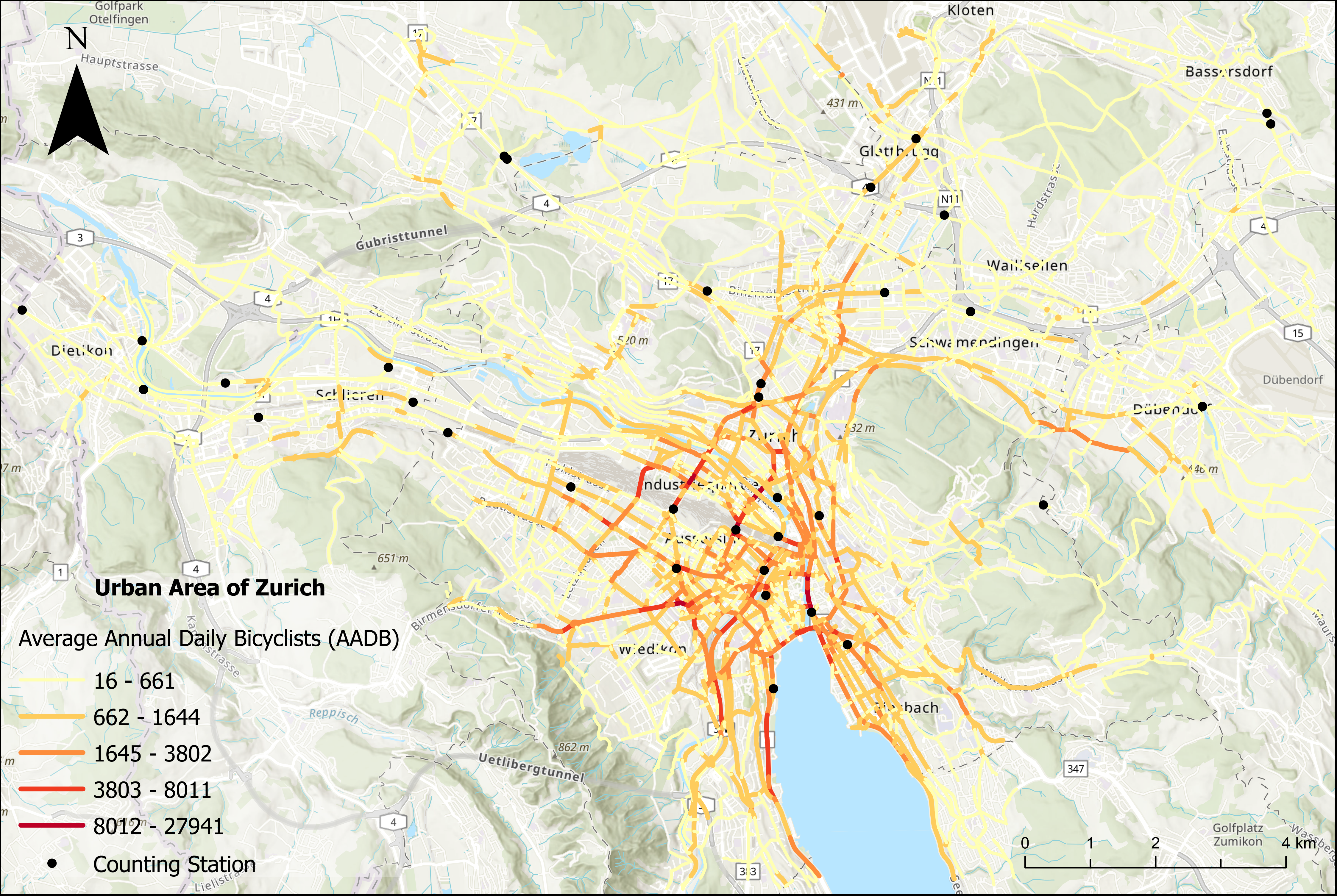


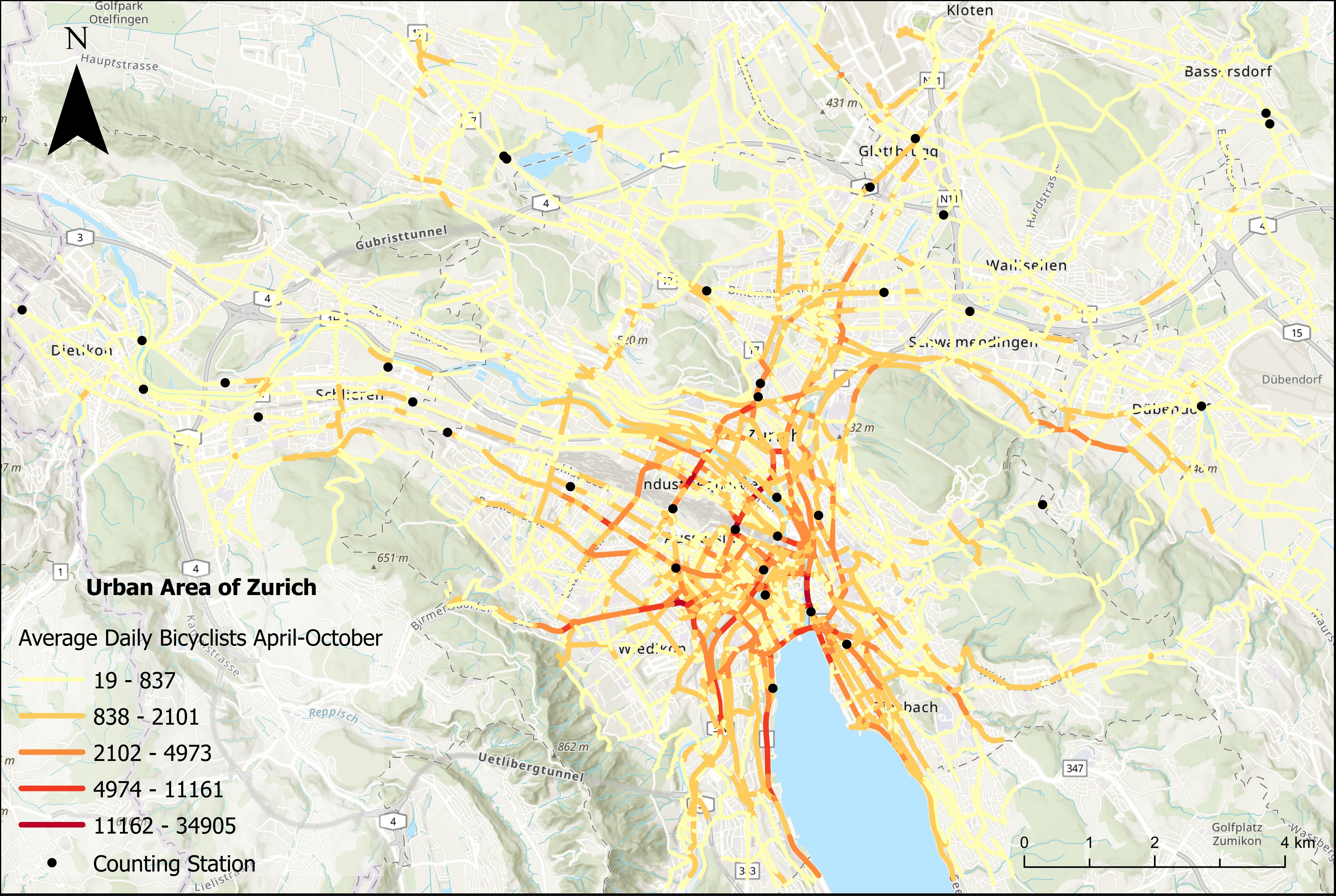
Figure 31: Histograms of all Variables.
Note: Stat. refers to the Counting Stations, while "All" refers to all (7760) street edges of the network.

