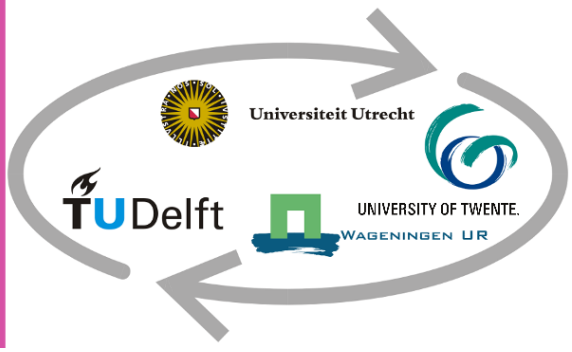
GIMA
Geographical Information Management and Applications

Estimating traffic intensity of cyclists using flow interpolation

Joachim Jochemsen
j.j.jochemsen@students.uu.nl

Supervisor: Simon Scheider

Professor: Stan Geertman



June 6, 2018

Estimating traffic intensity of cyclists using flow interpolation

Author:
Joachim Jochemsen

Student number:
s6030874 (UTwente), 3845265 (UU)

Date:
June 6, 2018

Abstract

Literature shows that there is still a lack of objective, quantitative information about cycling traffic for urban researchers and planners. At the moment, gravity models are the most standard models for traffic prediction. However, these models have some difficulties fitting to local measurements. Machine learning algorithms can fit to local measurements very well and have become viable in recent years due to the increase in computing power of modern hardware. Induction loops and manual counting by hand have traditionally been often used methods for (cycling) traffic counting and traffic data collection. Due to the increase in smartphone usage, GPS data has become a viable alternative to these traditional methods in recent years. Therefore, there lies a lot of potential in combining new and upcoming data sources for traffic information such as GPS with more conventional data sources such as traffic counts using machine learning. This research aims to investigate if and how cyclist traffic intensity can be estimated using machine learning algorithms to combine GPS tracks and local traffic counts. The municipality of Tilburg in the Netherlands is selected as the scope for this research, because of the availability of data for this area.

Possible input features for the machine learning algorithms were based on literature. GPS cyclist intensities, spatial distance, road surface type, the width of roads and attractivity were found to be suitable input features. The Fietsersbond network was chosen over the OpenStreetMap network, since the former contained most of the possible input features. GPS tracks from the B-Riders and Fietstelweek were mapmatched to the network to provide traffic intensities to combine with the traffic counts, which are based on the 'Fietstelprogramma Gemeente Tilburg'. Seven different machine learning regression algorithms are tested, and their outcomes were assessed using K-fold and Leave-one-out cross-validation.

There was found to be very little correlation between the GPS cyclist intensities and the traffic counts. The outcomes of every single machine learning algorithm show that the B-Riders and Fietstelweek GPS data are unsuitable for estimating the traffic intensity of cyclists using flow interpolation since all of them have r2 scores that are zero or negative. Future research, using more extensive and less biased GPS data samples, could provide further insight into the possibilities of combining GPS tracks and local traffic counts to estimate cyclists traffic intensity on a network.

Preface

Before you lies my thesis, which was written over a period of ten months for the MSc programme Geographical Information Management and Applications(GIMA). There are several people that have helped me greatly during the process of writing this thesis and whom I would like to thank.

Firstly, I would like to thank my supervisor Simon Scheider. His feedback and suggestions during our numerous meeting helped me very much during the course of my thesis and kept me motivated during the several struggles I had along the way.

Next, I need to thank the NHTV in Breda, and more specifically Joost de Kruijf for introducing me to the subject and giving me the opportunity to work with some exiting data. I also would like to thank Dirk Bussche for assisting me with the mapmatching of the GPS data and answering all the questions that I had regarding the data.

A lot of hard work was put into this thesis over the last ten months, during which I learned a lot. I hope that you enjoy reading it and learn something as well.

Joachim Jochemsen

Bennekom, 6 June 2018

Contents

List of Figures

List of Tables

List of abbreviations

DT	Decision Tree
GP	Gaussian Processes
GPS	Global Positioning System
KNN	k-Nearest Neighbor
KR	Kernel Ridge
LOOCV	Leave-one-out cross-validation
LpOCV	Leave-p-out cross-validation
LR	Linear Regression
MAE	Mean Absolute Error
MSE	Mean Square Error
OSM	OpenStreetMap
POI	Point of Interest
PLS	Partial Least Squares
RMSE	Root Mean Square Error
SV	Support Vector
TNO	Netherlands Organisation for Applied Scientific Research

Introduction

1

1.1 Context

The Netherlands is known for being one of the leading countries in the world regarding cycling. Currently, the number of bicycle trips that are taken each year in The Netherlands exceeds four billion. This amount accounts for approximately 27 percent of all trips made in a given year in the Netherlands. This percentage has remained relatively stable over the years (Klinkenberg and Bertolini, 2012).

Many organizations in the Netherlands realize the importance of cycling. For example, inspired by the Tour de France start in Utrecht in 2015, CROW-Fietsberaad (2017) began a collaboration with nearly all organizations in the Netherlands responsible for cycling policies. These organizations include authorities such as municipalities and provinces, civil society organizations such as the ANWB and the organizations such as Nederlandse Spoorwegen(NS). The collaboration is titled 'Tour de Force 2020', and recently a Bicycle Agenda 2017-2020 was released. This agenda had eight main goals, which they hope should lead to a 20 percent increase in the number of kilometers cycled between 2017 and 2020.

In addition to all the parties involved in such projects, even the new coalition agreement of the Dutch government mentions the importance of cycling for the Netherlands and suggests that extra budget will be made available to improve the cycling infrastructure during the next four years (Rijksoverheid, 2017).

To be able to support policies regarding bicycles, policymakers need to have information about the (cycling) traffic intensity in a particular area. According to Zhu and Levinson (2015), any new infrastructural initiatives or policies need to be built on reliable and precise traffic flow and travel time predictions.

Originally, the most common way for policymakers to get accurate traffic numbers is induction loops(in Dutch: tellus) which are installed on or beneath the roads on road intersections. In short, when traffic drives or cycles over these induction loops, the magnetic field is altered, which is then detected by the induction loop. Another method that is used to gain insight into traffic intensity is by performing manual counting at a specific road intersection. Those methods, however, are limited, and often do not give a complete and accurate representation of the traffic situation. While traffic counts are complete regarding flow measurement, their weak point is that they are not complete in terms of spatial distance. This is because when doing traffic counts like this, the traffic is counted for a specific location only and the route each counted cyclist has taken is not measured or taken into account.

As an upcoming alternative to counting traffic via induction loops, GPS data has recently become a viable data source for traffic research. In the last ten years, the volume of GPS data has risen significantly. The cause of this rise is the so-called 'smartphone revolution'. While in 2009 smartphones accounted for approximately 15 percent of the total number of phones, in 2014 the share of smartphones amongst all phones was over 35 percent (Romanillos et al., 2016). Compared to traffic counts, GPS tracks are spatially extended, but also incomplete and biased. This is because GPS tracks are almost never accurate representations of the behavior of all cyclists. Therefore combining traffics counts and GPS tracks is a solution that can lead to data of higher quality.

1.2 Problem statement and relevance

At the moment, the sustainability and livability of vulnerable neighborhoods in cities are increasingly getting under pressure, according to Uyterlinde and van der Velden (2017). Since the 1970's, there has been an increasing focus and investments in the bicycle, to improve both the sustainability and livability in cities (CROW-Fietsberaad, 2015).

Currently, urban researchers and planners still experience a lack of objective, quantitative information about bicycle traffic. Klinkenberg and Bertolini (2012) noted that, at the time of writing their paper in 2012, fundamental and scientific bicycle research was still in an early stage in the Netherlands, and that a disconnect might exist between research and current bicycle policies. They also note that there will be a payoff for effective bicycle policies, for instance in accessibility, the economy, but also in national identity. Because of this, there is a need for high-quality quantitative cycling information to help support bicycle policies. For example, better information about traffic flow can assist in aligning traffic lights. Since substantial structural investments are needed to increase and improve the cycling infrastructure in the Netherlands (Fietsersbond, 2017a), the importance of this kind of data has grown over the years. Without sufficient information about cycling traffic intensity and flows, bicycle policies may not be based on a detailed picture of the current state of traffic.

As mentioned in the previous section, GPS tracking has become a viable alternative for determining traffic flows, not just in the Netherlands, but also in other parts of the world. A relevant example of this is the B-Riders project in Noord-Brabant. This is a project aimed at stimulating bicycle usage instead of car usage. While traveling by bicycle, the users track their route via the special B-riders app on their smartphone (B-Riders, 2017). The resulting GPS tracks are then used for research on cyclists traffic flow and behavior. Another example of how the rise and popularity of GPS data can help cycling policies is the national Dutch 'Fietstelweek'. Each year during this week, cyclists are asked to register their cycling trips via a special app. The resulting tracks help in mapping and understanding the cycling network of the Netherlands and how it can be improved (Fietstelweek, 2017).

This was also noticed by TNO(Netherlands Organisation for Applied Scientific Research). In one of their research projects called the 'Fietsmonitor Zuid', they aimed at providing a better insight into the traffic flow of cyclists by combining several data sources, including GPS (Netherlands Organisation for Applied Scientific Research, 2014). Figure 1.1 shows how combining data and GPS tracks can help create new and better data about traffic intensity.

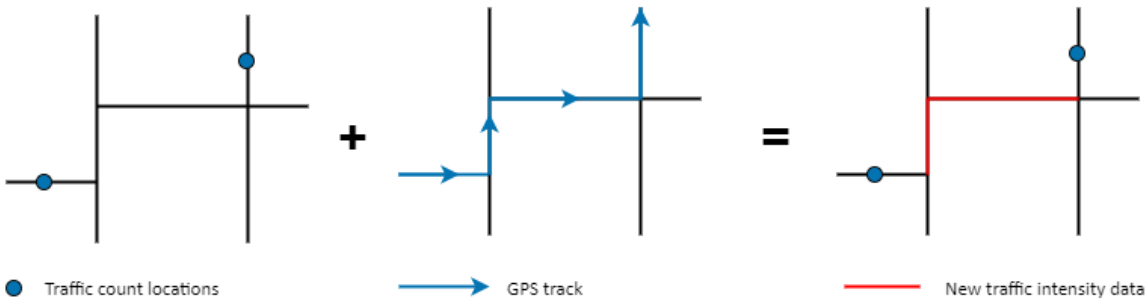


Fig. 1.1.: Schematic overview of the added value of combining traditional traffic counts and GPS tracks

The leftmost figure shows two existing traffic counts located on two different roads, while the middle figure shows a GPS track along with several different roads, including those with traffic counts. Finally, the third figure shows that traffic intensity of roads that do not have traffic counts can be obtained by combining the known traffic counts with the GPS tracks.

Currently, most standard traffic prediction approaches use gravity models. The basic idea of gravity models is that when the importance of two locations increases, the number of movement (or trips) between them also increases. The larger the distance between the two locations, however, the lower the amount of movements. A downside of these gravity models is that they are difficult to fit to local measurements (Anderson, 2010). An alternative to these traditional gravity-based models is machine learning, which can fit to local measurements very well. Partly due to the continuing increase in computing power of everyday hardware, this has become a viable alternative.

To conclude, there lies a lot of potential in combining new and upcoming data sources for traffic information such as GPS with more traditional data sources such as traffic counts. By using machine learning instead of traditional gravity models to combine this data, high-quality quantitative cycling information may be provided to help support bicycle policies. Therefore, this thesis aims at investigating how machine learning can be of use in combining GPS and traffic count data to gain high-quality cycling information in the form of traffic intensity.

1.3 Research questions

As mentioned in the previous section, this thesis aims to find, test and compare methods that can estimate the traffic intensity of cyclists across a network. To do this, the following main research question needs to be answered:

How can the traffic intensity of cyclists on a network be estimated by means of flow interpolation from local traffic counts and GPS tracks?

To support the answering of the main research objectives, several sub-questions have been defined, each focusing on a specific aspect of the research. The sub-questions that will be answered are:

- *Which machine learning methods are suitable for estimating the traffic intensity of cyclists across a network based on local traffic counts and GPS tracks?*
- *Which road characteristics should be taken into account as explanatory features for a cyclist traffic model?*
- *To what extent are biased GPS tracks useful as a variable in a cyclist traffic model?*
- *How can such a cyclist traffic model best be validated on local flow traffic count measurements?*

Chapter 3 explains the methodology that will be used to answer these research questions.

1.4 Research scope

In addition to clearly defining the Research Questions, it is also important to mention what this research is not about. Doing this early on in the research prevents potential scope creep and helps to make sure the research fulfills its objectives within the specified timeframe. The scope and several research limitations for this research are:

- This research will focus on the area of the municipality of Tilburg in Noord-Brabant, The Netherlands. One reason for this is the availability of cycling data for this area. Another reason is that limiting the research area also helps in keeping the research manageable regarding computing time. Finally, keeping the research area relatively small makes it easier to check the data for any errors or inconsistencies, something which is almost impossible with larger datasets.
- This research will not provide a whole functioning traffic model that can instantly be used by policymakers. Rather, it should be regarded as an exploration of methods which one can build upon to provide suitable solutions for different scenarios.

1.5 Reading guide

This section gives a brief explanation of the structure of this thesis. The following chapter 2 shortly examines the existing literature regarding cycling in the Netherlands, traffic data collection and machine learning and gives a theoretical framework. Chapter 3 then explains the used methodology based on the theoretical framework and conceptual model established in chapter 2. Afterward, chapter 4 details how the data is prepared for analysis. Next, chapter 5 contains the analysis and results. Finally, in chapter 6 a conclusion will be drawn and the research questions will be answered based on the results. This chapter also includes a discussion and suggestions for future research.

Theoretical Background

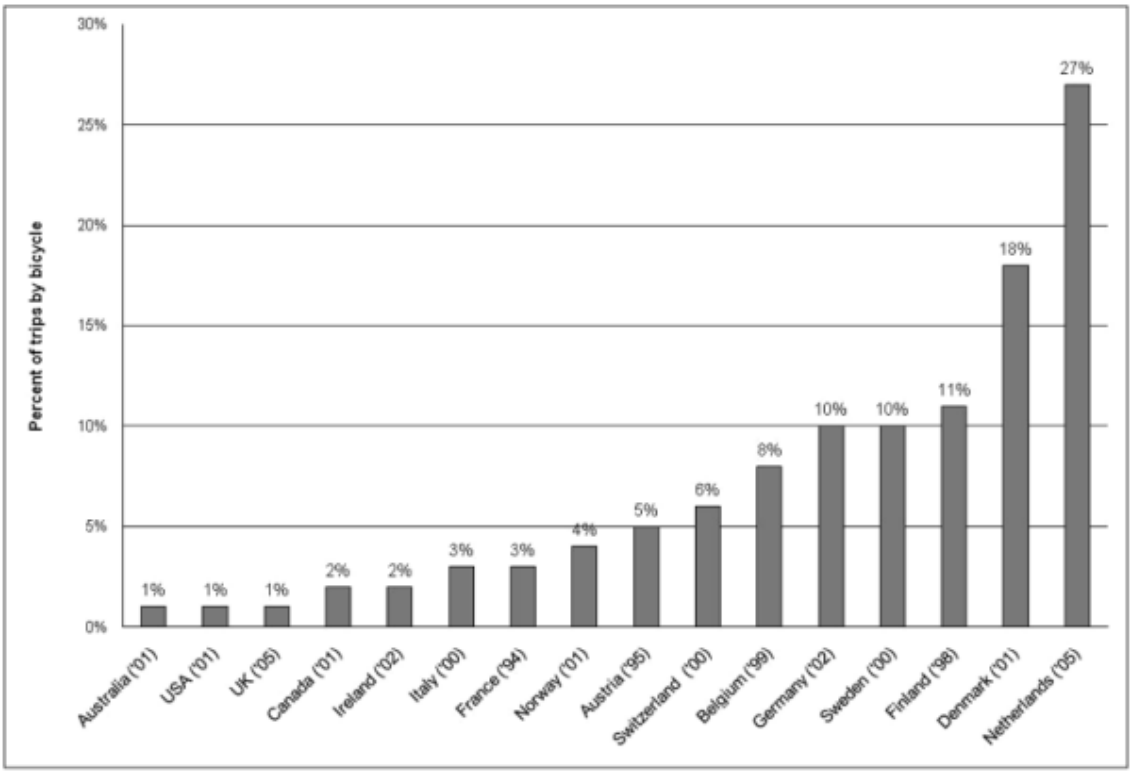
2

In this chapter, existing literature regarding cycling in the Netherlands, traditional and upcoming traffic data collection methods and machine learning that is relevant for this thesis is examined. This literature will then be used to create a conceptual model at the end of this chapter, which will be a starting point for setting up the methodology in chapter 3.

2.1 Cycling policies and traffic data collection

2.1.1 The popularity of cycling in the Netherlands

As already shortly mentioned in chapter 1.1, cycling is very popular in the Netherlands as compared to other countries, with approximately 27 percent of all trips being taken by bike (Pucher and Buehler, 2008). Figure 2.1 shows the large difference between the number of trips taken by bicycle in the Netherlands and various other countries around the world.



Sources: European Union (2003); German Federal Ministry of Transport (2003); U.S. Department of Transportation (2003); European Conference of the Ministers of Transport (2004); Department for Transport (2005); Organisation for Economic Cooperation and Development (2005); Netherlands Ministry of Transport (2006); Australian Bureau of Statistics (2007)

Fig. 2.1.: Bicycle share of trips in Europe, North America and Australia(percentage of total trips by bicycle). Source: (Pucher and Buehler, 2008)

In addition to the percentage of trips taken by bicycle, the distance traveled by bicycle per capita per day also differs significantly. They range from around 0.1 kilometers in Portugal and Spain up to 2.5km in the Netherlands.