D) Categorical Maps



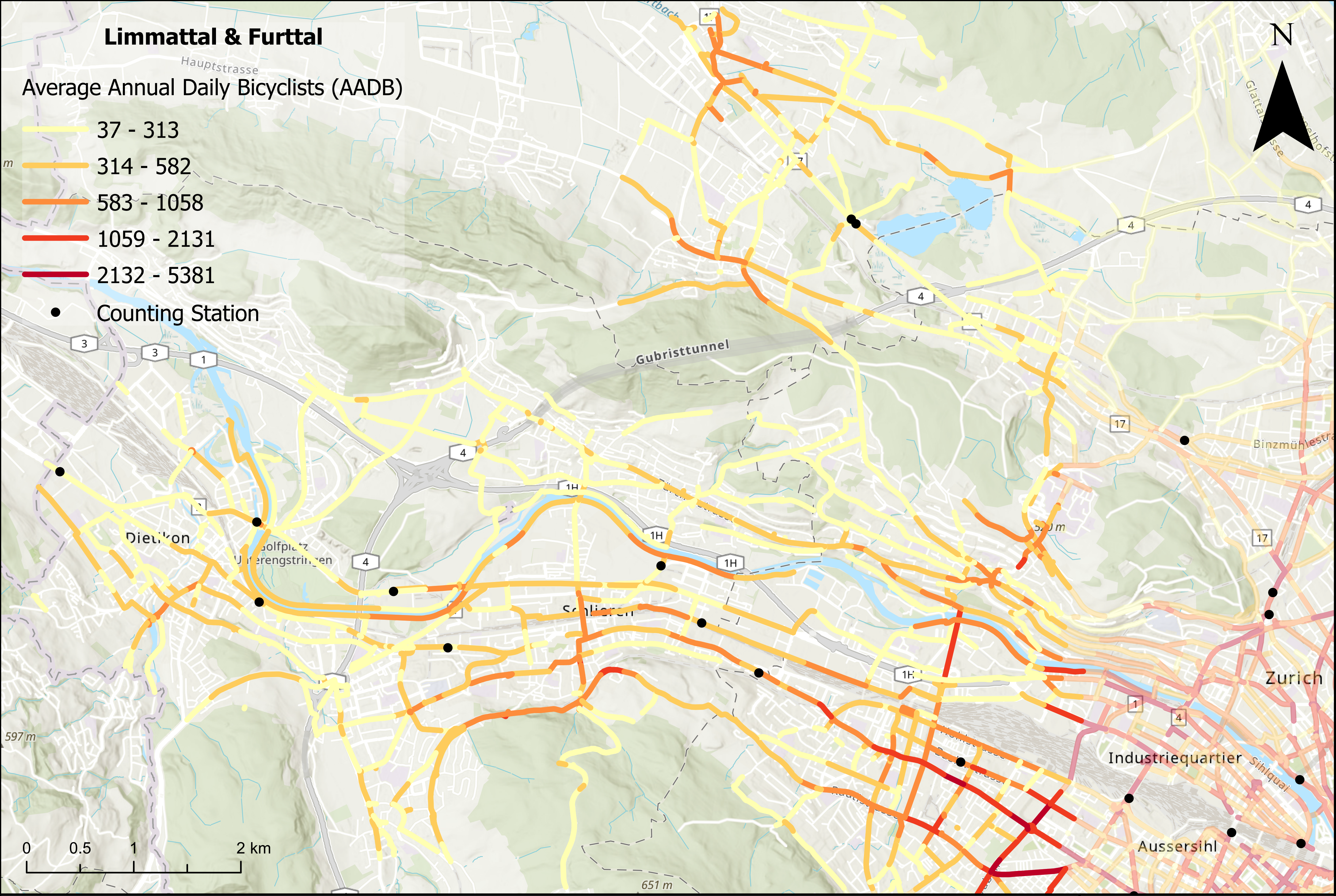
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



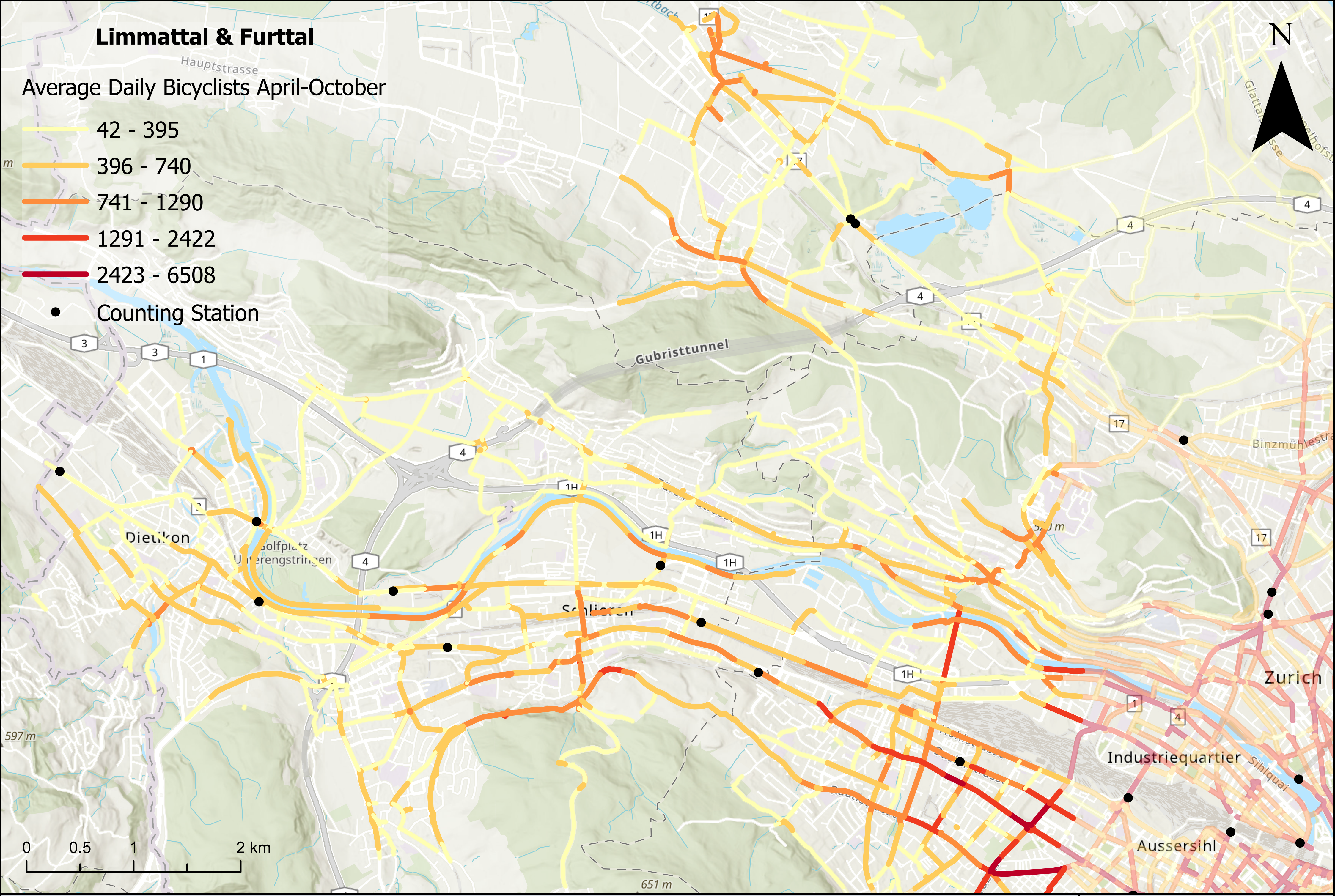