To understand why cycling is such a frequently used method of transport in the Netherlands, it is essential to look at the role the Dutch government has played in the past and present regarding cycling policies. According to Pucher and Buehler (2008), one of the most important reasons for the high levels of cycling in the Netherlands and several other European countries is that it is much safer to cycle in those countries when compared to, for example, the United Kingdom or the United States. The government of the Netherlands has spent large amounts of funding and planning on cycling facilities. An average amount of 60 million euro's per year was spent to finance various bicycle projects in the last few years (Pucher and Buehler, 2008).

2.1.2 Traditional and new traffic data collection methods

In order execute and securing funding for bicycle projects and policies, they need to be well-founded and supported by traffic data. In general, there are two main traffic data collection technologies. The first and more traditional technologies are called "in-situ" and refer to the collection of data by placing detectors on or alongside the road. Along with the induction loops as described in chapter 1.1, the most common "in-situ" methods according to Leduc (2008) are:

- *Magnetic induction loops*: As mentioned in chapter 1.1, magnetic induction loops have long been the most common way of measuring traffic. The loops are installed on the road and create a magnetic field that gets disrupted each time a vehicle passes.
- *Pneumatic tubes*: Pneumatic tubes are made of rubber and installed completely across a road. Whenever a vehicle drives over the tubes, the air pressure changes and a pulse of air is created and send towards the detector on the side of the road.
- *Piezoelectric sensors*: This are sensors that contain piezoelectric material(materials that produce a certain electrical discharge when under pressure). The piezoelectric material is deformed when a vehicle or bicycle rides over it, which causes a change in the electric charge that can be detected.

The methods mentioned above are all intrusive, which means that they consist of a sensor on the road and a detector that detects and records the signals sent by the sensor. In contrast to intrusive "in-situ" methods, non-intrusive methods rely on remote observations instead of sensors (Leduc, 2008). There are several common non-intrusive traffic data collection methods:

- *Manual counting*: Traditionally, manual counting was done by hand, and the counts were written down on paper. Nowadays, observers generally perform manual counts with specific applications on electronic devices such as phones and tablets.
- *Infra-red*: Every vehicle and person radiates a certain amount of infra-red energy. This energy can be detected by remote sensors to determine the type and amount of vehicles that have passed.
- *Video detection*: Video cameras can be aimed at a road to record vehicle license plates, type and speed. Systems like this can also be used to detect the total amount of vehicles that pass.

The second, new and upcoming, traffic data collection technique is through the use of GPS localization. Due to the 'smartphone revolution' as defined by Romanillos et al. (2016) and described in chapter 1.2, this has become a viable alternative or supplementary traffic data collection method for the more traditional methods listed above. Harvey and Krizek (2007) were the first to research cycling mobility using GPS in 2006 and 2007. They tracked 51 participants to study and analyze their cycling behavior. In the end, they noted that cleaning up the GPS data proved to be challenging due to positional errors that occurred during recording, and that matching the GPS tracks to existing road network could vastly improve the ability to analyze cycling behavior. In 2010, Reddy et al. (2010) were one of the first researchers to carry out research where smartphones were used to gather the GPS tracks of cyclists. This project aimed at improving the process of sharing routes with other people, as well as being able to view your own route.

2.2 State-of-the-art of gravity models

In section 1.2, it was already shortly noted that most standard traffic predictions approaches use gravity models. The use of a gravity model for researching trip distribution was originally proposed in the 1950s, such as by Casey (1955). Since then, not much as changed when it comes to gravity models. Sen and Smith (1995) state that gravity models can be described as a representation of mean interaction behavior. They report that every different independent spatial interaction process, also known as P, can be characterized by the mean interaction frequencies that it is associated with. Currently, the gravity model is still the most used method for spatial interaction. According to Rodrigue (2017), the basic formula for gravity models as they are used today is as follows:

$$T_ij = k * \frac{P_i * P_j}{d_ij}$$

Where Pi and Pj represents the weight or importance of the origin and destination locations, Dij represents the spatial distance between the origin and destination, and k represents a constant. Calibration parameters can be used to extend the gravity model. These parameters include beta for the friction between the origin and destination location, lambda for the potential of movement, which is often related to welfare, and alpha for the attractiveness.

2.3 Machine learning

2.3.1 Brief history of machine learning

While the main developments in machine learning have been going on since the 1950's, the groundwork for machine learning as it is used today was already being done in the 18th century. In 1763, Thomas Bayes published his work "*Essay towards solving a Problem in the Doctrine of Chance*", which contains the Bayes Theorem, a mathematical theorem that describes how probabilities can be calculated. This theorem remains an important part of machine learning to this day.

One of the first well known modern examples of machine learning and artificial intelligence is the was developed in 1952 by Arthur Samuel. He created a learning machine that was able to play checkers. It learned how to play by playing against itself and remembering which moves where good and which moves where bad. During the same decade, Frank Rosenblatt invented an algorithm called the perceptron, one of the earliest examples of a neural network. The perceptron could take an image and then process and recognize it to produce an output.

However, despite the hype that was generated by the perceptron at first, people soon realized that it was impossible to train it and make it learn new patterns. The lack of discoveries and progress in the field of AI and machine learning, coupled with the disappointment about existing technologies such as the perceptron, lead to a phenomenon called the "AI winter" in the 1970's. During this period, the number of resources and funding spent on AI and machine learning research was greatly lowered.

Since the late 1990's, machine learning and AI have grown increasingly popular and returned to the forefront, strengthened by several highly publicized events. In 1997, a computer called Deep Blue machine defeated famous chess champion Gary Kasparov in a match. Other famous examples of machine learning AI defeating humans are IBM's Watson, who defeated two champions in a show called Jeopardy, and AlphaGo, a system developed by Google that defeated the top Go player of the world in 2016 (The Royal Society, 2017).

2.3.2 Basics of machine learning

Nowadays, two main kinds of machine learning exist: unsupervised learning and supervised learning. The main difference is that unsupervised learning deals with unlabeled data, while supervised learning deals with trying to predict a label based on other variables and features (The Royal Society, 2017).

When using (supervised) machine learning, there are two types of variables: the predictor (or feature) variables and the target (or goal) variable. In this case, the target variable is the counted traffic intensity of cyclists. The predictor(or feature) variables are GPS tracks and network features. When the target variable is a continuous number (the traffic intensity of cyclists at a certain road segment in this case), the machine learning task is called a regression task (Hastie et al., 2001).

A vital part of machine learning is finding out how accurate the created model is. If the model is not accurate with respect to measurements, the results are less useful. The accuracy of a model with respect to measurements can be captured as the fraction of the correct predictions in classification tasks. For regression tasks, the accuracy of the model with respect to measurements can be captured by, among others, the coefficient of regression. When a model fits very well to a certain dataset, but very bad for a new and unseen dataset, overfitting may be taking place. The opposite of overfitting is underfitting, which is when the model does not fit to the training data and neither to new and unseen data (The Royal Society, 2017). To prevent the overfitting of a machine learning model and check the accuracy, the data needs can be split into two sets:

- **Train set:** The training dataset is used to make the model 'learn' and fit the weights of all predictor variables. In other words, the training set is used to build the model.
- **Test set:** This part of the dataset is used after the model has been trained on the Training dataset. By measuring the fitting of the model on the test dataset, an overestimation of the model accuracy, which comes from overfitting (to a particular dataset), can be prevented.

In the end, the accuracy and performance of the model are greatly dependent on how the split between training and test data is done. To combat this problem and make the split less arbitrary, cross-validation can be used. In cross -validation, random splits(folds) are done between training and test data, and for each split the accuracy is calculated. The downside of increasing the number of folds is that it takes more computing power and thus takes a longer time.

2.4 Input features

Using machine learning as described in the previous section, the model needs predictor features to accurately predict the traffic intensity as a goal variable. This section will focus on which input features, according to literature, may be of influence and can thus be used for a machine learning model focusing on traffic intensity.

GPS tracks
The main interesting input feature to predict cyclist traffic intensities are the intensities gained from GPS tracks. Currently, very little research has been done on the possibility of using GPS tracks. Some researches have attempted to predict traffic flows based on GPS data, such as Necula (2014). However, it still is a very new area of research, which means many aspects have not yet been thoroughly researched.

Spatial Distance
When trying to make predictions between different traffic counts for different road segments, spatial features are important. May et al. (2008) say that the spatial distance between the road segments may be useful to take the spatial dependence of traffic flows into account. For example, when trying to find out the traffic intensity for a certain road segment, it seems logical that the intensities measured by induction loops close to the road segment have a higher influence than the ones measured by traffic loops that are relatively far away. Thus, a distance matrix, which calculates the distance on the network between induction loops and/or road segments, could be used as an input feature.

Road surface type
In addition to spatial features such as distance, network features also can be used for interpolation. One such relevant network feature is the type of the road surface. Noland and Kunreuther (1995) say that the condition of the road surface may have a significant influence on whether or not cyclists decide to ride on it. When a road surface is bad and unsuitable for cycling, cyclists are forced to choose other routes due to their perceived lack of security. Similar results were also found by Stinson and Bhat (2004). They concluded that paved and smooth road surfaces are preferred by cyclists over unpaved roads.

Width of the road
Another network feature that may be of interest is the width of the roads on which people cycle, because of speed and security considerations. Petritsch et al. (2006) found that roads with two lanes are preferred by cyclists over wider roads with more than two lanes. They argue that this has to do with the fact that, on wide roads, drivers are focused more on other drivers than on cyclists, which leads to cyclists being more prone to have accidents on those wider roads.

Attractivity
Finally, the attractiveness factor of the roads is found to be usable as a predictor variable. According to Cerná et al. (2014), people are more likely to cycle along routes that they deem 'attractive'. They argue that the attractiveness of a certain road can depend on the number of points of interest along this road, such as service facilities, parks and nature, utilities or scenic landscapes.

2.5 Conceptual model

All theoretical concepts and principles described in this chapter can be combined into a conceptual model, that shows how they relate to each other and how they need to be combined to reach the desired results and answer the research questions. Figure 2.2 shows this conceptual model.

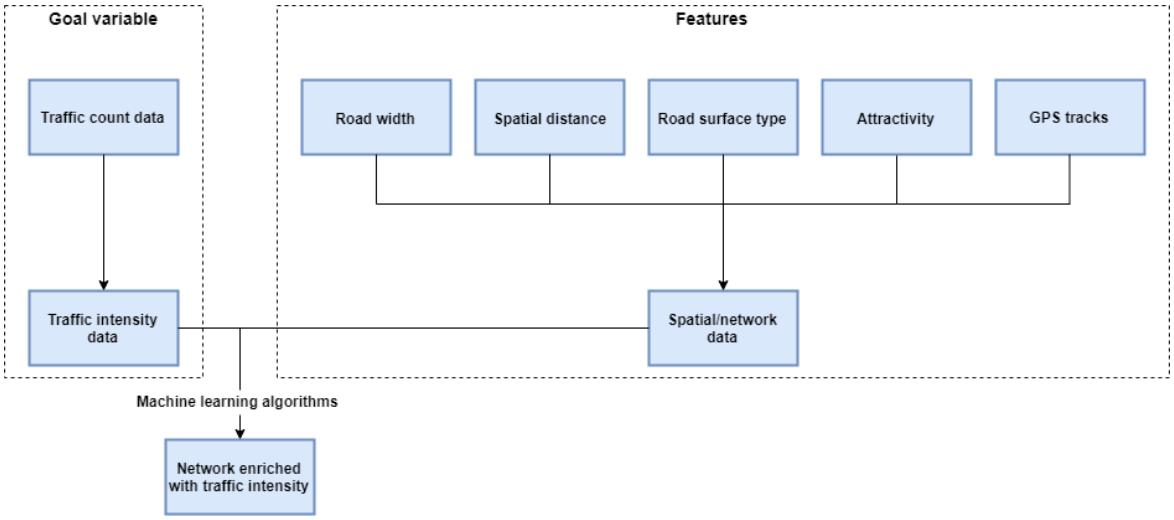


Fig. 2.2.: Conceptual model of the research based on existing literature.

The following chapter will elaborate and build upon this conceptual model and explain the methodologies that are used.

Research Methodology

3

In this chapter, the methodology that is used to carry out the research is detailed. First, all used data sources will be described, along with the used software packages. Next, an overview will be given about the quality of the data sources, followed by the specific machine learning methods that are used. The chapter ends with a list of steps that will be taken to produce results and a schematic overview of the entire research process.

3.1 Used data and software

3.1.1 GPS data

B-Riders
As already mentioned in chapter 1.2, the B-Riders project is based on the idea to get more people to use their bicycles instead of their cars. The original project was started back in September 2013 and ended in December 2014, and contains GPS tracks of over 700 people for this period of time. To motivate people to participate and as an incentive to track their trips, a financial compensation of around 10 to 15 cents could be claimed for each registered kilometer that was cycled while using the GPS application of B-Riders (B-Riders, 2017). Due to the voluntary nature and incentive of participating in the B-Riders project, it is extremely likely that the data is not representative for the whole of Noord-Brabant(or The Netherlands). This does not mean the data is unusable, but it is important to keep in mind to prevent drawing the wrong conclusions from the outcome of the analysis. It is important to note that while the original B-Riders dataset contains privacy details about the involved participants, the used data is anonymized before being used. This data anonymization ensures that the privacy of the participants is respected, while still keeping all the necessary information that is needed to help answer the research questions.

Fietstelweek
Another source of GPS data is the 'Nationale Fietstelweek' (in English: National Cycling Count week). This event is organized for a week each year by the Dutch Ministry of Environment(In 2017 the name of this ministry was changed to Ministry of Infrastructure, Public Works and Water Management), together with a majority of the provinces in the Netherlands including Noord-Brabant. During this week, participants can track their cycling activity via an app, in a very similar fashion to the B-Riders project (Fietstelweek, 2017). The first edition of the Fietstelweek took place in 2015, and since this is closest to the collection date of the B-Riders project, the 2015 data is used.

GPS track aggregation
Since the GPS data described above is measured as single points along a certain path for each track, the tracks that are described via these points need to be aggregated to be able to extract traffic intensity and flow information from them. One way in which this can be done is by 'map matching', in which the GPS points are matched to the road segments. Newson and Krumm (2009) describe a map matching method based on a Hidden Markov Model approach, which is a model that can be used for pattern recognition and is therefore very useful for matching GPS points to a network. They use a GPS track represented by points, just like the B-Riders dataset. From each node(GPS point in this case), the most likely route is calculated through the Hidden Markelov method. Ultimately, this method can lead to a path between

the GPS points that are matched to the underlying network. By doing this, for each road segment, the amount of generated paths that are matched to it can then simply be counted to get the traffic intensity of that specific road segment.

3.1.2 Traffic data

As mentioned earlier in chapter 1.4, the area of focus is Tilburg, partly due to the availability of suitable data. For counted traffic data, one of these data sources is the 'Fietstelprogramma Gemeente Tilburg'. For this program, the traffic count of cyclists (as well as mopeds) per hour was counted for various road segments across Tilburg, for both travel directions. Furthermore, next to traffic count data from a municipal level, provincial traffic count data from and around the region of Tilburg was also considered. However, it was decided to not use the provincial traffic count data since the amount of data collections points was very low, and the data was collected at a completely different time period, which made combining it with municipal traffic counts not preferable.

3.1.3 Network data

OpenStreetMap
When it comes to determining a suitable network, there are several options. Perhaps one of the most frequently used networks is that of OpenStreetMap(OSM). This free and open-source platform aims to create an extensive map of the entire world. Users can freely edit the map to change or add geographic information, which makes it a prime example of a volunteered geographic information (OpenStreetMap contributors, 2017). A potential downside of using OSM is that since it makes use of crowdsourcing, there is no official quality control, and the accuracy of the data can not always be guaranteed. However, the quality of the OSM network in the Netherlands is generally considered sufficient to perform the analysis.

Fietsersbond
In addition to OSM, the bicycle network maintained by the Dutch Fietsersbond(Cyclists Union) needs to be considered. This union defends the interests of all cyclists in the Netherlands. For over ten years, the Fietsersbond has had its own cycling route planner. The underlying network of this route planner can serve as a basis for machine learning. An advantage of this network is that it contains Points of Interests(POI's) for cyclists such as recreational sites, public transport and potential barriers such as traffic lights (Fietsersbond, 2017b).

3.1.4 Used software

The machine learning algorithms that are used to estimate traffic intensity (as described further along in chapter 3.3) are executed using the Python programming language. Several additional libraries are needed to perform the analysis. Firstly, to be able to properly manipulate and pre-process the datasets, the pandas library is used. The machine learning aspect is handled by the Python plugin scikit-learn. This is an open-source library containing the most common machine learning algorithms (Pedregosa et al., 2011). Additionally, the NumPy and SciPy Python libraries are used to enable numerical and scientific computing respectively. Both packages are fully integrated and compatible with scikit-learn. To visualize and plot results, both Matplotlib, a Python library designed for plotting data, and ArcGIS will be used.

3.2 Data quality

When working with many different kinds of data from various sources, it is important to look at the quality of the used data. There are many different ways of measuring data quality, Pipino et al. (2002) name 16 different dimensions along which the quality of data can be assessed. In table 3.1, all data sources that will be used, as described earlier in this chapter, are evaluated for the dimensions that are deemed relevant and might cause potential problems. The data quality assessment in the table is done using color coding. If a cell is green, it means that the quality of a data source is deemed sufficient for that specific dimensions, while an orange cell means that the quality of the data for that dimension might potentially not be sufficient.

For layout and readability purposes, the data sources have been labeled as follows:
1: B-Riders GPS tracks
2: Fietstelweek GPS tracks
3: Tilburg traffic counts
4: OpenStreetMap network
5: Fietsersbond network

Dimension	**1**	**2**	**3**	**4**	**5**
Accessibility					
Appropriate amount of Data					
Believability					
Completeness					
Free-of-Error					
Objectivity					
Representativeness					
Timeliness					

Tab. 3.1.: Assessment of data quality based on dimensions as described by Pipino et al. (2002).

Below, a more detailed explanation is given for each dimension in which the data quality for one or more sources might not be sufficient:

Accessibility
Both the Fietstelweek GPS tracks and the OpenStreetMap network are open data and freely accessible. The B-Riders tracks, Tilburg traffic data and the Fietserbond network are not open data however, and will be provided by the NHTV Breda.

Appropriate amount of data
The OpenStreetMap and the Fietsersbond network can both be considered of containing appropriate amount of data, which in their case are the cyclist roads of Tilburg. The main data source that could provide problems in terms of appropriate amounts of data are the Tilburg traffic counts. Since there are only 24 locations for which cyclist intensities were collected, this sample size could turn out to be to small to estimate cyclist intensities.

Believability
The GPS tracks and the traffic data are collected from well-functioning devices (smartphones and traffic counts respectively) and can therefore be regarded as credible. Both the OpenStreetMap and Fietserbond networks are maintained by volunteers, however, with no official quality control taking place. For the Netherlands, the data can be regarded as credible for most locations, but it is import to remain cautious for any potential errors.

Completeness
Regarding completeness, the same can be said about the OSM and Fietsersbond network as with its believability. The voluntary nature of the dataset can cause problems when it comes to completeness, such as missing information for certain roads. This can happen because it is not required to fill in all fields when editing the network. However, as said above, for The Netherlands the network can generally be regarded as complete and suitable for analysis.

Free-of-Error
All data sources always have a change of containing errors, but it can be said that the traffic counts of Tilburg have the lowest change of containing error, with the only possible errors occurring because of human error. The GPS tracks can contain numerous errors caused by an unstable GPS connection, which might lead to tracks having the wrong location or not being complete. Both networks may once again contain errors caused by the voluntary nature of the projects.

Objectivity
The used GPS data is biased, and not a completely accurate representation of the average cyclist. This has several reasons, such as requiring a smartphone and, in the case of the B-Riders tracks, being paid to participate. The network datasets may again not be complete objective due to them being voluntary.

Representativeness
Due to the nature of the B-Riders and Fietstelweek projects, it can already be said beforehand that these data sources are most likely not representative. The participants of those projects are not an average representation of the cyclists in Tilburg, which is also cause by the fact that the sample size is relatively small.

Timeliness
The GPS tracks and traffic data are recorded at a certain date in the past. Therefore, it is important to make sure that all data used is from roughly the same time period if possible, otherwise combining those data sources might lead to wrong results and conclusions. Since the road layout does not change frequently, the network data can be regarded as up-to-date and sufficient for the analysis.

3.3 Machine learning regression

Regarding machine learning, there are many different methods that can be used to predict and estimate values. Since it was established that, based on the literature, the research question can be regarded as a supervised machine learning problem, this section will shortly describe the supervised machine learning methods that will be used to predict the traffic intensity for road segments. It is important to note that, in addition to the seven methods listed below, there are many more methods that could be used to predict traffic intensity. Due to time constraints and to keep the scope manageable, only the methods listed below will be used in this research as they are deemed the most suitable.

3.3.1 k-Nearest Neighbor Regression

The k-Nearest Neighbor method was found to be a good candidate method for flow prediction by May et al. (2008). In k-Nearest Neighbor, the value of the point that is to be calculated is determined by looking at the mean values of its nearest neighbors in a multidimensional space by machine learning features. In deciding the neighboring points that should be taken into account, one can either determine a total number of neighbors or create a buffer with a certain radius around the target point and take all neighbors in that radius into account during the learning. The weights that each neighboring point gets can be uniform, based on distance so that nearby points weigh more heavily than point further away for example, or based on any other definition of distance (such as time) (Scikit-learn, 2017c).

3.3.2 Decision Trees Regression

Another supervised machine learning method that can be used for regression tasks is Decision Trees. A decision tree model predicts the target variable based on so-called decision rules. These decision rules are based on the input features, such as the ones described in chapter 2.4. Decision trees come with numerous advantages, such as being able to be visualized for easier interpreting, as well as the possibility to perform statistical tests to validate the model (Scikit-learn, 2017a).

3.3.3 Gaussian Processes Regression

The third machine learning method that will be used is called Gaussian Processes. These processes make use of interpolation and look at, among others, the kernel function(similarity) between points to estimate the target variable based on the predictor variables. Advantages of this kind of machine learning methods are that they are considered highly flexible, and can be optimized precisely for the tasks that are required (Scikit-learn, 2017b).

3.3.4 Kernel Ridge Regression

Kernel Ridge Regression is a form of regression where, as is implied by the name, normal Ridge Regression is kernelized. Kernel Ridge is very similar to support Vector Regression(see below), with the main difference being the use of other loss functions (Scikit-learn, 2018b).

3.3.5 Support Vector Regression

While Support Vector machines are often used for classification analysis, they can also be used for regression problems. Drucker et al. (1997) were one of the first to describe Support Vector Regression, based on earlier concepts of Vladimir Vapnik. The idea behind Support Vector Regression is that it always tries to minimize the errors and maximizes the margins. Since it works based on nonparametric techniques, it can be described as a kernel function.

3.3.6 Partial Least Squares Regression

Tobias (1999) mentions Partial Least Squares Regression as a suitable method for making predictions when the factors are relatively collinear. It differs from a lot of other regression methods in that it projects both the observed as well as the predicted variable to a new space. Because of this, it is an example of a cross decomposition method. PLS is especially useful if there are fewer observations than variables in the feature variables matrix.

3.3.7 Linear Regression

Linear regression is one of the most standard forms of regression. Freedman (2009) state that normal linear regression deals with only one feature variable. When using more than one feature variable, multiple linear regression is used. During linear regression, the predictive model is fitted to the observed data.

3.4 Cross validation

To assess results of the machine learning classifiers mentioned above, cross-validation is used. This technique is used to validate the accuracy of the predictions of the models and see how the analysis will perform on other datasets. In general, two main types that can be distinguished in cross-validation: Exhaustive cross-validation and non-exhaustive cross-validation. Exhaustive cross-validation methods, as implied by the name, look and test every single possible combination of training and test data set for the given sample. On the other hand, non-exhaustive cross-validation does not look and test every single possible combination but uses approximations. An example of exhaustive cross-validation is leave-one-out cross-validation, while an example for non-exhaustive cross-validation is k-fold cross-validation. Both of these methods and their use for this case are discussed below.

3.4.1 K-fold

When performing k-fold cross-validations, a k amount of subsamples are created from the original sample, with all of these subsamples having an equal size. One of the subsamples is randomly chosen as the validation data, while the other subsamples are used as training data. This can be repeated k times(each of those times being known as a fold), hence the name k-fold cross-validation.

3.4.2 Leave-one-out

Leave-one-out cross-validation(LOOCV) is a variant of leave-p-out cross-validation(LpOCV. During LpOCV, a specified number of p samples from the original sample are used for validation, while the other samples are used as training data. This process is repeated for every single possible combination of validation and training data. LOOCV consists of a $p = 1$, which means that only 1 sample is used for validation while all others are used for training. An interesting and important thing to note is that when k equals the total number of samples in K-fold cross-validation, it is essentially the same as LOOCV.

3.5 Model scores

By performing K-fold and leave-one-out cross-validation, several functions are used to measure the performance of the regression analysis. The two metrics that are used to measure the model score are R^2 and the Mean Squared Error. Both of these functions are available within the sk-learn package.

3.5.1 R^2-score

R^2, also known as R-squared measures the coefficient of determination. It measures the how much of the variability of the fitted model is explained. If the R^2-score is zero, that means that the model is constant and always predicts the expected value of y without taking into account the input feature variables. This is an example of a (very) poorly fitted model. On the other hand, if the model fits perfectly to the data, the R^2-score will be 1.0. Thus, the closer the R^2-score is to 1, the better the fit of the model (Walpole et al., 2013). It is also important to note that the R^2-score used in scikit-learn can get negative, which signals that the model is performing worse than the naive model. Walpole et al. (2013) also mention several dangers of (only) using R^2-score to measure the performance of the model. Firstly, they mention it is difficult to decide what value of R^2 is acceptable, as there is no decisive standard. Secondly, when comparing models by their R^2-score, overfitting can cause R^2 to be (artificially) high, while this may not directly imply that that model is better in prediction the target variables. Therefore, in addition to the R^2-score, the Mean Squared Error(MSE) is also measured for each model during cross-validation.

3.5.2 Mean Squared Error

Another metric that, according to Walpole et al. (2013) is often used to compare different estimators is the Mean Squared Error(MSE). As the name implies, the MSE calculates the mean of the squared errors of the estimator. Since the error is squared, the MSE of an estimator is always positive. The larger the error of an estimator, the larger the MSE of that estimator. Thus, the closer to zero the MSE gets, the better the model.

3.6 Feature selection

After acquiring the data sources mentioned above in chapter 3.1, the data needs to be explored and the possible input features need to be selected based on their availability and suitability for analysis. For each of the possible input features as found by literature in chapter 2.4, the available data sources are examined to see if the features are available and suitable. Only features whose variance is above a certain threshold, in this case 0, are selected. This is one of the simplest approaches to feature selection (Scikit-learn, 2018a).

The variance of each possible feature is determined by first calculating the standard deviation of each possible input feature that is available in one of the used data sources for the municipality of Tilburg. This is done by using the statistics function in ArcGIS on each relevant column. Then, by taking the square of the standard deviation, the variance of the possible input features is calculated. As mentioned above, only features that have a variance above 0 are deemed suitable for analysis, which means that all zero-variance features are not selected.

GPS cycling intensities
For the GPS cycling intensities, the 'intensitei' column of both the B-Riders and the Fietstelweek are considered as possible input features. For the B-Riders network, the standard deviation for the GPS cycling intensities is 385,12, which means the variance is 148317,19. This means the feature is suitable for analysis and therefore selected. Regarding the Fietstelweek network, the standard deviation for the GPS cycling intensities is 20,98, which means the variance is 440,28. This means that, in addition to the GPS cycling intensities, the Fietstelweek GPS cycling intensities are also selected.

Attractivity
The 'schoonheid' column from the Fietsersbond network is considered as a possible input feature for attractivity. The OSM network does contain a separate point file with Points of Interests, but this deemed unsuitable due to the difficulties of converting Points of Interests to attractivity values for each road segment. The standard deviation for the attractivity feature for all roads in the municipality of Tilburg is approximately 1,46, which means the variance is approximately 2,12. This means the feature is suitable for analysis and therefore selected. The exact breakdown of the amount of features per class can be found in table 4.2 in chapter 4.3.2.

Road surface type
The 'wegdeksrt' column from the Fietsersbond network is considered as a possible input feature for road surface type. The OSM network does not contain any information about road surface type. The standard deviation for the road surface type feature for all roads in the municipality of Tilburg is 1,20, which means the variance is 1,44. This means the feature is suitable for analysis and therefore selected. The exact breakdown of the amount of features per class can be found in table 4.3 in chapter 4.3.3.

Width of the road
The 'breedtekls' column from the Fietsersbond network is considered as a possible input feature for the width of the road. The OSM network does not contain any information about the width of the road. The standard deviation for the road surface type feature for all roads in the municipality of Tilburg is 0, which means the variance is also 0. Therefore, the width of the road feature is not selected as in input feature.

Spatial distance
When it comes to spatial distance, it becomes clear by projecting both the OSM and Fietsersbond network against a background map that the road segments have a spatial component. This also becomes clear due to the geometry field that each seperate road segment has. Chapter 4.3.4 further explains how the spatial feature variable is prepared and pre-processed.

Traffic counts
For the goal variable traffic counts, the measured intensities from the 'Fietstelprogramma Gemeente Tilburg' are checked to be sure that they can be used for analysis. The standard deviation for the 'r1_intens' column is 1089,78, which means that the variance is 1187626,69. Regarding the 'r2_intens', the standard deviation is 1044,56, which means the variance 1091113,80. The traffic counts from the 'Fietstelprogramma Gemeente Tilburg' are therefore suitable and selected as goal variables.

Chapter 4 describes the file structure of each of the selected features mentioned above in more detail and will describe how they are prepared and pre-processed to be suitable for analysis.

3.7 Research steps

Below, a more detailed overview is given of all the steps that need to be taken to conduct the research, based on the conceptual model and methodology. The research steps are divided into two sections. First, the necessary steps that are needed to prepare the data are listed, followed by the steps that have to be done to perform the analysis.

Data preparation

- Firstly, all the data sources mentioned earlier in chapter 3.1 need to be obtained. When possible this is done via open data sources. The data that is not available via open data sources, such as the Tilburg traffic data, will be provided by the NHTV Breda.
- Afterward, a decision was made on whether to use the OSM network, the Fietsersbond network or a combination of both. This will be done based on the availability of potential feature variables that are identified during the literature study and the feature selection.
- The B-Riders and Fietstelweek GPS tracks then need to be aggregated, and the resulting traffic intensity value needs to be matched to the correct road segment on the selected network. This is done via map-matching as described further along in chapter 4.2.
- The input features from the road network need be checked and any potential errors, such as missing values or different types of notation need to be corrected or deleted. They then need to be numerical values to be suitable for analysis.
- Next, the traffic count data from the region of Tilburg needs to be added to the network. This is done by projecting each traffic count point based on its spatial location, and then manually matching them to the corresponding road segment.
- Finally, the mapmatched GPS tracks containing the GPS intensities need to be joined to the original network containing all the other variables. If possible, this needs to be done by using common identifiers, otherwise

Analysis

- To examine the ratio between the traffic intensity aggregated from GPS and the traffic intensity measured by traffic counts, the ratio will be calculated for each road segment where both are available.
- Next, before performing the analysis using machine learning regressors, the correlation between the goal variable and all feature variables needs to be shown using scatterplots.
- Afterward, the machine learning algorithms are run for several different combinations of goal and feature variables.
- The performances of each of the machine learning algorithms for each combination of variables are described and compared by calculating the R2-score and MSE using 5-fold cross-validation and LOOCV.

3.8 Schematic overview of research process

To summarize and gain a clear overview of all the steps that need to be taken to realize the objectives and research questions as described in this thesis, figure 3.1 shows a schematic overview of the entire research process, based on the research steps mentioned in the previous section.

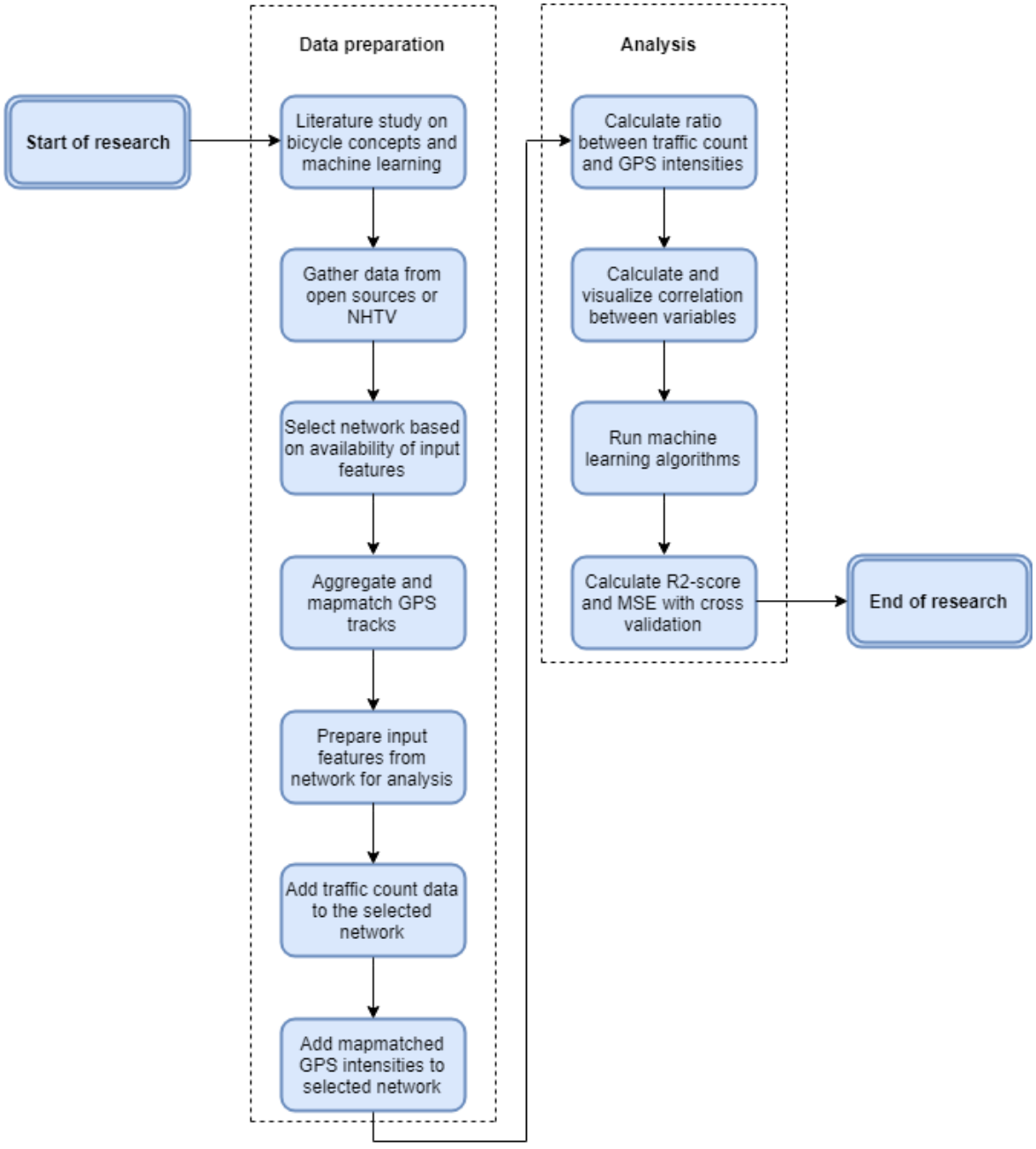


Fig. 3.1.: Schematic overview of all steps of the research process.