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



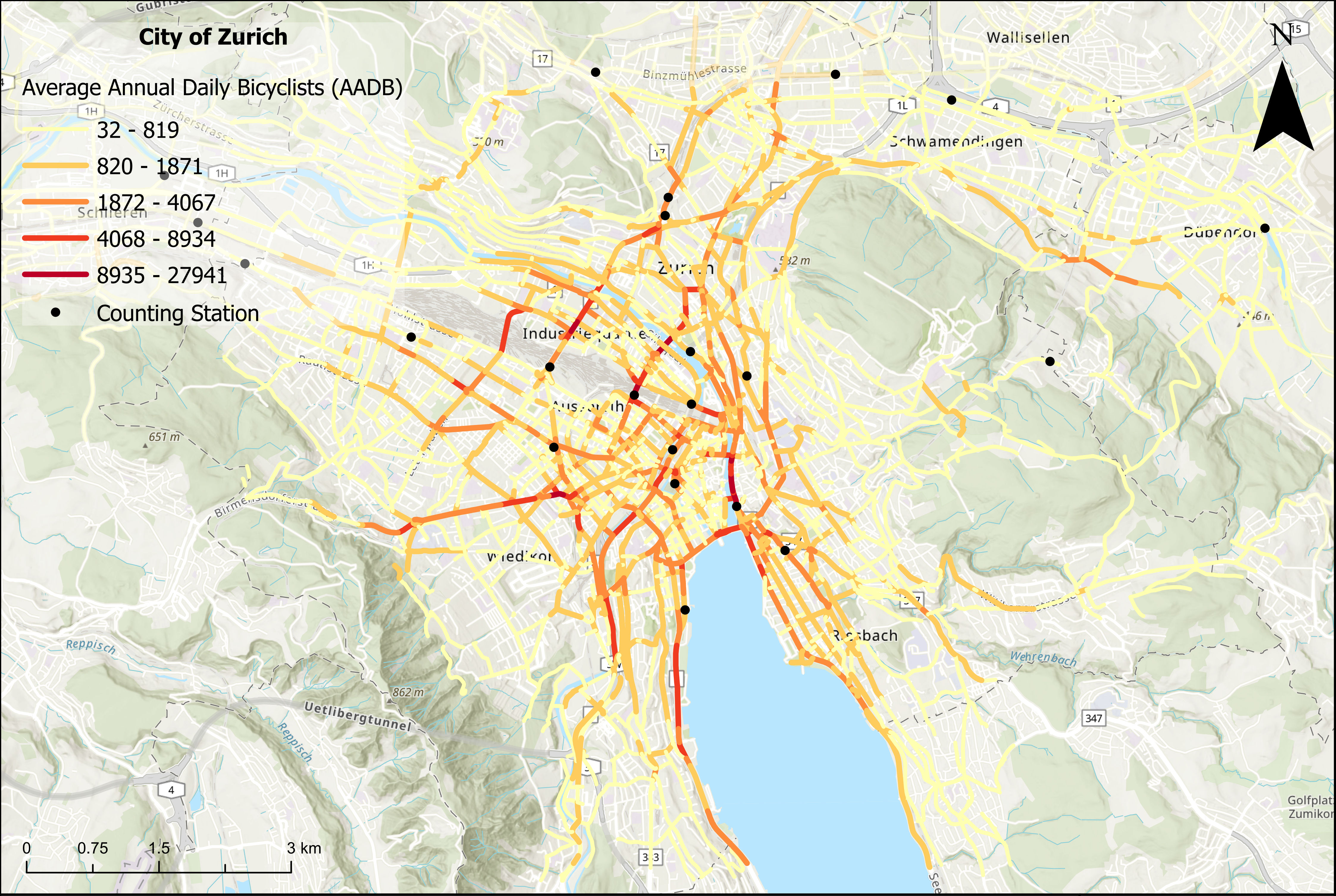
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



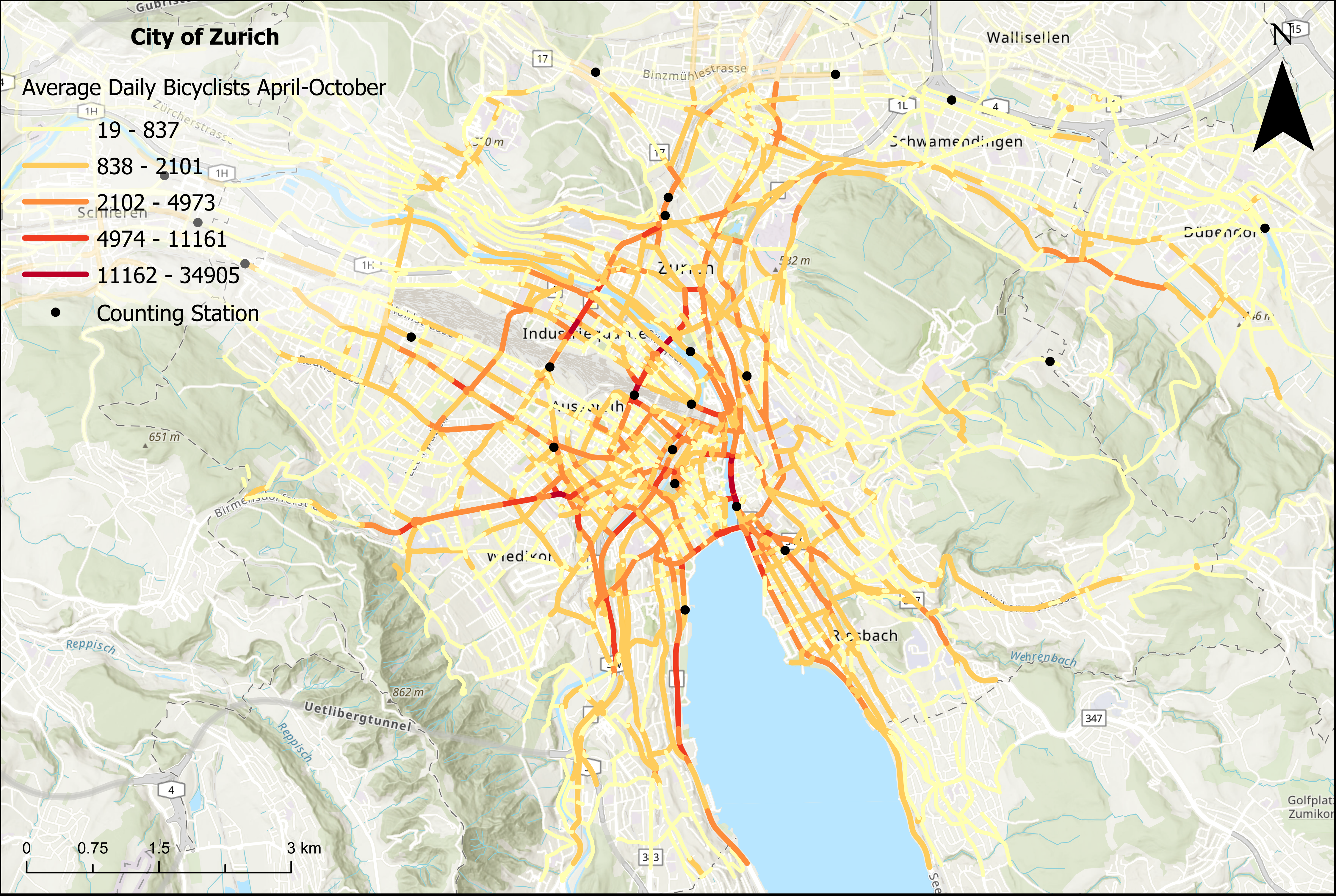