Data preparation and pre-processing

4

This chapter details the pre-processing steps that are taken to make the data suitable for analysis. All scripts used during the analysis can be found on Github by clicking HERE, or by using the following link: `https://github.com/JJochemsen/GIMA-M7-Thesis`. Whenever a script was used, its name in the repository will be mentioned in the concerned text.

4.1 Network selection

The first step of data preparation and pre-processing is that a selection of the network to be used needs to be made. Both the Fietsersbond and the OSM network were obtained and compared to find the one most suitable for the analysis. Combining both networks, to take advantage of each others variables is also an option that was considered. However, it quickly became clear that this is not feasible for multiple reasons. First, since both networks are created and maintained via the principle of VRI, they differ slightly on a spatial level. This means joining the networks based on spatial location is nearly impossible. Secondly, the networks do not contain a common identifier which could be used to combine them. Thus, after reviewing both networks, the Fietersbond network is chosen since it contains most of the feature variables listed in Chapter 2.4, and as explained in more detail in chapter 3.6.

As mentioned in Chapter 1.4, this research will focus on the Municipality of Tilburg. Since the Fietsersbond network covers the entire Netherlands, a selection is made to only select the road segments that lie within the municipal borders of Tilburg. Since the data contains a field with the municipality of each road segment, this selection is easily made by selecting all features where the "Gemeente" (municipality) attribute equals Tilburg. By doing this, the original Fietsersbond network containing 1.554.923 road segments is reduced to only 14.308 road segments for the municipality of Tilburg. An overview of the resulting network, along with its location within the Netherlands can be seen in figure 4.1 on the following page.

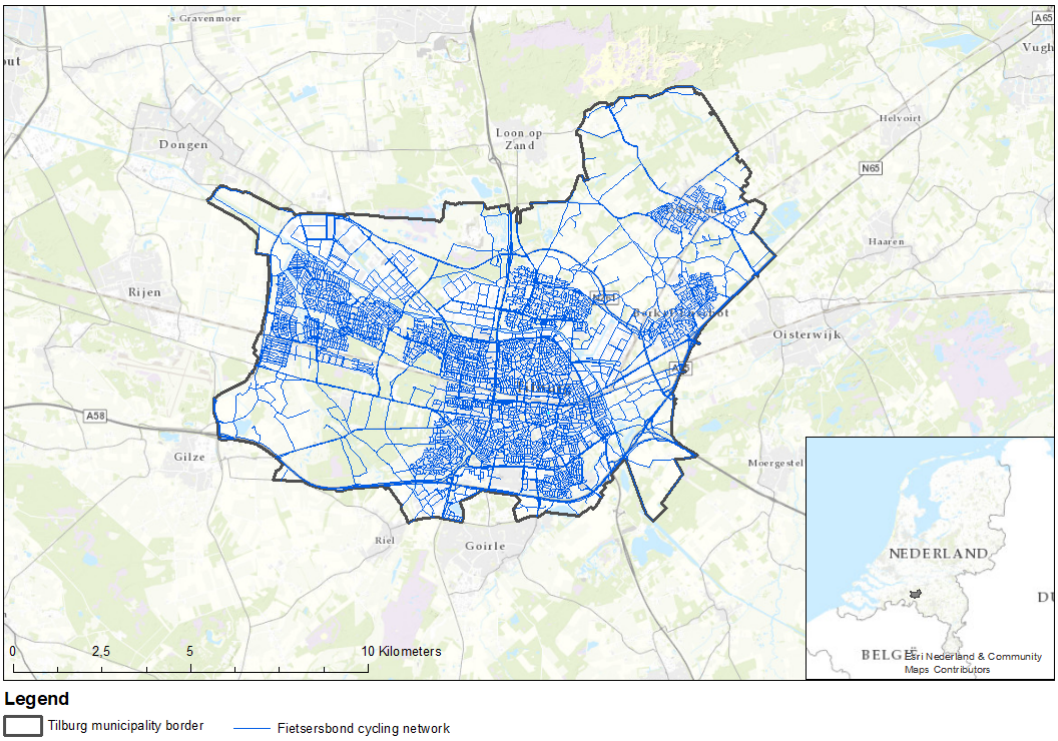


Fig. 4.1.: Overview map of Fietsersbond cycling network of Tilburg

4.2 Mapmatching the GPS tracks

The mapmatching of both the B-Riders data and the Fietstelweek tracks onto the networks was done by the NHTV. The script uses the raw GPS data collected by the B-Riders and Fietstelweek and the unmodified cycling networks of the Fietsersbond and OSM as input.

For the start and end of each route, mapmatching looks for three nodes("knopen") in the network which are located close to the start and end point of the route and are not connected with each other. Every link(road segment) in the network gets a resistance factor which increases the further away the link is from the perceived GPS points.

Next, for all possible combinations of selected nodes(3x3, so nine nodes in this case) the route with the lowest resistance is decided. Per definition, this calculated route lies close to the original GPS line, but on a logical route on which cycling is possible and allowed. For example, whenever a GPS-line(straight lines between all GPS points of a route) enters or leaves a parallel road the mapmatching script makes sure to not assign the route to (inaccessible) the main road, even though the GPS points, through inaccuracy, may lie closer to the main road than to the parallel road.

For every link, the start- and endpoint are then projected onto the GPS-line and the exact time, with the accuracy down to seconds, is interpolated since in most cases the link is located between two GPS points. Based on those two calculated times as well as the length of the link, the absolute speed is calculated, including the possible waiting time spend on nodes caused by, for example, traffic lights.

Sometimes links are missing in the network. When this happens, either no route is calculated, or a route with a noticeable detour is calculated. In the first case, there is nothing that the mapmatching can do to fix it, so the route is 'rejected' and not used. This rarely happens, but when it does, it is mostly in areas outside of the target area, such as countries abroad. The second case, routes with detours, are a more common occurrence. This is detected by the mapmatching script if there is an illogically large difference between the calculated route and the GPS-line. The links that contain the detour are removed from the mapmatching process in this case, since it is clear that the person in question did not cycle on those links.

Additionally, the newer versions of the mapmatching script also look at the type of links. Several types of links are excluded by default, such as railroads for cyclists. Others are included, but with varying resistance. For example, bicycle paths have a lower resistance than pedestrian paths. Concretely, this means that when a GPS-line follows a certain type of link for its entire length, it is mapmatched on those links whether they are bicycle paths or normal roads. But in situations where bicycle, car, and pedestrian roads are located right next to each other, within the standard GPS-inaccuracy distance, the script will prefer the bicycle roads since those roads have the lowest resistance. (Bussche and van de Coevering, 2015)

4.3 Preparation of variables

4.3.1 Feature variable: GPS cycling intensities

As mentioned in the previous section, the mapmatching process gives two output files. The first is a table which contains all individual routes that were used to calculate the intensity on the road segments, which will later be used to calculate the intensity per weekday. The second output file is a shapefile of the Fietsersbond network, containing the intensity of cyclists for each road segment based on the B-Riders or Fietstelweek GPS data. Figure 4.2 and 4.3 show the file structure of the mapmatched B-Riders GPS data onto the Fietsersbond network.

FID	Shape	LINKNUMMER	SOURCE	TARGET	CYCLE	ONEWAY	SNELHEID	INTENSITEI	ROUTEID	INTENSI_01	SNELHEID_R	VERHOUDING
17415	Polyline	1037224	1127371	1003079	100		18,1612	4711	0	3857	0,756668	1,20267

Fig. 4.2.: File structure of the shapefile with intensities of the mapmatched Fietsersbond + B-Riders file

id	routeid	linknummer	distance_from_start	distance_from_end	richting	sequence	snelheid	uur	weekdag	month	year
1	293162	934511	3025	322	t	38	19,9053	7	4	0	114

Fig. 4.3.: File structure of the route table of the mapmatched Fietsersbond + B-Riders file

The file structure of the mapmatch of the Fietstelweek GPS data and the Fietsersbond network looks roughly the same, with the most noticeable difference that it does not contain a source and target field.

The mapmatched dataset contains the intensity of cyclists on road segments for the total period the B-riders and Fietstelweek project was active. However, since the traditional traffic counts were done on a single day, the intensity of cyclists per day could also provide new insights during the analysis. The route files for each of the mapmatches contains information about the day on which each specific cycling trip is taken, as seen above. This 'weekdag' column contains a value between 0 and 6, with each number corresponding to a specific day of the week. Table 4.1 below shows which number corresponds to which weekday.

Code	Weekday
0	Sunday
1	Monday
2	Tuesday
3	Wednesday
4	Thursday
5	Friday
6	Saturday

Tab. 4.1.: Codes and corresponding weekday of mapmatched tracks

Each specific road segment in the spatial network file has a unique 'linknummer' attribute. Additionally, the routes that travel over a road segment have the corresponding 'linknummer' in the routes file. This means that by counting the amount of 'linknummers' for each day in the route file, the intensity of cyclists for that specific day of the week on each specific road segment is acquired. The query *intensity_per_day.sql* on Github is used to calculate the daily intensities. The query counts the intensity on Sundays for the B-riders dataset, as it only selects rows where the 'weekdag' value is 0. By running this query seven different, once for each weekday, for both the B-riders and the Fietstelweek data, tables are created containing the intensity of cyclists per day per dataset. The resulting tables containing the intensity for each specific day were then joined back to the main mapmatched network, which contains the total intensity, using the query *join_intensity.sql* which can be found on GitHub.

4.3.2 Feature variable: Attractivity

One of the used feature variables is attractivity. The Fietersbond network contains a column called "Schoonheid" that is used as a measure for attractivity. It contains six unique text values that describe the level of attractivity for each road segment. To use these values in machine learning, they will be weighted, ranging from a weight of 1 being assigned to the least attractive road segments and a value of 5 to the most attractive road segments. The original values and their assigned weights are found in table 4.2.

Original text value	Translation	Weight	Amount
Zeer lelijk	Very ugly	1	20
Lelijk/saai	Ugly/boring	2	138
Neutraal	Neutral	3	8712
Mooi	Pretty	4	913
Schilderachtig	Picturesque	5	7
Onbekend/geen waarde	Unknown/missing value	-	4518

Tab. 4.2.: Attractivity values and weights

The *replace_values* query first creates two new columns in the Fietserbond network copying both the Attractivity column and the road surface type column. Then, it replaces the original text values with the weights that have been specified in table 4.2 and 4.3.

4.3.3 Feature variable: Road surface type

Another variable from the Fietsersbond network that is used for analysis is the "Wegdeksrt", or road surface type. Six different values of road surfaces are categorized. Just like the attractivity features, the road surface type values are weighted based on how suitable the road surface type is for cycling, with 1 being the least suitable and 3 being the most suitable. Table 4.3 contains the six different values, along with their weights.

Original text value	Translation	Weight	Amount
Onverhard	Dirt/gravel	1	75
Halfverhard	Semi-dirt	2	29
Klinkers	Clinker bricks	2	5490
Overig(hout/kinderkopjes)	Other(wood/setts)	2	32
Tegels	Tiles	3	1140
Asfalt/beton	Asphalt/concrete	3	3018
Onbekend/geen waarde	Unknown/missing value	-	4524

Tab. 4.3.: Road surface type values and weights

Both table 4.2 and 4.3 show that there is a significant amount of road segments for which the attractivity and/or the road surface type is unknown or not submitted. When visualizing the spatial distribution of the road segments with unknown variables, an interesting pattern is shown as seen in figure 4.4

One part of the road segments with missing values consist of highway and motorways, which makes sense since cycling is not possible and allowed on those roads. The second part of road segments with missing values is almost entirely located in the town of Berkel-Enschot, as seen in the mid-right of the map. An explanation for this might be that, since the Fietsersbond network is maintained by volunteers and no official guidelines exist, the person(s) that mapped this area simply did not submit values for the attractivity and road surface type, among others.

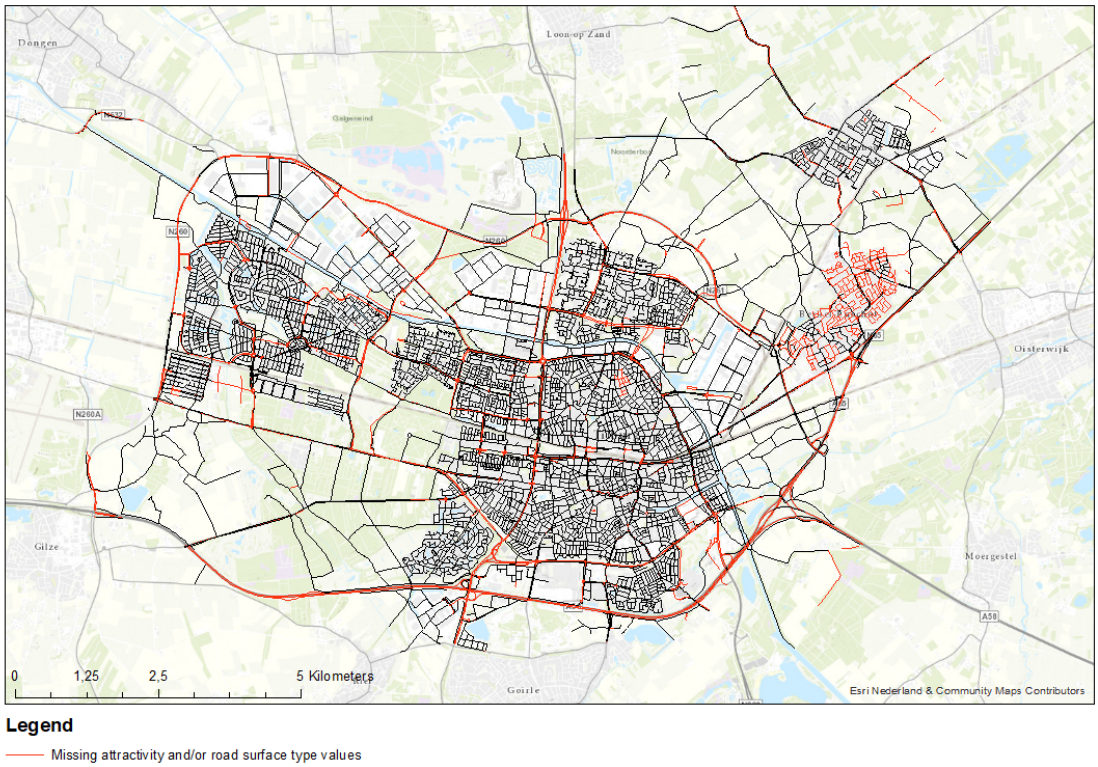


Fig. 4.4.: Location of the road segments with missing values for attractivity and/or road surface type

4.3.4 Feature variable: Spatial distance/location

To use spatial distance as a feature, the spatial location of each row(road segment) needs to be calculated. Since these consist of lines, they contain the LineString geometry for each feature. The X and Y coordinates of the centroid of each LineString are calculated and added as 2 separate columns(X and Y) to the dataframe containing the features. However, this means that the feature is not exactly about spatial distance anymore, since the distances between the individual coordinates are not taken into account. General spatial location was chosen over network distance, since after pre-processing the final analysis file leaves only 27 road segments, none of which are connected to each other anymore. This makes it impossible to calculate the distance over the network. Therefore, the feature variable changes slightly from spatial distance to spatial location, which is an important distinction to make.

4.3.5 Goal variable: Traffic count data

Next to the mapmatched traffic intensities based on GPS tracks, the 'official' traffic counts from the municipality of Tilburg are needed. For this, the bicycle traffic counts from the so-called "Fietstelprogramma Gemeente Tilburg" are used, as mentioned earlier in chapter 3. These traffic counts are conducted every year by company Groenlicht Verkeersadviezen for the municipality of Tilburg. This datasource consists of 24 different traffic count points for cyclists, for which the intensity of bicycles is counted between 7am and 7pm on a single day. Figure 4.5 shows the location of these traffic count points within the municipality of Tilburg. Appendix A shows a more precise map with the two directions which the intensities are measured.

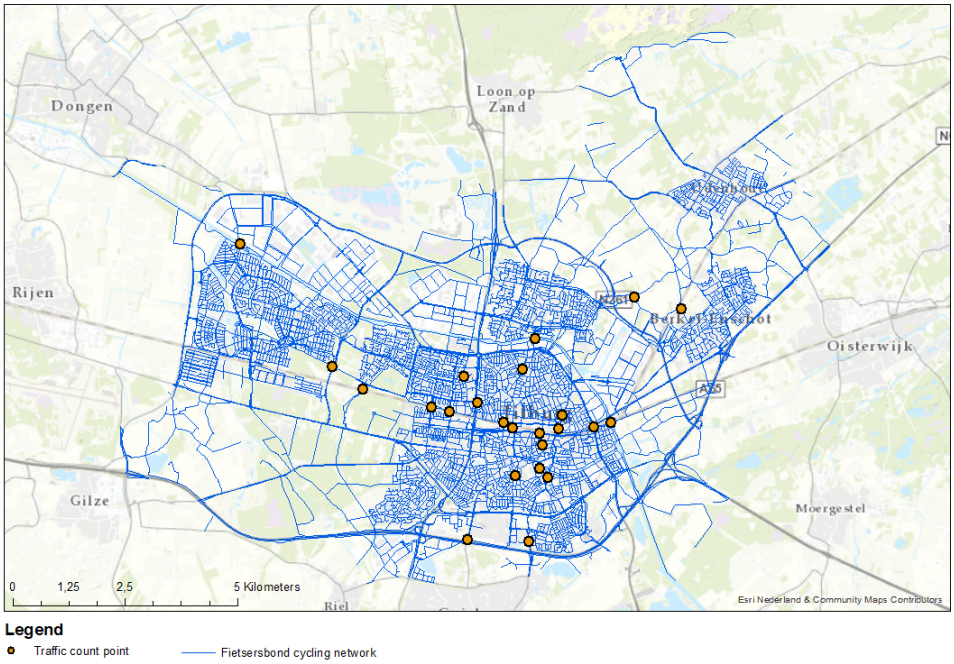


Fig. 4.5.: Overview of bicycle traffic count points from "Fietstelprogramma Gemeente Tilburg"

Since the file structure of the original Excel file is unsuitable for analysis purposes, the necessary data(in this case bicycle traffic intensity) needs to be extracted. The full script used for the extraction of the traffic count data can be found on Github and is titled *extract_count.py*. Table 4.4 below gives an overview of all the traffic count locations, along with the measured intensity, names, dates and weather obtained by performing the *extract_count.py* query.

Nr.	Location	Latitude	Longitude	Direction 1	D1 int	Direction 2	D2 int	Date	Weather
3	Piusstraat	51.551959	5.089297	Piusplein	2023	Broekhovenseweg	2196	September 15 2015	Cloudy, rain
5	B. Zwijssenstraat	51,553775	5,086699	Stadhuisplein	1640	Stadstraat	1606	September 15 2015	Cloudy, rain
9	Trouwlaan	51,55231	5,078729	Jan van de Leestraat	1832	Nieuwstraat	1754	September 15 2015	Cloudy, rain
14	Reitsehoevenstraat	51,572121	5,062	Dr. Ahausstraat	1716	Lage Witsiebaan	1523	October 8 2015	Cloudy
26	Koestraat	51,564414	5,093942	NS-Plein	1550	Leonard van Veghelstraat	1470	September 17 2015	Cloudy, rain
30	Goirkestraat	51,573582	5,081019	Wilhelminapark	1220	Julianapark	1177	October 15 2015	Cloudy, rain
40	Spoorlaan	51,560757	5,086592	Willem II straat	2642	Stationsstraat	2609	October 8 2015	Cloudy
51	Oude Lind	51,57984	5,084997	Ringbaan Noord	1336	Von Weberpad	1233	September 17 2015	Cloudy, rain
56	Spoordijk	51,563076	5,109582	Ringbaan Oost	1329	Oisterwijk	1365	October 8 2015	Cloudy
61	Zwartvenseweg	51,569399	5,029756	Bredaseweg	375	Reeshofdijk	242	October 8 2015	Cloudy
62	Statenlaan	51,565922	5,051758	Prof. Verbernelaan	1674	Wandelboslaan	1459	October 6 2015	Cloudy
63	Ringbaan West	51,566836	5,066575	Hart van Brabantlaan	641	Wandelboslaan	721	October 6 2015	Cloudy
64	St Cecliliastraat	51,562918	5,07505	Hart van Brabantlaan	2634	Jan Heijnstraat	2420	October 6 2015	Cloudy
65	Gasthuisring	51,56192	5,077813	Hart van Brabantlaan	4906	Burg. Brockxlaan	4804	October 6 2015	Cloudy
66	NS-Plein	51,561629	5,092566	Spoorlaan	4940	Enschotsestraat	4608	October 6 2015	Cloudy
67	Ringbaan Oost	51,562105	5,104033	Spoorlaan	989	Bosscheweg	820	October 6 2015	Cloudy
77	Riddershofpad	51,588119	5,116916	Hazennest	340	De Kraan/Udenhout	351	September 15 2015	Cloudy, rain
78	Rauwbrakenweg	51,58592	5,132042	Rauwbrakenweg	166	Berkel-Enschot	82	October 8 2015	Cloudy
82	Stappegoorweg	51,539023	5,083367	Oeralweg	726	Goirle	887	September 17 2015	Cloudy, rain
83	Goirleseweg	51,539298	5,063628	Ringbaan Zuid	1884	Goirle	1805	September 17 2015	Cloudy, rain
104	Pieter Vreedeplein	51,558404	5,087601	Pieter Vreedeplein	2615	Willem II straat	2225	October 8 2015	Cloudy
113	Fietsbrug Voldijk	51,598459	4,989908	Tilburg (Reeshof)	83	Industrieterrein/ Dongen	94	September 15 2015	Cloudy
118	Reeshofdijk	51,574025	5,019838	Burg Baron v V tot Vweg	1999	Heyhoef	1951	October 13 2015	Cloudy, rain
121	Academielaan	51,565079	5,05741	Academielaan	2876	Wandelboslaan	2919	October 6 2015	Cloudy

Tab. 4.4.: Overview of traffic counts

4.4 Combining the data-sources and variables

Once all the different data sources have been properly pre-processed, they need to be combined into a single file to be suitable for analysis.

4.4.1 B-riders intensities

To prevent any errors that might occur during a Spatial Join caused by small differences in spatial location between the original and mapmatched networks. The join was made by using the Source and Target values of each road segment. Both the original and the mapmatched Fietsersbond network contain these values, and they always correspond to the exact same road segment. A new column was created for both networks, containing a textual sum of the Source and Target values for each field. The reason for using this method of joining the network is that the source and target fields are not unique. For example, a source value can exist multiple times, but each time with a different target node. However, each combination of a source and target node is unique, which makes it ideal for using it as a join field. The two networks were then joined using this column as the join field, which resulted in a single Fietsersbond network containing both the original feature variables as well as the B-riders GPS intensity.

4.4.2 Traffic counts

Next, the traffic counts have to be joined to the network. They are first projected as points using the latitude and longitude columns values of each point. Several obstacles are encountered when joining the points containing the traffic counts to the network with the intensities. Firstly, as expected since they are two different data sources, the point locations do not exactly intersect the road segment. Spatially joining them to the closest road segment on the network was considered. However, this is not viable because the official traffic intensities, as mentioned earlier and shown in Appendix A, are measured separately for both directions. However, the spatial network does not always contain two separate road segments for both directions. This means that when the network only contains a single road segment for both directions, the sum of both directions(r1 and r2) needs to be added. Also, in some cases, it even occurs that no road segments exists on the network for a point location. This is simply an error that is caused due to the voluntary nature of the Fietsersbond network. Therefore, the presence of road segments was checked at the location of each road segment using the maps created for each traffic point that can be found in Appendix A. Table 4.5 shows, for each traffic point, whether the corresponding road segments exists, and in what way.

Telpunt nr.	Road segments availability
3	Separate road segment for each direction
5	1 road segment for both directions
9	1 road segment for both directions
14	1 road segment for both directions
26	1 road segment for both directions
30	Separate road segment for each direction
40	Separate road segment for each direction
51	1 road segment for both directions
56	1 road segment for both directions
61	1 road segment for both directions
62	Separate road segment for each direction
63	Separate road segment for each direction
64	Road segment does not exist
65	Only road segment for direction 2 exists
66	Only road segment for direction 2 exists
67	Separate road segment for each direction
77	1 road segment for both directions
78	Road segment does not exist
82	Separate road segment for each direction
83	Separate road segment for each direction
104	1 road segment for both directions
113	Road segment does not exist
118	1 road segment for both directions
121	Road segment does not exist

Tab. 4.5.: Availability of road segments for traffic counts

Afterward, the points containing the traffic counts are manually snapped to the corresponding road segments. When two separate road segments exist for both directions, the traffic count point is duplicated and snapped to both road segments. They were then joined to the network via a one-to-one intersecting spatial join. If, as mentioned earlier, no corresponding road segment exists for a certain traffic count point, that traffic count point is not used and removed from the data.

4.4.3 Fietstelweek intensities

Lastly, the Fietstelweek intensities need to be joined to the network. As mentioned earlier, this data source does not contain a Source and Target field, so it cannot be joined to the network the same way as the B-riders intensities. Since the 'linknummers' from this data source do not match the 'linknummers' from the B-riders dataset, joining them using those variables is also not an option. The spatial location of the network also differs very slightly from the network containing the other data, which makes a spatial join unsuitable. Therefore, the only option is to manually add the correct Fietstelweek 'linknummers' to the original network, and then join the network containing the Fietstwelweek intensities to it. This method is not ideal for larger datasets, but for the purpose of this thesis, it is still feasible since there are only 27 relevant 'linknummers' that are added.

4.5 Complete analysis file

The result of the data-preparation and preprocessing is a shapefile containing all the goal and feature variables that will be used for machine learning. Figure 4.6 on the next two pages shows all values in the analysis file. Because the intensity per day is incredibly low, and for some days even zero, it is decided to only use the total intensities of the B-Riders and Fietstelweek. Below is a list that explains what each of the headers stands for:

telpuntnr: Traffic count point number.
wegdek_num: Road surface type weight.
schoon_num: Attractivity weight.
r1_intens: Traffic count point intensity for direction 1.
r2_intens: Traffic count point intensity for direction 2.
intens_cor: Corrected traffic counts, taking into account the directions.
brid_int: Total B-riders cyclist intensity.
brid_INT0: B-riders cyclist intensity for Sundays.
brid_INT1: B-riders cyclist intensity for Mondays.
brid_INT2: B-riders cyclist intensity for Tuesdays.
brid_INT3: B-riders cyclist intensity for Wednesdays.
brid_INT4: B-riders cyclist intensity for Thursdays.
brid_INT5: B-riders cyclist intensity for Fridays.
brid_INT6: B-riders cyclist intensity for Saturdays.
ftw_int: Total Fietstelweek cyclist intensity.
FTW_INT0: Fietstelweek cyclist intensity for Sunday.
FTW_INT1: Fietstelweek cyclist intensity for Monday.
FTW_INT2: Fietstelweek cyclist intensity for Tuesday.
FTW_INT3: Fietstelweek cyclist intensity for Wednesday.
FTW_INT4: Fietstelweek cyclist intensity for Thursday.
FTW_INT5: Fietstelweek cyclist intensity for Friday.
FTW_INT6: Fietstelweek cyclist intensity for Saturday.

telpuntnr	wegdek_num	schoon_num	r1_intens	r2_intens	intens_cor	brid_int	BRID_INT0	BRID_INT1	BRID_INT2	BRID_INT3
3	3	3	2023	2196	2023	280	5	46	62	70
3	3	3	2023	2196	2196	225	5	49	41	46
5	3	3	1640	1606	3246	1382	23	273	290	269
9	3	3	1832	1754	3586	248	1	70	47	34
14	3	3	1716	1523	3239	538	4	111	124	88
26	3	3	1550	1470	3020	1281	18	284	243	192
30	3	3	1220	1177	1177	250	2	63	59	32
30	3	3	1220	1177	1220	307	3	71	72	57
40	3	3	2642	2609	2642	341	4	70	59	65
40	3	3	2642	2609	2609	230	4	48	41	38
51	3	3	1336	1233	2569	838	7	188	209	119
56	3	3	1329	1365	2694	2407	25	534	548	340
61	3	3	375	242	617	359	5	57	110	62
62	3	3	1674	1459	1459	148	5	29	28	31
62	3	3	1674	1459	1674	585	11	104	158	130
63	3	3	641	721	641	276	3	59	69	52
65	3	3	4906	4804	4804	1275	8	233	308	227
66	3	3	4940	4608	4608	1266	17	249	251	211
67	3	3	989	820	989	317	4	68	78	54
67	3	3	989	820	820	545	6	110	129	94
77	3	3	340	351	691	1123	19	234	280	129
82	3	3	726	887	726	265	9	67	60	52
82	3	3	726	887	887	310	7	86	70	62
83	3	3	1884	1805	1884	792	21	129	147	151
83	3	3	1884	1805	1805	869	14	126	187	200
104	3	3	2615	2225	4840	259	12	40	58	47
118	3	4	1999	1951	3950	4711	58	885	1092	1028

BRID INT4	BRID INT5	BRID INT6	ftw int	FTW INT0	FTW INT1	FTW INT2	FTW INT3	FTW INT4	FTW INT5	FTW INT6
47	41	9	15	1	2	4	0	3	4	2
45	30	9	25	2	4	3	5	5	3	2
296	206	25	54	8	5	14	10	9	8	1
72	20	4	69	1	8	19	9	15	13	6
126	78	7	26	0	3	6	4	7	3	1
273	233	38	36	4	3	7	7	5	8	3
46	43	5	32	3	3	7	3	8	5	0
57	41	6	32	3	3	7	3	8	5	0
75	58	10	6	1	2	1	0	0	1	0
50	42	7	12	1	0	0	3	0	5	1
151	150	14	49	4	10	13	5	10	13	1
551	378	31	93	5	9	14	15	16	16	16
103	19	3	24	2	4	2	5	5	2	4
33	17	5	159	8	29	30	24	28	25	11
120	56	6	32	7	3	6	6	6	3	2
64	25	4	5	2	0	0	1	1	0	1
300	179	20	76	3	16	17	9	13	18	3
272	228	38	52	4	3	10	8	11	7	6
72	36	5	19	0	2	4	1	5	3	0
130	66	10	9	2	0	2	1	3	1	0
216	217	28	32	1	6	8	4	7	7	1
52	23	2	22	3	1	7	3	2	4	2
53	23	9	12	1	2	1	2	5	1	0
162	152	30	37	1	2	9	6	6	6	8
196	126	20	44	3	4	5	11	8	10	4
51	42	9	48	4	6	10	9	11	8	2
923	648	77	140	10	26	28	27	27	16	6

Fig. 4.6.: File resulting from pre-processing that is used for analysis

Analysis and results

5

In this chapter, the results of the analysis described in the Methodology chapter are detailed. First, the ratio between the GPS intensities(For both B-riders and Fietstelweek) is calculated and discussed to see whether it is plausible given the spatial background. Next, the correlation between the goal variable(traffic counts) and the feature variables is shown using scatterplots and correlation coefficients. Finally, for each machine learning classifier specified in chapter 3.3, the outcomes of the cross-validation are given, along with the regressor quality(r2 score) and the MSE.

5.1 Plausibility and ratio of GPS intensities

When looking at the pre-processed file containing all values(figure 4.6), at first glance, it looks like there is very little correlation between the GPS intensities of the B-riders and Fietstelweek and the traffic counts. To see whether or not the intensities make sense given the spatial background, both the B-riders and Fietstelweek are visualized, based on four quantiles, in figures 5.1 and 5.2 against a background map of the municipality of Tilburg. Note that all road segments have been visualized to get a better understanding of the plausibility, and not just the 27 that also contain traffic count data and being used in the analysis.

Both figures show relatively believable and plausible intensities given the spatial background. The main roads contain the highest intensities, while the smaller roads within each neighborhood have lower intensities. It is logical that the GPS intensities are (significantly) lower than the actual measured traffic counts, since obviously only a minimal amount of the people cycling on a certain road will be participating in either the B-riders of Fietstelweek project, even though the B-riders project collects data over a period of several months. In an attempt to (partly) solve this problem, the ratio is calculated between the traffic counts and the GPS intensities. Table 5.1 shows the ratio between the traffic counts and the B-riders intensities.