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap

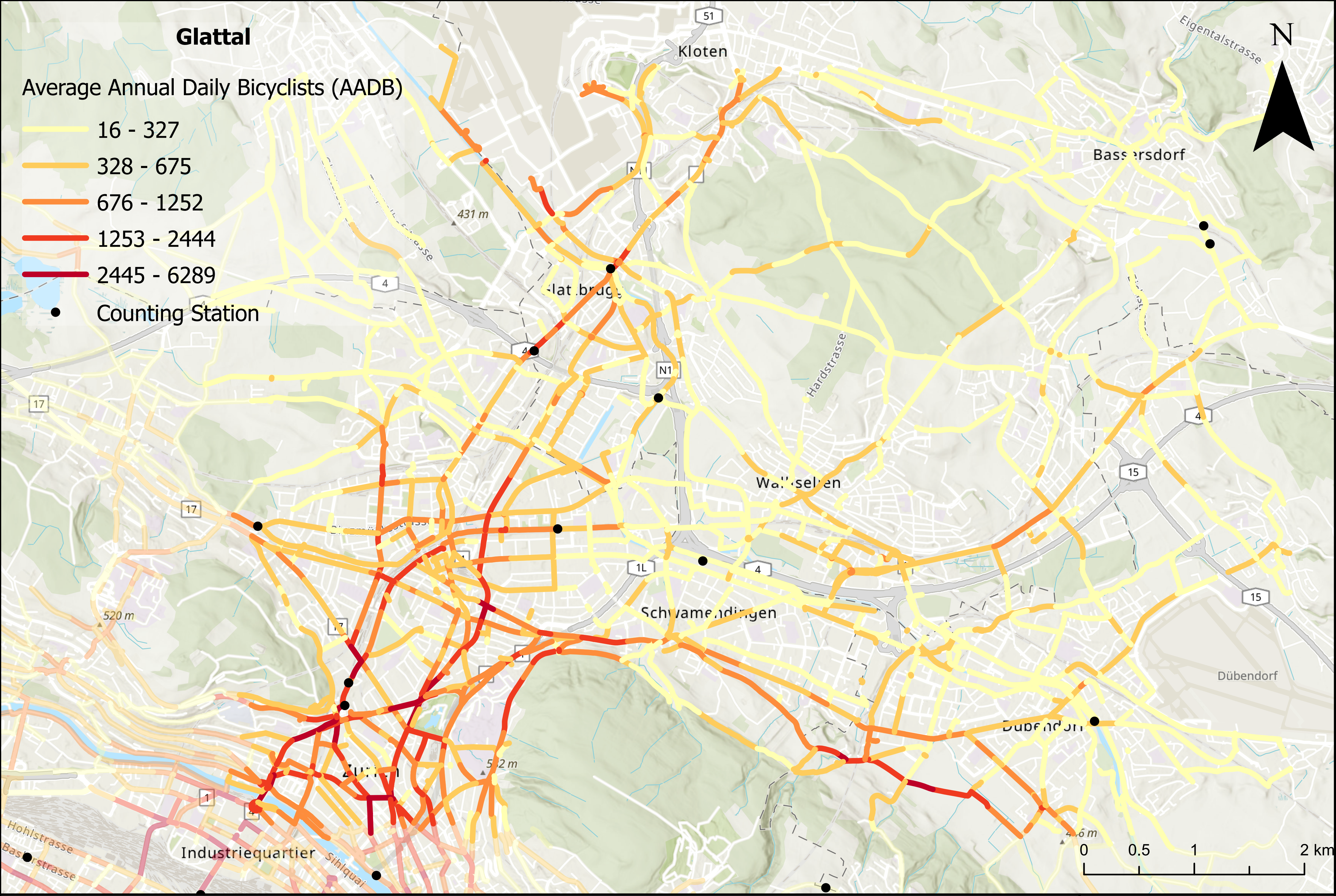


City of Zurich

Average Daily Bicyclists April-October