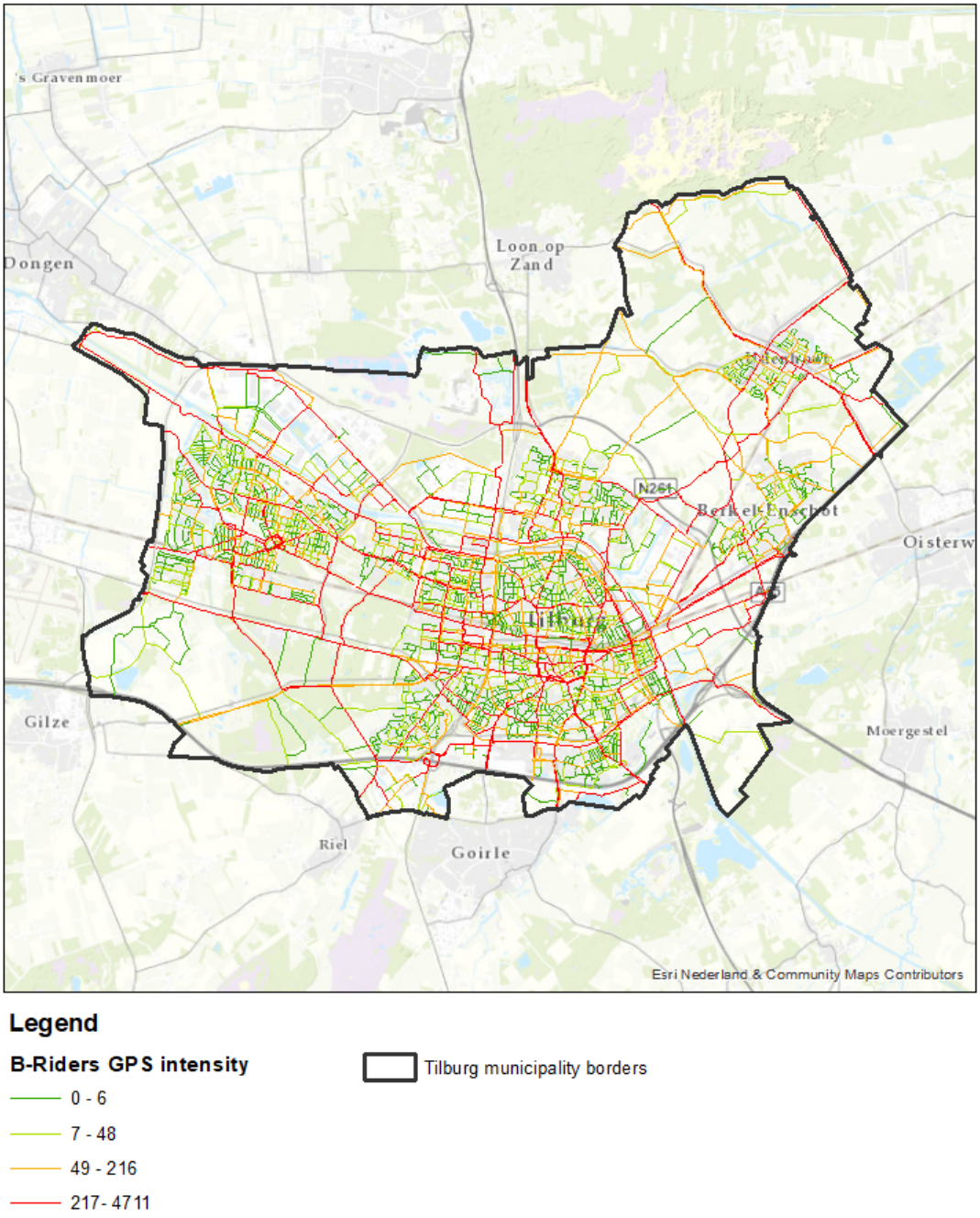


Fig. 5.1.: B-riders intensities against Tilburg background map

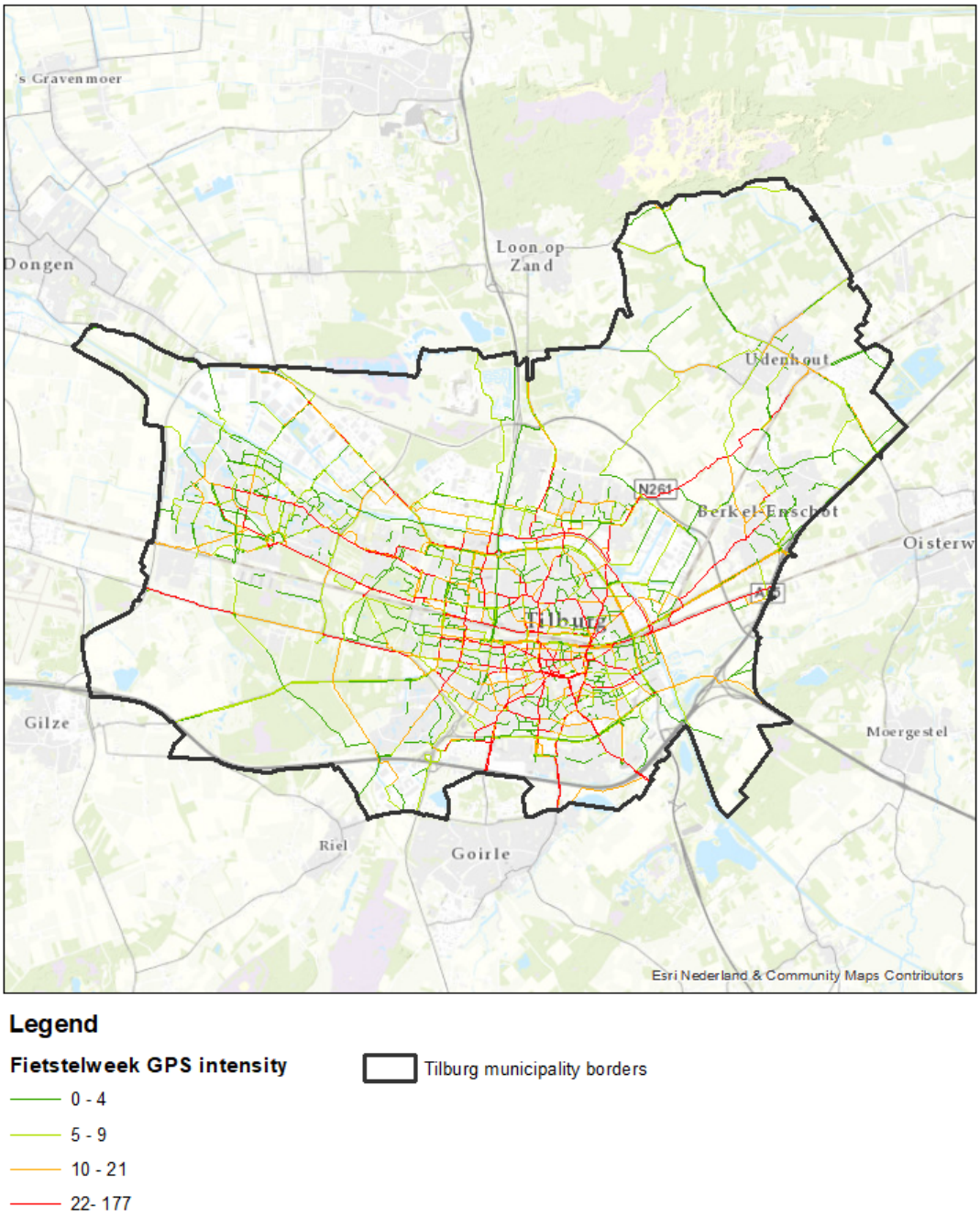


Fig. 5.2.: Fietstelweek intensities against Tilburg background map

Telpunt nr.	Traffic count	B-Riders int.	Ratio	Corrected B-riders int.
3	2023	280	7,225	1381
3	2196	225	9,76	1109
5	3246	1382	2,34877	6814
9	3586	248	14,4597	1223
14	3239	538	6,02045	2653
26	3020	1281	2,35753	6316
30	1177	250	4,708	1233
30	1220	307	3,97394	1514
40	2642	341	7,7478	1681
40	2609	230	11,3435	1134
51	2569	838	3,06563	4132
56	2694	2407	1,11924	11867
61	617	359	1,71866	1770
62	1459	148	9,85811	730
62	1674	585	2,86154	2884
63	641	276	2,32246	1361
65	4804	1275	3,76784	6286
66	4608	1266	3,63981	6242
67	989	317	3,11987	1563
67	820	545	1,50459	325
77	691	1123	0,615316	274
82	726	265	2,73962	288
82	887	310	2,86129	352
83	1884	792	2,37879	747
83	1805	869	2,0771	716
104	4840	259	18,6873	1919
118	3950	4711	0,838463	1566
Mean ratio: 4,930382				

Tab. 5.1.: Ratio between traffic counts and B-riders intensities, along with corrected intensities

The minimum ratio between the traffic counts and the B-riders intensities is 0,615316, while the maximum ratio is 18,6873. One thing that becomes clear, is that there are two occurrences where the B-riders intensity is higher than the traffic count intensity. The mean of all ratios is 4,930382, and the 'corrected' B-riders intensities are calculated by multiplying the original B-riders intensities with this average ratio. To better understand the ratio's between the B-riders intensities and traffic counts, figure 5.3 shows the spatial distribution of the ratio.

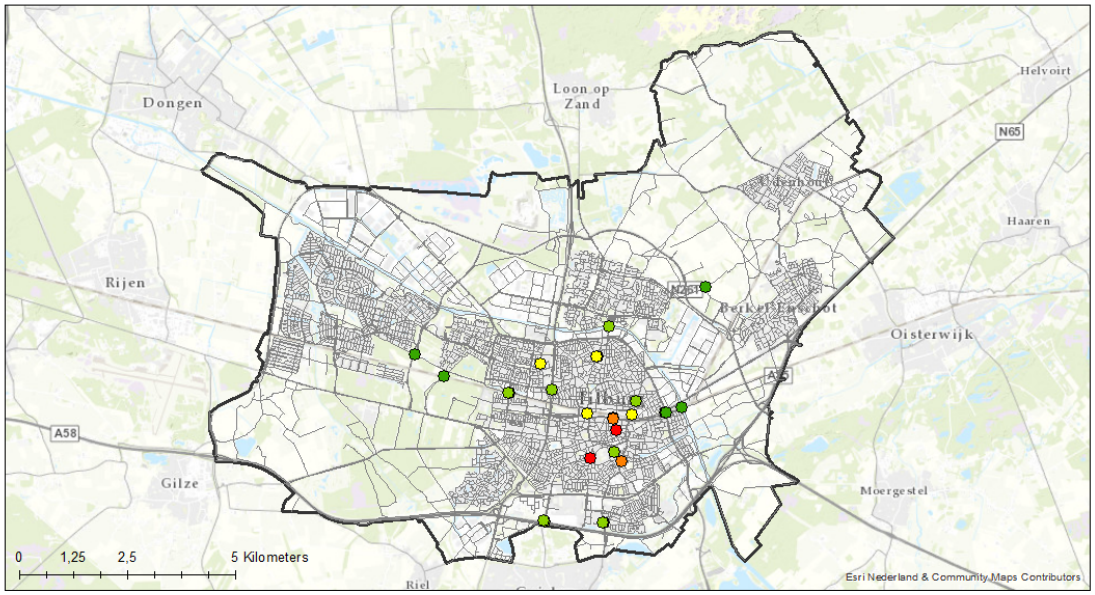


Fig. 5.3.: Map of ratio between traffic counts and B-riders intensities

The spatial distribution shows that, in general, the largest ratio's between the traffic count and the B-riders intensities occur near the city center, while the lower ratio's mostly are further away from the city centre. This makes sense since there more people cycling near the city center. Therefore, the effect of the relatively small sample size of B-riders participants becomes more clear there.

Table 5.2 on the next page shows the ratio between the traffic counts and Fietstelweek intensities.

Telpunt nr.	Traffic count	Fietstelweek int.	Ratio	Corrected Fietstelweek int.
3	2023	15	134,867	1232
3	2196	25	87,84	2053
5	3246	54	60,1111	4434
9	3586	69	51,971	5666
14	3239	26	124,577	2135
26	3020	36	83,8889	2956
30	1177	32	36,7813	2628
30	1220	32	38,125	2628
40	2642	6	440,333	493
40	2609	12	217,417	985
51	2569	49	52,4286	4024
56	2694	93	28,9677	7637
61	617	24	25,7083	1971
62	1459	159	9,1761	13057
62	1674	32	52,3125	2628
63	641	5	128,2	411
65	4804	76	63,2105	6241
66	4608	52	88,6154	4270
67	989	19	52,0526	1560
67	820	9	91,1111	739
77	691	32	21,5938	2628
82	726	22	33	1807
82	887	12	73,9167	985
83	1884	37	50,9189	3038
83	1805	44	41,0227	3613
104	4840	48	100,833	3942
118	3950	140	28,2143	11497
Mean ratio: 82,118279				

Tab. 5.2.: Ratio between traffic counts and Fietstelweek intensities, along with corrected intensities

The minimum ratio between the traffic counts and the Fietstelweek intensities is 9,1761, while the maximum ratio is 440,333. The ratios between the Fietstelweek intensities and the traffic counts are much higher than the ratios between the B-riders intensities and the traffic counts. This is because the Fietstelweek, as implied by the name, only ran for a week while the B-riders program ran for several months. Therefore, the intensities for the B-riders, and also the ratios, are much higher than for the Fietstelweek. Once again, to better understand the ratio's between the Fietstelweek intensities and traffic counts, figure 5.4 shows the spatial distribution of the ratios.

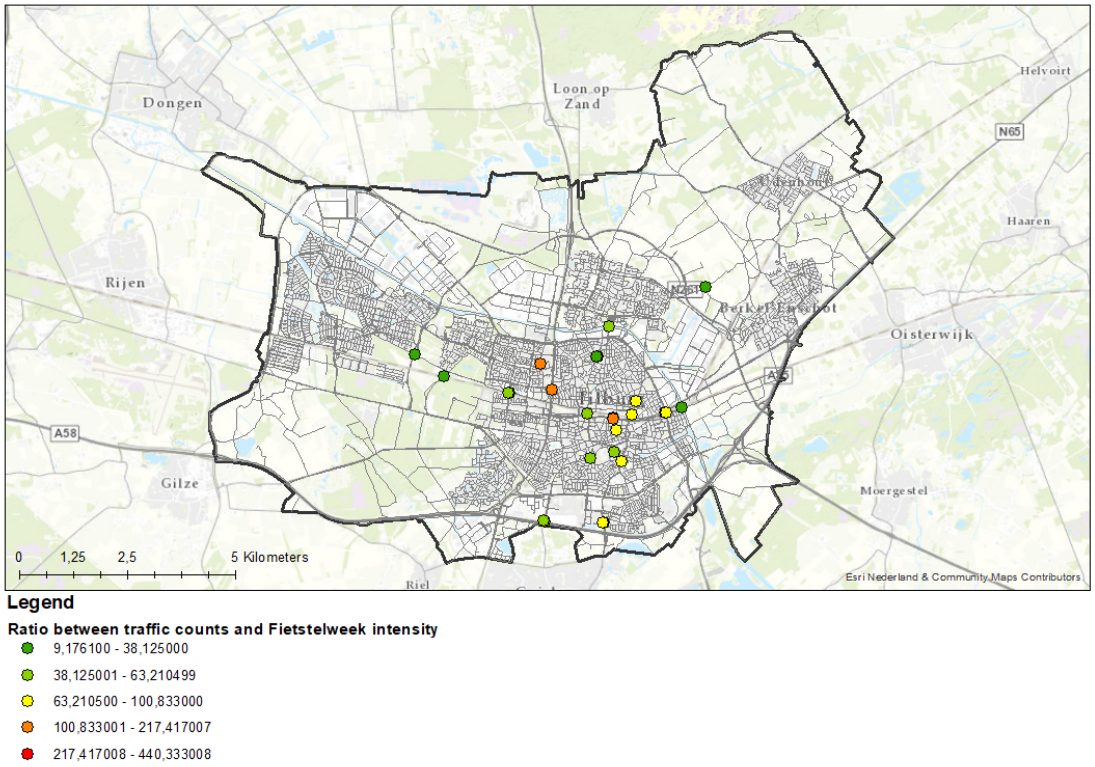


Fig. 5.4.: Map of ratio between traffic counts and Fietstelweek intensities

The pattern shown in the spatial distribution of the ratios is pretty similar to that of the B-riders intensities. The largest ratio's occur near the city centre, while further away from the city centre the ratio's are significantly lower. The same explanation can be given here, in that there are simply more people cycling near the city centre, and thus the effect of the relatively small sample size of Fietstelweek participants becomes more clear.

5.2 Correlation between variables

To show the correlation between the goal variable and all feature variables for which it is relevant, scatterplots are made, along with the correlation coefficient. The script used to generate the scatterplots and calculate the correlation coefficient is named *scatterplot.py* and can be found on Github.

5.2.1 Correlation between traffic counts and B-riders GPS intensities

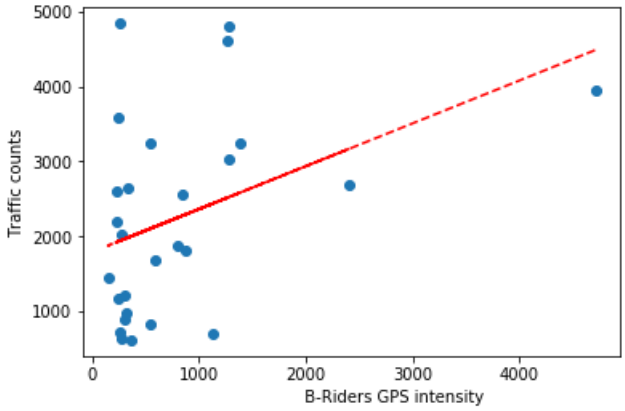


Fig. 5.5.: Scatterplot of traffic counts and B-riders GPS intensities

The correlation coefficient between the traffic counts and the B-riders GPS intensities is 0.40716848131988526, which means the two variables have a low positive correlation, as shown by the line of best fit.

5.2.2 Correlation between traffic counts and corrected B-riders GPS intensities

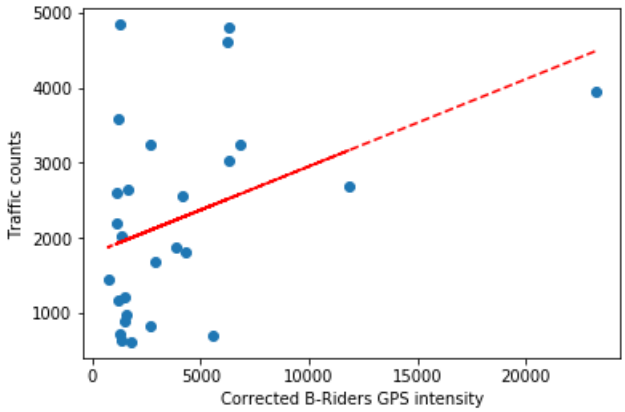


Fig. 5.6.: Scatterplot of traffic counts and corrected B-riders GPS intensities

The correlation coefficient between the traffic counts and the corrected B-riders GPS intensities is 0.4071658116735912, which means the two variables have a low positive correlation, as shown by the line of best fit. The scatterplot, and thus the correlation coefficient are almost identical to the non-corrected B-riders GPS intensities which means that the correction sorted very little effect.

5.2.3 Correlation between traffic counts and Fietstelweek GPS intensities

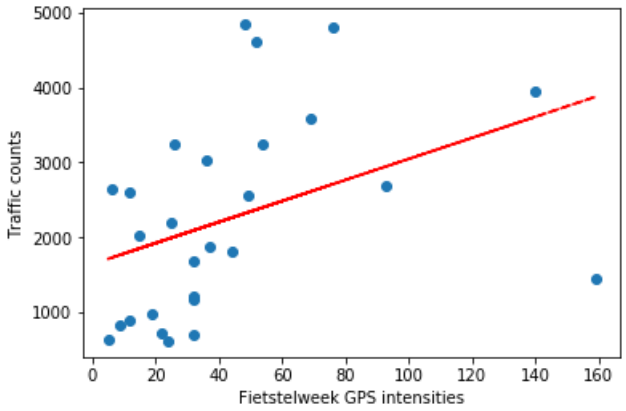


Fig. 5.7.: Scatterplot of traffic counts and Fietstelweek GPS intensities

The correlation coefficient between the traffic counts and the Fietstelweek GPS intensities is 0.39785878228848215, which means the two variables have low a positive correlation, as shown by the line of best fit. The correlation coefficient is only slightly lower than the correlation coefficient between the traffic counts and the B-riders GPS intensities.

5.2.4 Correlation between traffic counts and corrected Fietstelweek GPS intensities

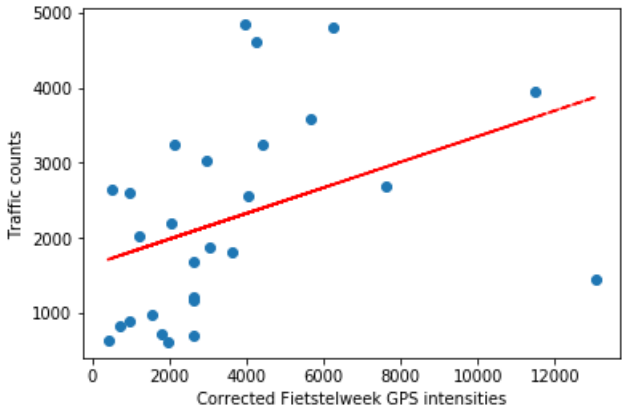


Fig. 5.8.: Scatterplot of traffic counts and corrected Fietstelweek GPS intensities

The correlation coefficient between the traffic counts and the corrected Fietstelweek GPS intensities is 0.39784335386865616, which means the two variables have a low positive correlation, as shown by the line of best fit. Just like with the corrected B-riders intensities, the scatterplot and correlation coefficient from the corrected Fietstelweek barely differ from the normal Fietstelweek intensities.

5.2.5 Correlation between traffic counts and Attractivity

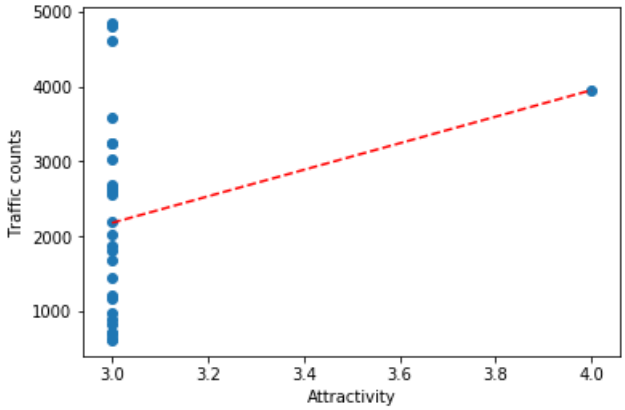


Fig. 5.9.: Scatterplot of traffic counts and Attractivity

The correlation coefficient between the traffic counts and the corrected Fietstelweek GPS intensities is 0.25710286957749356, which means the two variables have a low positive correlation, as shown by the line of best fit. However, all but one of the attractivity values are the same.

5.2.6 Correlation between traffic counts and Road surface type

The correlation coefficient between the traffic counts and the corrected Fietstelweek GPS intensities is 0, since all cases of road surface type in the dataset have the same value(3). This means it is not possible to calculate a line of best fit, and therefore it is also not relevant to create a scatterplot.

5.3 Machine learning results

In this section, the results for each of the machine learning classifiers are listed. Firstly, for each classifier, 5-fold cross validation is performed for the following four variations of goal and target variables

- **Goal variable:** Traffic counts. **Feature variable:** Attractivity, Road surface type, Spatial location and B-riders intensities
- **Goal variable:** Traffic counts. **Feature variable:** Attractivity, Road surface type, Spatial location and corrected B-riders intensities
- **Goal variable:** Traffic counts. **Feature variable:** Attractivity, Road surface type, Spatial location and Fietstelweek intensities
- **Goal variable:** Traffic counts. **Feature variable:** Attractivity, Road surface type, Spatial location and corrected Fietstelweek intensities

Due to the intensities per day being very low(sometimes even being zero) as shown in the previous chapter, it has been decided to only use the total GPS intensities and the corrected GPS intensities when it comes to GPS intensities.

Next to 5-fold cross-validation, LOOCV is performed for each classifier, again using the four variations of goal and target variables mentioned above. The python script used to perform the machine learning, which contains all the parameter settings for each of the classifiers, is named *MP.py* and can be found on Github.

5.3.1 k-Nearest Neighbor

Table 5.3 shows the results of 5-fold cross validation for the k-Nearest Neighbor regressor with k=5. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.4 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	-0.25940	1371820
2	-4.72266	1461839
3	-1.70481	1585510
4	-7.22015	1877289
5	-4.77737	5980563
R2 mean: -3.73688 MSE mean: 2455404		

Tab. 5.3.: KNN with B-riders intensities

Fold	R2	MSE
1	-0.84682	2420899
2	-1.83817	1429990
3	-4.79975	2510248
4	-25.91648	1869795
5	-20.56347	3740681
R2 mean: -10.79294 MSE mean: 2394323		

Tab. 5.4.: KNN with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.5 shows the results using the normal intensities, while table 5.6 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-0.19152	1332838
2	-4.47796	1931572
3	-3.41385	3560650
4	-12.11435	1846075
5	-5.63278	4812176
R2 mean: -5.16609 MSE mean: 2696662		

Tab. 5.5.: KNN with Fietstelweek intensities

Fold	R2	MSE
1	0.33837	1160411
2	-0.68242	2167534
3	-7.65212	2038180
4	0.17835	1304644
5	-18.14355	1916950
R2 mean: -5.19227 MSE mean: 1717544		

Tab. 5.6.: KNN with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.2 Gaussian Process

Table 5.7 shows the results of 5-fold cross validation for the Gaussian Process regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.8 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	0	7001083
2	0	7099107
3	0	3132607
4	0	12584308
5	0	3713373
R2 mean: 0 MSE mean: 6706095.746666667		

Tab. 5.7.: GP with B-riders intensities

Fold	R2	MSE
1	0	7001083
2	0	7099107
3	0	3132607
4	0	12584308
5	0	3713373
R2 mean: 0 MSE mean: 6706095.746666667		

Tab. 5.8.: GP with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.9 shows the results using the normal intensities, while table 5.10 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-2.62080e+60	7001083
2	0	7099107
3	-7.25514e+172	3132607
4	-4.81858e+60	12584308
5	0	3713373
R2 mean: -1.45103e+172 MSE mean: 6706096		

Tab. 5.9.: GP with Fietstelweek intensities

Fold	R2	MSE
1	0	7001083
2	0	7099107
3	0	3132607
4	0	12584308
5	0	3713373
R2 mean: 0 MSE mean: 6706096		

Tab. 5.10.: GP with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative or zero, and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.3 Decision Tree

Table 5.11 shows the results of 5-fold cross validation for the Decision Tree regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.12 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	-0.79775	3611403
2	-3.85664	4041753
3	-0.40795	3821205
4	-1.16930	2458753
5	-2.88759	7468864
R2 mean: -1.82384 MSE mean: 4280396		

Tab. 5.11.: DT with B-riders intensities

Fold	R2	MSE
1	-0.78870	3676881
2	-3.85664	4041753
3	-0.18176	2166165
4	-0.84791	4275047
5	-2.11561	8429499
R2 mean: -1.55812 MSE mean: 4517869		

Tab. 5.12.: DT with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.13 shows the results using the normal intensities, while table 5.14 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	0.38965	1836020
2	-6.28379	5413014
3	-7.27120	2912464
4	-2.46860	3299555
5	-17.01741	3180876
R2 mean: -6.53027 MSE mean: 3328386		

Tab. 5.13.: DT with Fietstelweek intensities

Fold	R2	MSE
1	0.36565	1833215
2	-9.35961	3900655
3	-8.80182	2693368
4	-2.52781	3209739
5	-12.70382	2896827
R2 mean: -6.60548 MSE mean: 2906761		

Tab. 5.14.: DT with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.4 Kernel Ridge

Table 5.15 shows the results of 5-fold cross validation for the Kernel Ridge regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.16 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	-7.68317	1699586
2	-10.52080	2156969
3	-15.26225	846566
4	-40.40558	2825879
5	-268.82810	3043283
R2 mean: -68.53998 MSE mean: 2114457		

Tab. 5.15.: KR with B-riders intensities

Fold	R2	MSE
1	-7.68525	1699321
2	-10.52443	2156999
3	-15.25248	846347
4	-40.39021	2825989
5	-268.65707	3043290
R2 mean: -68.50189 MSE mean: 2114389		

Tab. 5.16.: KR with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.17 shows the results using the normal intensities, while table 5.6 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-6.95878	1237147
2	-1.11477	5225868
3	-17.71480	785940
4	-23.66280	2892578
5	-22.72679	3675570
R2 mean: -14.43559 MSE mean: 2763421		

Tab. 5.17.: KR with Fietstelweek intensities

Fold	R2	MSE
1	-6.95707	1237252
2	-1.11464	5226649
3	-17.70188	785678
4	-23.67010	2892727
5	-22.72938	3675326
R2 mean: -14.43462 MSE mean: 2763527		

Tab. 5.18.: KR with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.5 Support Vector

Table 5.19 shows the results of 5-fold cross validation for the Support Vector regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.20 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	0	1433260
2	0	1850096
3	0	1315474
4	0	3879676
5	0	2364163
R2 mean: 0 MSE mean: 2168534		

Tab. 5.19.: SV with B-riders intensities

Fold	R2	MSE
1	0	1433260
2	0	1850096
3	0	1315474
4	0	3879676
5	0	2364163
R2 mean: 0 MSE mean: 2168534		

Tab. 5.20.: SV with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.21 shows the results using the normal intensities, while table 5.22 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-5.54466e+30	1433259
2	-3.57861e+31	1850096
3	0	1315474
4	-1.17256e+31	3879676
5	0	2364163
R2 mean: -1.06112e+31 MSE mean: 2168534		

Tab. 5.21.: SV with Fietstelweek intensities

Fold	R2	MSE
1	0	1433259
2	0	1850096
3	0	1315474
4	0	3879676
5	0	2364163
R2 mean: 0 MSE mean: 2168534		

Tab. 5.22.: SV with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative or zero, and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.6 Partial Least Squares

Table 5.23 shows the results of 5-fold cross validation for the Partial Least Squares regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.24 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	-4.92793	2197322
2	-5.22090	2567818
3	-9.86570	781707
4	-94.65645	2933716
5	-65.30736	3609364
R2 mean: -35.99567 MSE mean: 2417985		

Tab. 5.23.: PLS with B-riders intensities

Fold	R2	MSE
1	-4.92797	2197294
2	-5.22250	2567711
3	-9.86515	781730
4	-94.64255	2933751
5	-65.31294	3609389
R2 mean: -35.99422 MSE mean: 2417975		

Tab. 5.24.: PLS with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.25 shows the results using the normal intensities, while table 5.26 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-4.58224	1867226
2	-0.92840	7425289
3	-13.72738	732240
4	-30.78669	2941629
5	-11.14462	4888783
R2 mean: -12.23387 MSE mean: 3571033		

Tab. 5.25.: PLS with Fietstelweek intensities

Fold	R2	MSE
1	-4.58206	1867344
2	-0.92845	7425447
3	-13.73291	732291
4	-30.79012	2941671
5	-11.14404	4888880
R2 mean: -12.23551 MSE mean: 3571126		

Tab. 5.26.: PLS with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

5.3.7 Linear Regression

Table 5.27 shows the results of 5-fold cross validation for the Linear regressor. The traffic counts are used as goal variables, while the attractivity, road surface type, spatial distance and the B-riders intensities are used as feature variables. Meanwhile, table 5.28 shows the results with the ratio-corrected B-riders intensities instead of the original B-riders intensities.

Fold	R2	MSE
1	-4.76149	2030331
2	-3.41002	3958771
3	-3.85971	777847
4	-32.30871	2767496
5	-74.19664	3948768.87045
R2 mean: -23.70732 MSE mean: 2696643		

Tab. 5.27.: LR with B-riders intensities

Fold	R2	MSE
1	-4.761605	2030284
2	-3.41052	3958306
3	-3.85939	777900
4	-32.30006	2767536
5	-74.18711	3948829
R2 mean: -23.70373 MSE mean: 2696571		

Tab. 5.28.: LR with corrected B-riders intensities

The tables below show the results when using the Fietstelweek intensities instead of the B-riders intensities. Table 5.29 shows the results using the normal intensities, while table 5.30 shows the results using the ratio-corrected intensities.

Fold	R2	MSE
1	-4.67512	1908814
2	-1.38704	9249897
3	-13.56084	725759
4	-34.43537	2947422
5	-11.00408	4912970
R2 mean: -13.01249 MSE mean 3948973		

Tab. 5.29.: LR with Fietstelweek intensities

Fold	R2	MSE
1	-4.67504	1908892
2	-1.38706	9249988
3	-13.56626	725811
4	-34.43856	2947460
5	-11.00360	4913050
R2 mean: -13.01410 MSE mean: 3949040		

Tab. 5.30.: LR with corrected Fietstelweek intensities

Since for all four variations of feature variables the R2 mean is negative and the MSE is very high, it can be concluded that this regressor is unsuitable for predicting traffic intensities based on the given feature variables. When performing LOOCV, each of the four variations of feature variables results in an R2 score of 0, which leads to the conclusion that the model predicts as good as a naive model that has a constant value.

Conclusion

6

In this chapter, the main conclusions and new insights that can be gathered from the results are described by answering the main research question and sub-questions that were established at the beginning of this research. It also contains a discussion that gives suggestions for potential future research.

6.1 Results in the context of research questions

The research goal of this thesis has been to find ways to estimate traffic intensities of cyclists on a network using interpolation between local traffic counts and intensities from GPS tracks. This thesis attempts to answer the research question as described in chapter 1.3, which is as follows:

"*How can the traffic intensity of cyclists on a network be estimated by means of flow interpolation from local traffic counts and GPS tracks?*"

Four sub-questions have been formulated to help answer the main research question, with each sub-question focusing on a specific area of the main research question. Therefore, each sub-question is answered below using the results and insights gained from the research. Together, all these answers can be used to answer the main research question.

Results on suitable methods for estimating traffic intensity
The first sub-question encompassed finding out which methods are suitable for estimating traffic intensity of cyclists based on local traffic counts and GPS tracks. According to literature, gravity models are still the most used models but are unsuitable for this type of estimating since they do not fit well to local measurements, which the traffic counts are. Therefore, machine learning methods are used since they do fit well to local measurements. Since flow interpolation is a regression problem, regression classifiers are suitable for estimating the traffic intensity. All seven used regression classifiers perform equal or worse than the naive model, which makes them all unsuitable estimating traffic intensity in the context of this thesis. This can be explained given the lack of correlation that was shown between the data. This means that either the used data sample is too small, or that the GPS data is too unrepresentative.

Results on road characteristics as feature variables
Next, the second sub-question deals with which road characteristics should be taken into account as feature variables. According to literature, road surface type, the width of a road and the attractivity all influence people behavior and route change when cycling and are therefore suitable variables. All of these road characteristics are available in the existing road network and used as features variables in the machine learning models, with the exception the road width, which turned out to not be available on an acceptable level in both of the used networks. The pre-processing and analysis shows that both the road width and the attractiveness consist of the same value for (almost) all of the road segments and that there is little to no correlation. Therefore, within the scope of this research, they do not provide much added value as feature variables. However, when executing the machine learning models on a much larger scale with many roads having varying attractiveness and road surfaces, they might turn out to be useful as feature variables.

Results on the usefulness of biased GPS tracks as a variable
Regarding the third sub-question, which focuses on the usefulness of biased GPS tracks as a variable in the machine learning model, several things can be concluded. First, the results of the machine learning regression show that for the scope of this research, the GPS tracks of both the B-riders as well as those of the Fietstelweek are not suitable as variables. For all seven machine learning classifiers, the model performed equal or worse than the naive model which predicts with a constant value. This is also the case even when taking into account the different measurement periods by looking at the ratios.

Results on the validation of machine learning algorithms
The last sub-question deals with the validation of the machine learning algorithms. Literature shows that, when using regression methods to estimate, the regression score(r2) and the mean square error(MSE) are suitable. The r2 scores and the MSE for all of the machine learning classifiers can be used to compare them. However, as said before, the r2 score for each of the machine learning classifiers is zero or negative, which means that they are all almost equally unsuitable. Meanwhile, the MSE scores are very high for all classifiers, which means that the predicted traffic intensity differs significantly from the actual value.