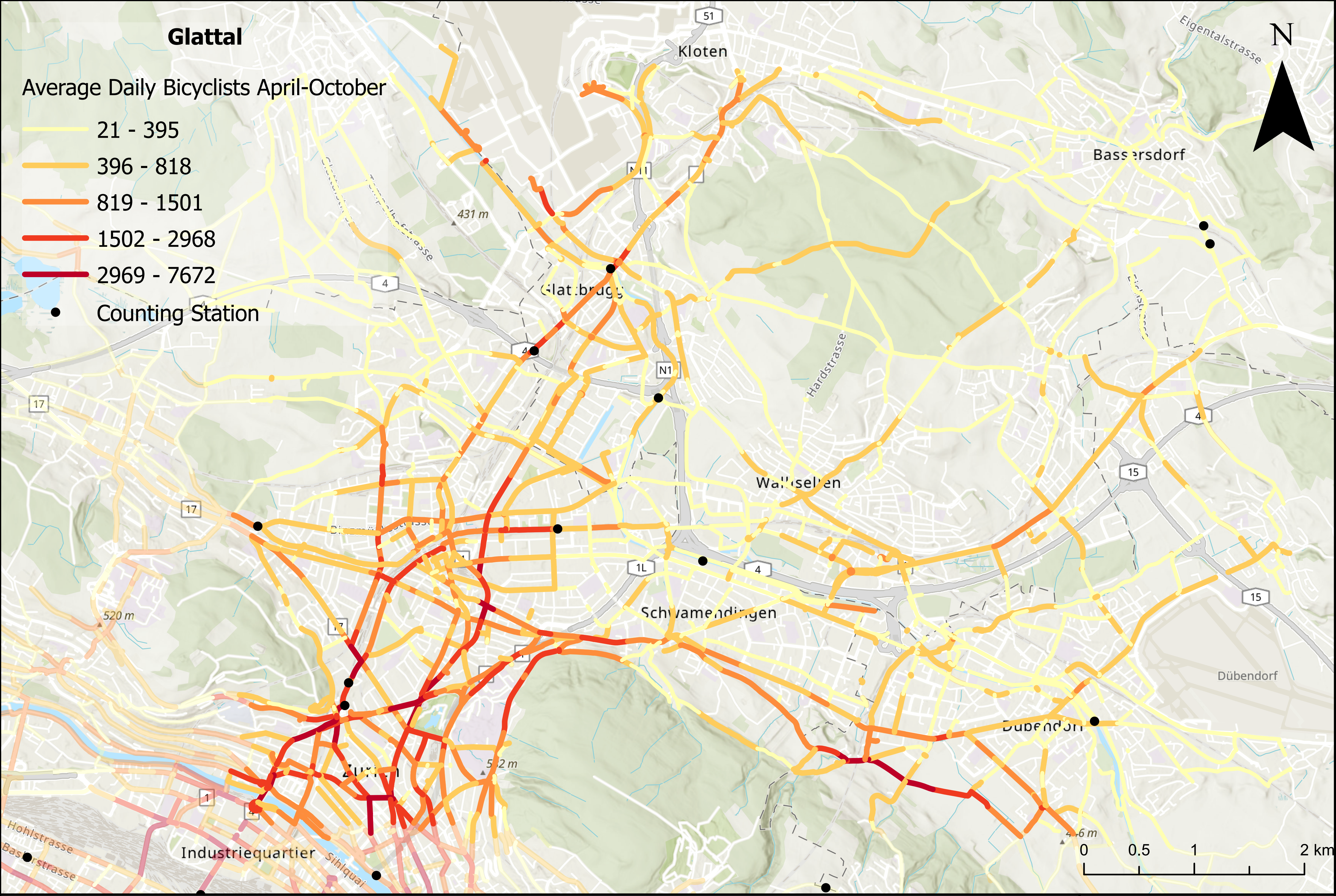
MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap



MSc Thesis: "Modelling Bicycle Ridership using Crowdsourced Data in the Urban Area of Zurich", Geo 511, University of Zurich
27.08.2023; ©Thomas Özvegyi; Contact: t.oezvegyi@gmx.ch

Sources: ESRI, Strava, Kanton Zürich, Stadt Zürich, Bundesamt für Statistik, swisstopo, OpenStreetMap

Personal Declaration

I hereby declare that the submitted thesis is result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Thomas Özvegyi, August 2023