To conclude, regarding the main research question, one can say that, when using the B-riders and Fietstelweek GPS data, it is not possible to accurately predict the traffic intensity of cyclists on a network scale. Machine learning regression might still be a suitable method for predicting traffic intensities, but to achieve more desirable results, a significantly larger and less biased sample of traffic count points than the amount used in this research is needed. Also the GPS tracks need to be representative and have a significantly larger sample size than the ones used during this research in order to achieve more desirable results.

6.2 Discussion

When a research project is finished, it is important to look back and reflect on the results and process of the research itself and discuss the new insights that became clear. By doing this, any encountered shortcomings or limitations can be discussed, which can offer help and possibilities for potential feature research on this topic. In this section, several limitations that were encountered during the research are discussed. Based on these encountered limitations, several suggestions for future research are made.

The results and conclusion of this thesis show that the used GPS tracks are not suitable for making traffic flow estimations. However, as mentioned in the introduction, there still lies a lot of potential in combining new data sources such as GPS with traditional traffic counts. The major limitation turned out be that the GPS tracks of the B-Riders and Fietstelweek are not representative for the average behavior of cyclists. A potential solution to this could be to de-bias or re-sample the GPS tracks to obtain more representative data. Another possible solution is to greatly increase the sample size of GPS tracks, which could lead to a reducing of the bias.

It also became clear that there is a danger in using non-representative GPS samples, which has to be kept in mind during this kind of research. In the context of this thesis, it became very obvious that the GPS samples were non-representative and biased because of the negative R^2-scores and very large MSE. However, it is unlikely that the non-representativeness of GPS sample is always shown as clearly as it is in this thesis. Using non-representative GPS samples could therefore lead to drawing the wrong conclusion if the researcher is not fully aware that the GPS samples are non-representative.

Another limitation that arose during the the research is that, while according to existing literature, the spatial distance may have an influence, this research ended up using spatial location instead of spatial distance. Due to only using the centroids of the road segments as input features and not the distance between them, the spatial location input feature may not correctly line up with literature.

6.2.1 Suggestions for future research

As shown in the previous section, several limitations and points for improvement were encountered during this research. The suggestions below might be useful to take into account for future research about estimating traffic intensity of cyclists on a network using GPS data and local traffic counts, to prevent the same limitations and shortcomings that happened during this research. Taking these suggestions into account might result in models that can estimate the intensity of cyclists.

Larger sample size for traffic count data
An important point for potential improvement concerns the 'official' traffic count data. For this research, only 24 locations were able to be used, from the 'Fietstelprogramma Gemeente Tilburg', of which several had to be removed due to empty values or non-existent data. While it turned out to still be possible to perform analysis and interpolation using so few data points, it is far from ideal. Attempts were made to add additional traffic count data points from other data sources as the province of North-Brabant or TNO, but since these were measured over different time periods and during different years and dates, the data did not correlate. Therefore the first suggestion is to, if feasible, significantly increase the sample size of bicycle traffic count locations. In addition to increasing the sample size, to get a representative view of the cycling intensities of a certain city or region, the spatial distribution of traffic count points is important. The 24 points used were not ideally distributed, with most of them being located near the busy city center, which can lead to bias. Having a better spatial distribution of traffic counts, with them being evenly distributed, would increase the believability of the analysis and lead to it being less biased.

Additional spatial input features
Furthermore, an interesting aspect that future research could look in to is a more comprehensive use of spatial input features. In the end, for this thesis, the spatial location of each traffic count point was used by looking at each traffic count points X and Y coordinates. This is different from using the spatial distance or network distance, and using these values as input features might provide different results that provide new insights into the use of GPS tracks for traffic flow measurement. An interesting possibility might be to calculate the distance to the city center(which could be the market, central station or town hall for example) for each traffic count point, and use it as an additional input feature during the machine learning.

Increased amount of feature variables
The machine learning algorithms used four variables for estimating the cyclist traffic intensity, mostly due to the availability of these variables. However, by adding additional variables, the performance of the model might be improved. Some other features that could be added to the model and that might improve the model are socioeconomic features or safety features such as the number of traffic lights. It could also be interesting to use more than one variable for certain input features. For example, the input feature attractivity could be represented by two different variables, one from the Fietsersbond network and one from the OSM network.

Different parameters for machine learning algorithms
The machine learning algorithms that were used during this research were used using mostly their default parameters. By looking at the effect that changing these parameters has, the ability of the machine learning classifiers to estimate cyclists could be improved.

Bibliography

Anderson, J. E. (2010). The gravity model. Working Paper 16576, National Bureau of Economic Research.

B-Riders (2017). Zo werkt b-riders company. http://www.briders.nl/zo-werkt-b-riders-company. [Date accessed: 17-10-2017].

Bussche, D. and van de Coevering, P. (2015). Bikeprint: in depth analysis of cyclist behaviour and cycle network/ performance using extensive gps track data.

Casey, H. J. (1955). Applications to traffic engineering of the law of retail gravitation. *Traffic Quarterly*, 9(1):23–35.

Cerná, A., Černý, J., Malucelli, F., Nonato, M., Polena, L., and Giovannini, A. (2014). Designing optimal routes for cycle-tourists. *Transportation Research Procedia*, 3:856 – 865.

CROW-Fietsberaad (2015). Fietsen en lopen smeerolie van onze mobiliteit. *Fietsverkeer*, 37:6–7.

CROW-Fietsberaad (2017). Bicycle agenda 2017-2020. http://www.fietsberaad.nl/index.cfm?lang=nl&repository=Bicycle+Agenda+2017+2020. [Date accessed: 19-10-2017].

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press.

Fietsersbond (2017a). 100 miljoen niet genoeg voor grote ambities fietsbeleid. https://www.fietsersbond.nl/nieuws/100-miljoen-genoeg-grote-ambities-op-fietsbeleid/. [Date accessed: 09-11-2017].

Fietsersbond (2017b). Fietersbond Routeplanner. https://routeplanner.fietsersbond.nl/pagina/handleiding. [Date accessed: 11-11-2017].

Fietstelweek (2017). Fietswelweek. http://fietstelweek.nl/data/faq/. [Date accessed: 06-11-2017].

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Harvey, F. J. and Krizek, K. (2007). Commuter bicyclist behavior and facility disruption. https://www.lrrb.org/pdf/200715.pdf. [Date accessed: 02-12-2017].

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc.

Klinkenberg, J. and Bertolini, L. (2012). There are still opportunities for Dutch cycling. pages 1–9.

Leduc, G. (2008). Road traffic data: Collection methods and applications.

May, M., Scheider, S., Rösler, R., Schulz, D., and Hecker, D. (2008). Pedestrian flow prediction in extensive road networks using biased observational data. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08*.

Necula, E. (2014). Dynamic traffic flow prediction based on gps data. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 922–929.

Netherlands Organisation for Applied Scientific Research (2014). Fietsmonitor Zuid-Nederland - Intensiteiten schatten.

Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness.

Noland, R. B. and Kunreuther, H. (1995). Short-run and long-run policies for increasing bicycle transportation for daily commuter trips. *Transport Policy*, 2(1):67 – 79.

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`. [Date accessed: 09-11-2017].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petritsch, T., Landis, B., Huang, H., and Challa, S. (2006). Safety Model: Bicycle Sidepath Design Factors Affecting Crash Rates. *Transportation Research Record: Journal of the Transportation Research Board*, 1982:194–201.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM - Supporting community and building social capital*, 45(4):211–218.

Pucher, J. and Buehler, R. (2008). Making cycling irresistible: Lessons from the netherlands, denmark and germany. *Transport Reviews*, 28(4):495–528.

Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., and Srivastava, M. (2010). Biketastic: Sensing and mapping for better biking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1817–1820. ACM.

Rijksoverheid (2017). Vertrouwen in de toekomst: Regeerakkoord 2017-2021. `https://www.kabinetsformatie2017.nl/documenten/publicaties/2017/10/10/regeerakkoord-vertrouwen-in-de-toekomst`. [Date accessed: 18-10-2017].

Rodrigue, J.-P. (2017). *The geography of transport systems*. Routledge, Taylor & Francis Group.

Romanillos, G., Zaltz Austwick, M., Ettema, D., and De Kruijf, J. (2016). Big Data and Cycling. *Transport Reviews*, 36(1):114–133.

Scikit-learn (2017a). Decision Trees. `http://scikit-learn.org/stable/modules/tree.html#tree`. [Date accessed: 06-11-2017].

Scikit-learn (2017b). Gaussian Processes. `http://scikit-learn.org/stable/modules/gaussian_process.html`. [Date accessed: 02-12-2017].

Scikit-learn (2017c). Nearest Neighbors Regression. `http://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-regression`. [Date accessed: 06-11-2017].

Scikit-learn (2018a). Feature selection. http://scikit-learn.org/stable/modules/feature_selection.html. [Date accessed: 06-4-2018].

Scikit-learn (2018b). Kernel Ridge Regression. http://scikit-learn.org/stable/modules/kernel_ridge.html#m2012. [Date accessed: 02-04-2018].

Sen, A. and Smith, T. (1995). *Gravity Models of Spatial Interaction Behavior*. Springer Berlin Heidelberg.

Stinson, M. A. and Bhat, C. R. (2004). Frequency of bicycle commuting: Internet-based survey analysis. *Transportation Research Record*, 1878:122 – 130.

The Royal Society (2017). Machine learning: the power and promise of computers that learn by example. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf. [Date accessed: 06-11-2017].

Tobias, R. (1999). An introduction to partial least squares regression.

Uyterlinde, M. and van der Velden, J. (2017). Kwetsbare wijken in beeld. *Platform31*.

Walpole, R., Myers, R., Myers, S., and Ye, K. (2013). *Probability and Statistics for Engineers and Scientists*. Pearson custom library. Pearson.

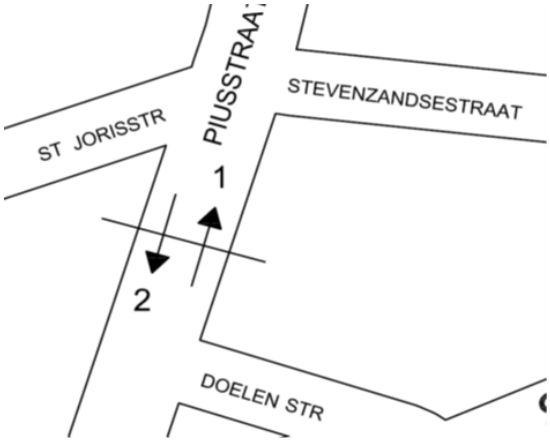
Zhu, S. and Levinson, D. (2015). Do people use the shortest path? an empirical test of wardrop's first principle. *PLOS ONE*, 10(8):1–18.

Appendix: Telpunt maps

A

Below, the official maps for each of the 24 count locations are shown, along with both directions for which intensities were measured.

Telpunt 3:

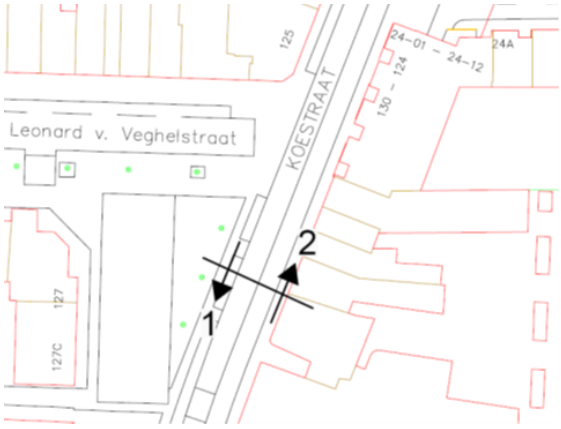


Telpunt 5:

Telpunt 9:

Telpunt 14:

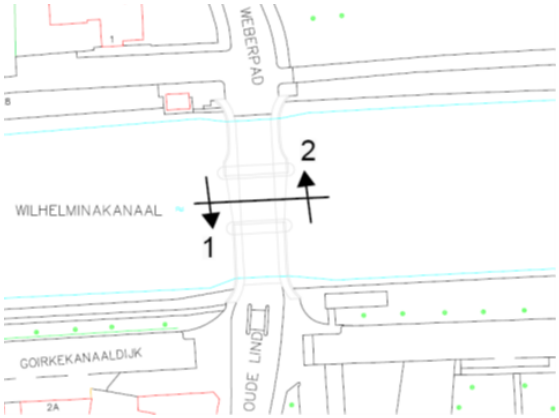
Telpunt 26:



Telpunt 30:

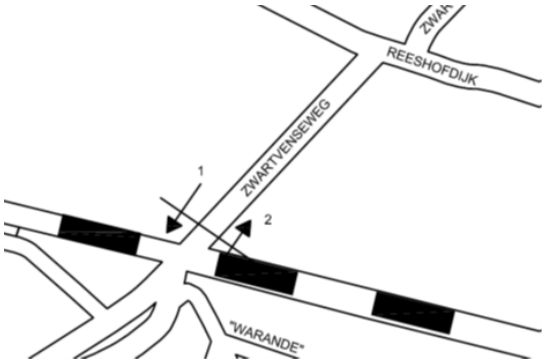
Telpunt 40:

Telpunt 51:



Telpunt 56:

Telpunt 61:

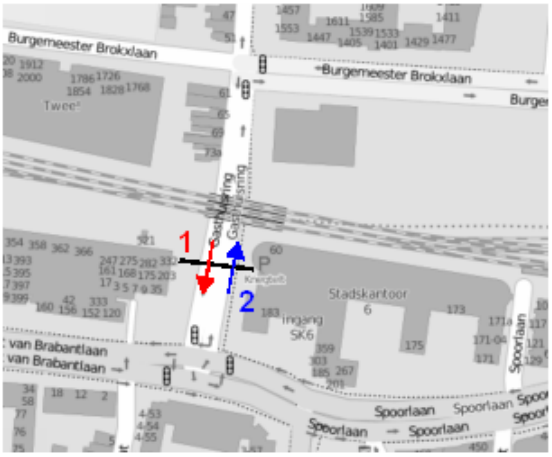


Telpunt 62:

Telpunt 63:

Telpunt 64:

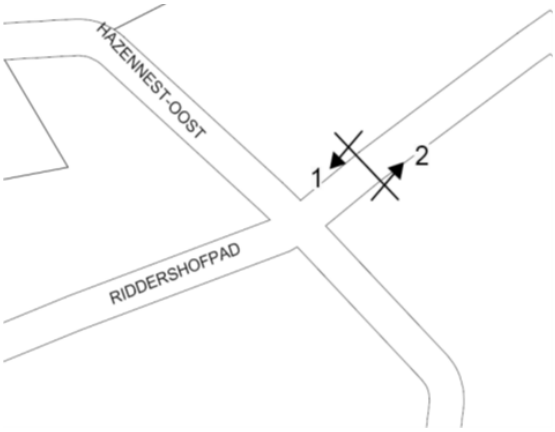
Telpunt 65:



Telpunt 66:

Telpunt 67:

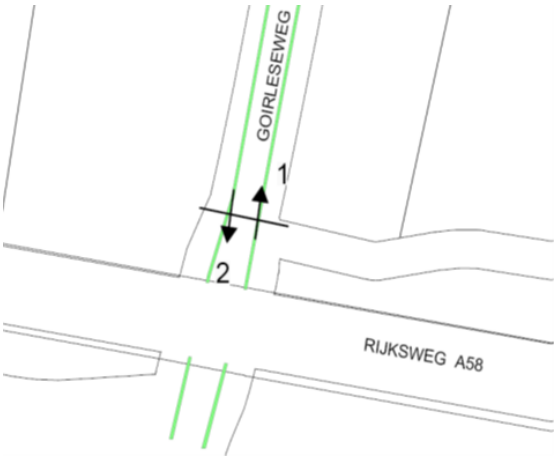
Telpunt 77:



Telpunt 78:

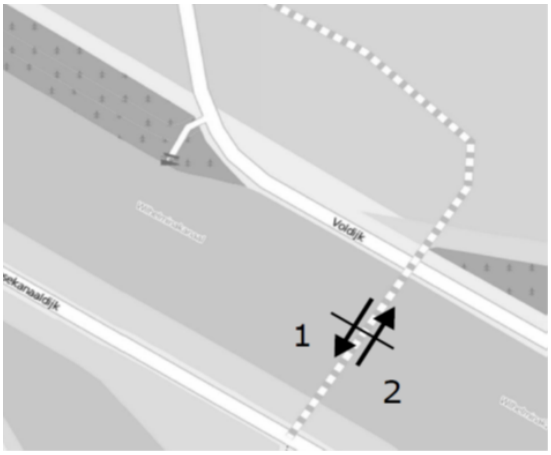
Telpunt 82:

Telpunt 83:

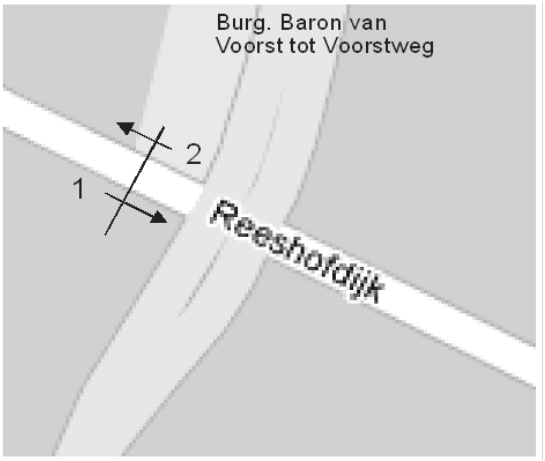


Telpunt 104:

Telpunt 113:



Telpunt 118:



Telpunt 121: