March 29, 2023

BACHELOR THESIS

Effect of the road safety on the cyclists route choice: Do cyclists consider road safety when deciding on their route?

UNIVERSITY OF TWENTE.

Witteveen + Bos

YAKUTA, R. (RYHOR, STUDENT B-CE)

University supervisor: Baran Ulak

External supervisor: Sander Veenstra

Preface

It is my great pleasure to present this Bachelor thesis on the topic of "Effect of the road safety on the cyclists route choice". The study was conducted as the final component of my Bachelor's degree in Civil Engineering at the University of Twente during the academic year 2022-2023. This project was executed externally under the guidance of "Witteveen+Bos" and the University of Twente. During this project I had an opportunity to work in a great company environment and have gained invaluable experience in road network data analysis.

Firstly, I would like to thank Witteveen+Bos for providing me with the opportunity to work on this project. Working in a Dutch-speaking company for an international student felt challenging at the beginning, however a friendly and responsive team provided me with a costless internship experience.

I would also like to express my gratitude to my external supervisor Sander Veenstra, who has provided me with his extensive knowledge and expertise in the field of working with spatial data and Python programming. His feedback, support and guidance during my project have helped me to develop my skills and knowledge in these fields.

Furthermore, I am extremely grateful to my internal supervisor Baran Ulak for his guidance during my thesis project, his feedback and for sharing his knowledge regarding regression modelling with me.

At the beginning of this project, I had limited prior knowledge of regression models and Python programming. However, with the help of my supervisors, I was able to acquire these skills and apply them to my research. This internship has not only allowed me to gain valuable experience but has also helped me to enhance my analytical, planning and problem-solving skills as well as gave me a great opportunity to work in a Dutch company environment.

I hope that this thesis will be interesting for the reader and it will contribute to the future researches in the field of road safety effect on the cyclist's route choice.

Ryhor Yakuta

Enschede, March 2022

Summary

Netherlands has the highest number of bicycles per capita in Europe, and the government and municipalities are constantly working on the improvement of the cycling network. Municipalities use traffic simulation models to improve cycling networks, including the "Fiets Monitor" developed by Witteveen+Bos company. However, the model's "all-or-nothing" approach for trip assignment may not reflect cyclists' preferences accurately, and Witteveen+Bos is working on its improvement. The research will be focused on the city of Enschede and its cycling infrastructure. The research aims to examine the influence of the road safety on the cyclists route choice, taking into account the impact of environmental and external factors. The study will involve examining existing literature application, required data collection and utilising regression modelling to determine whether road safety should be regarded as a critical factor in route attractiveness. The research gap outlined within this study is that while the influence of environmental factors on cyclists' route choice has been researched, the safety aspect of route choice was not considered and requires further investigation.

The safety of cyclists on the road is affected by various environmental factors such as road design, weather conditions, and traffic volume. Cycling safety can be measured by both objective safety (based on empirical data) and subjective safety (based on an individual's perception of safety). In this study, the objective safety was considered and road safety was defined as the probability of a crash occurring on a particular segment of the route, Negative binomial regression modelling was used to estimate the predicted number of crashes for each individual road segment.

Several environmental factors that can affect the risk of cycling crashes, including speed limits, road function class, traffic volumes, pavement type, intersections, and safety islands were included in the road safety analysis. In this study, various data sources were used, including the "Fietselweek data" for cycling counts, the "BRON" database for road accidents involving cyclists, "OpenStreetMap" for infrastructural information, "Basisregistratie Grootschalige Topografie" for greenery distribution, "Wijken en Buurten" for information on neighborhoods, and "RUDIFUN1" for data on land use classes. The limitations and benefits of each data source were discussed, with a focus on accuracy and completeness.

Negative binomial regression modelling was utilised to develop a safety performance function that can estimate the expected number of crashes. Meanwhile, logistic regression modelling was used to quantitatively analyse the influence of road safety on the cyclists' route choice. Python tools and packages were applied to execute all the modelling.

To build a negative binomial regression model, required data was gathered and processed. The list of factors that influence road safety was outlined and factors were properly categorised. The categorization was based on previous studies and the most popular road segments associated with the risk factors identified. The first step of the negative binomial regression model was to process the data as dummy variables, followed by selecting an appropriate distribution for the model. The dispersion parameter was used to estimate over-dispersion in the data. The Empirical Bayes method was fyrther applied to improve the precision of estimates by incorporating prior information and correcting for the regression-to-mean bias. This study found that pavement type, road function classes, and the location of road intersections have a significant impact on cyclist safety. The road safety map showed that major road links and the ring road surrounding the city of Enschede have the highest predicted number of crashes.

After the road safety was estimated a logistic regression model to analyse the influence of road safety on the overall network attractiveness was used. The segment approach and a logistic regression model were used to project the road safety effect on the route attractiveness. After, the route attractiveness maps of the network of Enschede were tested with and without consideration of the predicted number of crashes. Validation and verification procedures were executed to ensure the quality and accuracy of the results. Statistical tests such as Kolmogorov-Smirnov and Chi-square tests were applied to see if the road safety implementation has any significant influence on the road attractiveness. Furthermore, model key

performance indicators were checked to indicate if the road safety influenced the accuracy of the road attractiveness model. The impact of road safety on route segments' attractiveness was examined, and it was found that road safety had a minor effect on the overall network attractiveness. Statistical tests showed a difference between the datasets, but the difference was not significant enough to to considered. Overall, the addition of road safety data had a small effect on the model's predictions compared to other environmental factors.

This study investigated the relationship between road design factors and cyclist safety in Enschede using statistical models. Results showed that certain factors significantly influenced cyclist safety on road segments, but used methodology showed that road safety had a minor influence on the attractiveness of road segments for cycling. The study concluded that according to the applied methodology, road safety does not significantly affect cyclists' route selection, but further research is necessary to investigate the relationship between road safety and route selection for cyclists using alternative methodologies or integrating additional factors.

The recommendations suggest using more accurate and detailed data for analysis, conducting further research on other factors that influence cyclist safety and route choice decisions, incorporating alternative approaches to interpreting road safety, and exploring alternative methods for incorporating road safety into route attractiveness models. Additionally, the recommendations suggest examining other factors that may impact cyclists' safety and using the findings of this study as a starting point for future research.

Table of Contents

1. Introduction

1.1. Project Context

Cycling has been a popular mode of transportation in the Netherlands for over a century, with the construction of bicycle paths beginning in the 1890s (Reid, 2022). By 1911, the Netherlands had the highest number of bicycles per capita in Europe, and today it can be considered the world leader in cycling, with over 25 percent of all trips being made by bicycle (Wendel-Vos, Berg, & Giesbers, 2022). This high level of cycling is not only due to practical considerations, but also reflects the cultural significance of cycling in Dutch society, with the lowest mean average age of independent cycling in the world (Cordovil, Merce, Branco, & Lopes, 2022). Consequently, cycling is a key element in the Dutch transportation system, and the government and municipalities are committed to improving the cycling network by increasing the number of cycling lanes and enhancing the overall cycling environment.

In order to meet these goals, various models are being developed and used by different municipalities to simulate different scenarios and identify areas for improvement in the cycling network. One such model is the “Fiets Monitor” developed by Witteveen+Bos company, which simulates bicycle traffic and represents it as a visualized flow (Veenstra, 2022). The model uses a four-step procedure, with trips first being generated in an origin destination matrix, followed by the implementation of the Dijkstra algorithm to set the shortest routes for all OD pairs. The final step assigns all trips to the shortest possible routes using an “all-or-nothing” approach, where the shortest path is assumed to be the most preferred option for cyclists (Nijen, 2022). However, this approach does not always reflect the reality of cyclists' preferences, and Witteveen+Bos is searching for the ways to improve the trip assignment algorithm.

Route choice is also an important factor in the cycling experience, with cyclists typically choosing the path that they perceive to be the most attractive, regardless of distance. Environmental factors, such as location within a residential land use zone or paved infrastructure, can influence the attractiveness of a specific route and often override the popularity of the shortest possible route (Nijen, 2022). Additionally, the design of the infrastructure plays a significant role in the safety of the paths, which in turn influences the attractiveness of the routes for cyclists (Song, Ni, & Li, 2017). Therefore, this study will be focused on the investigation of the safety effect on the cyclists' route choice.

1.2. Research Aim

The aim of this research is to examine the influence of road safety on the route choice of cyclists, taking into account the impact of environmental factors. While previous studies have analysed the influence of environmental factors on cyclists route choice, the safety aspect of route choice was not considered and requires further investigation. Therefore, the general assumption of this study is that the safety of cyclists has a decisive influence on cyclists perception of the attractiveness of a road. The current method of route assignment relies on the "all-or-nothing" principle, where the shortest possible path is assumed to be the most preferred option. However, this is not always the case in reality as cyclists have different perceptions and preferences regarding route choice.

Cyclists have various route choices available to them, and they may not always choose the shortest path. As a result, this study will analyse the impact of road safety on cyclists' route choices, while also considering environmental factors. The study will involve examining existing literature and utilizing regression modelling to determine whether safety should be regarded as a critical factor in route attractiveness analysis and how it affects the appeal of road segments.

1.3. Research Scope

This section is going to introduce the boundaries of this research. Study area and required databases for this study will be outlined in this section.

1.3.1. Study Area

In this study, the city of Enschede and its cycling infrastructure will be selected as the primary study area, as the coefficients for the environmental factors were derived through regression models specifically developed for Enschede. Therefore, using the existing regression model in the same setting as previously employed is the most appropriate approach to analyse the safety effect on cyclists' route choice. The geographical location of Enschede is shown in Figure 1.

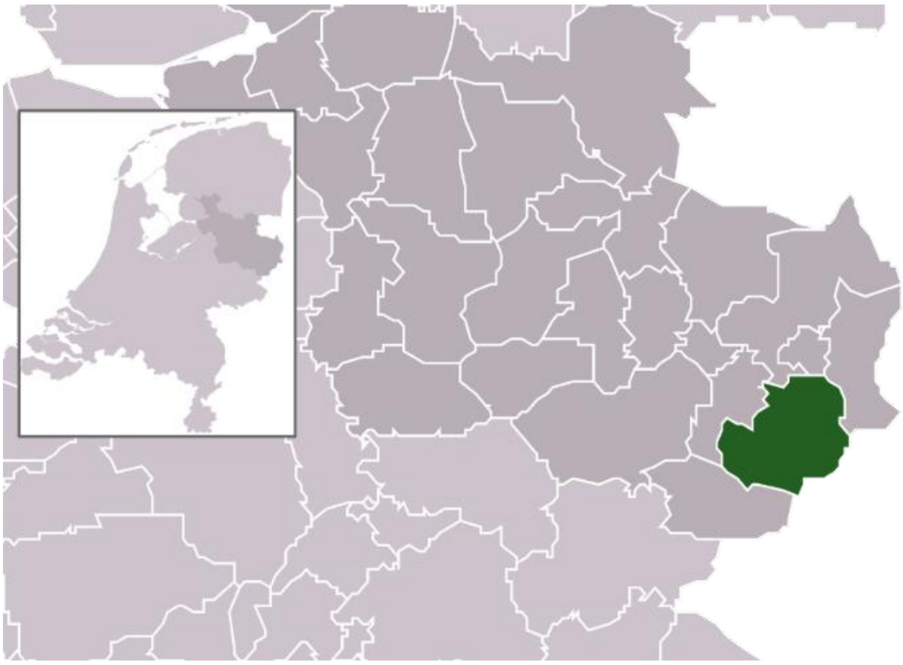


Figure 1:Enschede geographical location

1.3.2. Research Limitations

In order to develop a safety model for this study, it is necessary to consider multiple factors that contribute to the characteristics of road elements. While some safety models for cyclists have attempted to incorporate personal characteristics, the collection of data required for such modelling is often not available and can be complex to obtain (Elsayed, Sayed, & Brown, 2016). As a result, environmental characteristics of the roads will be used in this study to develop the safety model. Previous research has shown that environmental factors, such as speed limit, artificial lightning, surface pavement and infrastructure design, can significantly impact the safety of cyclists (Kim, Ulfarsson, & Porrello, 2021) .By incorporating these factors, the safety model developed in this study will provide valuable insights into the safety of cycling infrastructure in the study area.

1.4. Research Gap

Previously, the influence of the environmental factors on the cyclists route choice was researched and quantitatively analysed in the context of bachelor assignment in the Witteveen+Bos company. However, the safety influence of the road attractiveness was out of the scope of that study (Nijen, 2022). Moreover, majority of the available studies regarding the cyclists road safety research did not analyse its effect on the route attractiveness for the cyclists.

A generalised linear model was already used in order to predict traffic accidents, and certain conclusions were drawn from the analysis. However, the impact of the predicted safety was not projected on the attractiveness of the routes. (Reurings & Jassen, 2006).

Majority of studies that were investigating the safety effect, making the safety prediction models and stated that safety has a significant impact on the cyclists route choice. However, there is a lack of quantitative analysis of the safety effect on route choice in the literature. Despite the agreement among researchers that safety plays a crucial role in cyclists' choice of routes, few studies have attempted to quantify this effect.

1.5. Scientific and Societal Relevance

This study aims to investigate the relationship between road safety and road attractiveness for cycling. By analysing various environmental characteristics of road segments and modelling road safety, the study will identify the extent to which safety influences cyclists route choice. This research is scientifically relevant since it fills a gap in the literature, which is the lack of a quantitative analysis on the safety effect on route choice. Moreover, the results of this study can help to improve "Fiets Monitor" algorithm of the route assignment. In this way, this study can be considered as socially relevant due to the fact that its final outcomes can be theoretically used in order to improve cycling infrastructure in the Netherlands.

1.6. Research Questions

As it was previously stated, the aim of this research will be to quantitatively analyse the road safety effect on the route segments attractiveness. Therefore main question was formulated as following.

Main question:

"To what extent does the road safety influence route attractiveness for cyclists?"

In order to answer the main question, a number a sub questions were formulated that require to be answered.

Sub-questions:

1. *"How cyclists safety can be quantified?"*

Due to the fact that safety has a crucial impact on the cyclists behaviour, it is essential to consider it while trying to develop a procedure of a route assignment with regards to a external factors and not applying the "all-or-nothing" approach (Riggs, 2018). Therefore, first of all it needs to be quantified.

2. *"How can effects of cycling safety on the route selection can be modelled"*

In order to analyse the effect of the road safety on the route attractiveness it is required to add road safety results to the route selection model and analyse its impact.

3. *"Is there any statistical significant effect of the safety on route attractiveness?*

This will be the last sub question of this study. In order to give an objective assessment of the performed work, the end result of this study needs to be validated and verified. Developed methodology will be tested against the previously obtained results and a number of statistical tests for the data sets will be performed. This will help to understand if the method is reasonable and whether or not it gives reliable results.

2. Theoretical Framework

The following section is going to describe theory involved in this research. The modelling will be performed using Python and build up on the basis of already developed logistic regression model with its further improvement. Road safety interpretation, environment effecting cyclists safety and regression modelling will be discussed in this chapter.

2.1 Cyclists Safety

Cycling safety can be influenced by various environmental factors such as road design, weather conditions, and traffic volume. Poor road design, such as lack of bike lanes or poorly maintained road surfaces, can increase the risk of accidents for cyclists (Teschke, Harris, Reynolds, & Winters, 2012).Adverse weather conditions such as rain or snow can also reduce visibility and traction, making it more difficult for cyclists to control their bikes and avoid collisions (Elvik, 2017). Additionally, high traffic volume and speed limit can increase the risk of accidents for cyclists (Aultman-Hall & Kaltenecker, 1999). Therefore, cycling safety depends on a variety of environmental and personal factors the totality of which predetermines the level of safety of a particular cyclist on the road. There are two main ways different ways of looking at the concept of safety. Objective safety provides an empirical basis for making decisions and implementing safety measures, while subjective safety reflects the human experience of safety and informs how people interact with their environment.

Subjective safety refers to an individual's perception of how safe they feel while cycling. This perception is influenced by various factors such as the presence of dedicated cycling infrastructure, traffic volume, road conditions, and weather (Heesch, Garrard, & Sahlqvist, 2012). It is a subjective measure that reflects an individual's confidence in their ability to cycle safely in a particular environment (Moudon, Lee, Cheadle, & Collier, 2005). Moreover, some research papers state that the way of the road safety perception depends on the gender or age group of cyclists (Misra & Watkins, 2018). Since it is complicated to implement the subjective safety as it was stated in the Limitations chapter and due to the fact that every person has his or her own perception regarding the route safety, the objective safety will be implemented using the data about traffic accidents on the roads where the cyclist were involved. Therefore, the officially registered number of accidents that happened on the road segments will be formulating the segment safety.

Objective safety refers to the actual risk of injury or death while cycling in a particular environment (Lusk, Furth, & Morency, 2011). This measure is based on empirical data, such as the number of accidents, injuries, and fatalities per unit of cycling distance or time. Objective safety takes into account the physical characteristics of the environment, such as the quality and design of cycling infrastructure, traffic volume, speed limits and some other factors that will be discussed in the next chapter (Teschke, Harris, Reynolds, & Winters, 2012).

The level of 'objective risk' can be a crucial factor in the route selection for cyclists. One way of measuring this objective risk is through the probability of being involved in a crash on a certain route. In a study by von Stülpnagel and Lucas (2020), the researchers analysed the crash risk and subjective risk perception of urban cyclists. They found that there was a significant correlation between the actual crash risk and the cyclists' perception of risk. This suggests that the objective risk can be a reliable measure for assessing the safety of cycling routes. By incorporating the objective risk in the route selection process, cyclists can make informed decisions based on the level of risk associated with each route (von Stülpnagel, 2020).

Hence, the road safety in this study will be defined as the probability of a crash occurring on a particular segment of the route. By utilising negative binomial regression modelling and taking into consideration the factors associated with the adjacent road sections, the projected number of crashes for each individual road segment will be estimated.

2.2. Factors Increasing Crash Risk

This chapter will be dedicated to the environmental factors that were considered as significant in terms of the effect on the cycling safety.

Speed limits, road function class, motorised vehicles or cyclists volumes, surface pavement type, presence of intersections, and presence of safety islands were considered to be important environmental factors that can influence the risk of cycling crashes. Higher speed limits and higher traffic volumes are associated with an increased risk of cycling crashes (Teschke, Harris, Reynolds, & Winters, 2012).

The study conducted by van Petegem and Wegman (2014) also examined the impact of traffic flow on the risk of run-off-road crashes. The findings indicated that high traffic flow rates were associated with an increased risk of such crashes. The authors suggested that this may be due to a number of factors, including reduced space for cyclists to manoeuvre, increased difficulty in detecting hazards, and decreased visibility. Therefore, the traffic volume is considered to be a valuable factor effecting the crash risk within the scope of this study. Cycling volume and motorised vehicle volumes will be considered separately in the analysis.

Segment length is an important factor to consider when modelling a crash prediction model per route segment, as it can significantly affect the estimation of crash risk. The length of a road segment can influence the number of crashes that occur on that segment, with longer segments generally having a higher number of crashes than shorter segments. Additionally, the risk factors that contribute to crashes may vary depending on the length of the segment, with different risk factors being more or less important on longer or shorter segments. For example, on a short road segment, a sharp curve may be a more significant risk factor for crashes, while on a longer segment, high traffic volume or poor road surface conditions may be more important risk factors.

Another important factor that effect the crash risk is the road pavement type. Study conducted regarding surfaces pavements showed that pavement conditions can significantly influence the risk of crashes (Dozza & Werneke, 2014). In particular, smooth pavement surfaces were associated with a lower risk of crashes compared to rough pavement surfaces.

Next factor that will be used in a binomial regression model construction is the speed limit. Speed limit is a crucial factor in determining the potential severity of a crash and the likelihood of fatalities (Ribeiro, 2014). Lower speed limits have been shown to reduce the frequency and severity of crashes, particularly in areas with a high volume of vulnerable road users, such as pedestrians and cyclists.

Regarding the road function class, the related research was conducted in 2014 in Brussel. In the study by Vandenbulcke, Thomas, and Int Panis (2014), road function was found to have a significant effect on cycling crash risk in Brussels. Specifically, the researchers found that cycling on roads with a primary road function was associated with a higher crash risk compared to roads with a secondary or local function. This could be due to several factors, such as higher traffic volumes, higher vehicle speeds, and less cycling infrastructure on primary roads. Additionally, the study found that cycling on roads with public transport services was also associated with a higher crash risk, likely due to the presence of buses and trams sharing the road with cyclists (Vandenbulcke, Thomas, & Int Panis, 2014).

Safety crossing islands, which are designed to provide a safe crossing point for cyclists and pedestrians were also considered as a decisive environmental factors effecting the cycling crash risk. The results of the study aimed on the investigation of the road facilities on the road safety showed that the presence of a safety crossing island had a significant effect on reducing the risk of crashes with motor vehicles (Schepers & Stipdonk, 2017). Specifically, the odds of a bicycle crash with a motor vehicle were 74% lower at intersections with safety crossing islands compared to those without. Figure 2 illustrates how the safety island looks on the road segment.

Figure 2: Cyclists safety island

In the same research, authors found that two-way streets had a significantly higher incidence of bicycle crashes than one-way streets, particularly in urban areas. This could be due to the increased complexity and unpredictability of interactions between bicycles and motor vehicles on two-way streets. Additionally, the study found that crashes at intersections accounted for a higher proportion of fatal bicycle crashes on two-way streets compared to one-way streets. Therefore, traffic movement and intersections will be included as a factors in the regression analysis.

In this way, the most important factors that potentially increase the cycling crash risk were listed and their effect was assumed based on the reviewed articles. Table 1 represents discussed factors.

Factor	Assumed effect on the crash risk
Cycling volume	higher cycling volume increases crash risk
Motorised vehicle volume	higher motorised vehicle volume near the cyclists or on the shared roads increases crash risk
Segment length	the longer the route segment the higher the crash risk
Pavement type	Smooth pavement types reduce crash risk
Speed limit	higher speed limits increase the crash risk
Road function class	residential and cycling roads refer to lower crash risk
Safety island	road segments near the safety islands are less vulnerable for the crash risk

Traffic movement(one way/two way)	two way traffic movement road segments are refer to a higher crash risk
Intersection	road segments near the intersections refer to a higher crash risk
Artificial lighting	Artificial lighting decreases crash risk

Table 1: Factors influencing the cycling crash risk overview

2.3. Regression Modelling

The use of regression models in analysing and interpreting data has become increasingly important across various fields, including public health and transportation. Logistic regression modelling is commonly used in predicting a binary outcome, such as the occurrence of a crash, while negative binomial regression modelling is often used in predicting count data, such as the number of crashes in a specific area or time period. These models have been applied in various transportation safety studies, including those involving cycling and pedestrian safety. In this chapter, these models will be discussed.

2.3.1. Negative Binomial Regression Modelling

The use of statistical models in traffic accident modelling has become an increasingly popular tool in research. One such model that has been applied in traffic accident modelling is the Generalized Linear Model (GLM), which includes the negative binomial regression (NBR) method (Wei & Lovegrove, 2013). GLM is a flexible statistical model that can handle different types of response variables, including count data. Negative binomial regression is a type of GLM that is used to model count data. Negative Binomial Regression can model the count data with greater accuracy and is therefore more suitable for modelling traffic accidents, where count data such as the number of accidents is required by the research aim.

Crash prediction models, also known as safety performance functions, can be useful for the purpose of evaluating the safety performance of a network by estimating the expected crash risk. The relative safety of a road segment in a network can be determined by using the NBR model in relation to traffic flow, road length, and risk factors such as design elements (Petegem & Wegman, 2014).

From the other side, linear regression models are usually used to establish relationships between dependent and independent variables. However, a conventional linear regression model may not be appropriate for crash prediction modelling due to the requirement that the dependent response variable must be normally distributed, which is inaccurate for crash data that are discrete and non-negative. Additionally, some factors may strongly correlate with each other, leading to multicollinearity (Arhin & Gatiba, 2020). Therefore, it was considered to use generalised linear model, especially negative binomial regression model within the scope of this study in order to estimate road safety.

This research considers using Python tools and packages in order to perform negative binomial regression analysis. Therefore, all the modelling will be executed with the help of specialised Python packages.

Several studies that were focused on the safety model creation developed the following formula in order to analyse risk factors and based on them make safety model estimations (Petegem & Wegman, 2014):

Equation 1: Expected number of crashes formula

$$E(Y) = \alpha Q^{\beta} L e^{\Sigma \delta j x j}$$

Where:

(Y) = expected number of crashes,

α = constant,

Q = traffic flow,

L = road segment length,

xj = risk factors affecting the crash risk,

β, γ, δj = model parameters of the different model variables.

As crash data are based on counts, the Poisson distribution is the basic distribution (Van Petegem & Wegman, 2014). However, the assumption that the distribution of variance is equal to the mean is often violated in crash prediction models, as variance is often significantly larger than the mean, leading to over-dispersion (AmohGyimah & Saberi, 2016). Negative binomial models are commonly used in bicycle crash frequency research, as they can handle data over-dispersion (AmohGyimah & Saberi, 2016). Negative binomial models are preferred over Poisson distributed models for this reason. Therefore, safety model will be constructed using GLM tools in Python, specifically negative binomial regression.

2.3.2. Logistic Regression Modelling

In order to achieve the outlined goals of this study, it is necessary to quantitatively analyse the road safety influence on the cyclists route choice. For this purpose, the previously developed model that analyses environmental factors effect on the road segment attractiveness will be used. Furthermore, this model will be improved in order to consider road safety since it was not previously accounted. This model is based on the logistic regression, therefore it is important to shortly discuss the reasons of logistic regression being used and theory behind it.

For the problems where it is required to analyse the behaviour of people during making a choice discrete choice models are considered as one of the best available options (Khandelwal, 2020). Although there is a variety of discrete choice models, logistic regression model is applied in this study due to its simplicity, and practical applicability. Discrete choice models are useful for analysing the behaviour of individuals when they have to choose between different alternatives (Huber, Lindemann, & Muthmann, 2022). Attractiveness of the options can be represented by utility, which is defined as what an individual tries to maximise (de Dios Ortúzar & & Willumsen, 2011). The goal of the logistic regression modelling in this case is to find a relationship between the attractiveness of a certain route and the infrastructural and land use allocation characteristics of the routes. Logistic regression modelling is used because the dependent variable is binary, and the coefficients of each factor can be estimated through regression analysis.

Logistic regression takes the natural logarithm of the probability as a regression function of the independent variables. Equation 2 represents the formula of a logistic regression.

Equation 2: Logistic regression formula representation

$$Y_i = logit(P_i) = Ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

Furthermore, some preliminary measures need to be taken regarding the data that will be used. Outliers should not be present, and multicollinearity should be absent among the independent variables. Due to the fact, that model is already calibrated, adding a road safety in it will require outliers elimination and multicollinearity check of a road safety versus previously used environmental factors. In order to exclude multicollinearity, the Pearson correlation test is performed. Since the logistic regression model will be used for the basis of the modelling, data for the modelling has to be binary. Therefore, the roads observed with cycle counts will be considered as value one and the shortest paths will be assigned to a zero value. However, the output can not be binary, and it will be in a form of probability that results will take value of one. Equation 3 represents the formula of how the probability will be calculated. In this formula, β stands for the coefficient of a certain factor and x stands for the prediction factor of a certain factor (LaValley, 2008).

Equation 3: Probability of choosing a route

$$P_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}}}{1 + e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\cdots+\beta_k x_{k,i}}}$$

Talking about logistic regression model and environmental factors that were considered with their coefficients that are going to be included in this study, 13 factors presented on the Table 2 and previously considered as the factors that have the most impact.

Factor	Standardised β
Constant	0
Infrastructural	
Distance to traffic control installation	-0.124
Cycle lane	0.095
Separate cycle path	1.072
Artificial lighting	-0.246
Paved infrastructure (asphalt)	0.699
Motorised vehicle intensities	-0.675
Bicycle intensities	1.108
Land use allocation	
Residential land use zone	0.689
Commercial land use zone	-0.446
Greenery land use zone	-0.227
Industrial land use zone	-0.232
Land use mix	0.034
Degree of urbanisation	0.365

Table 2:Previosly researched factors with standardised regression coefficients

Furthermore, the conceptual framework of the logistic regression model that was previously developed, tested and validated is present on the Figure 3 (Nijen, 2022). Some of its sections will be discussed in 3.3.1. Logistic Regression Modelling.

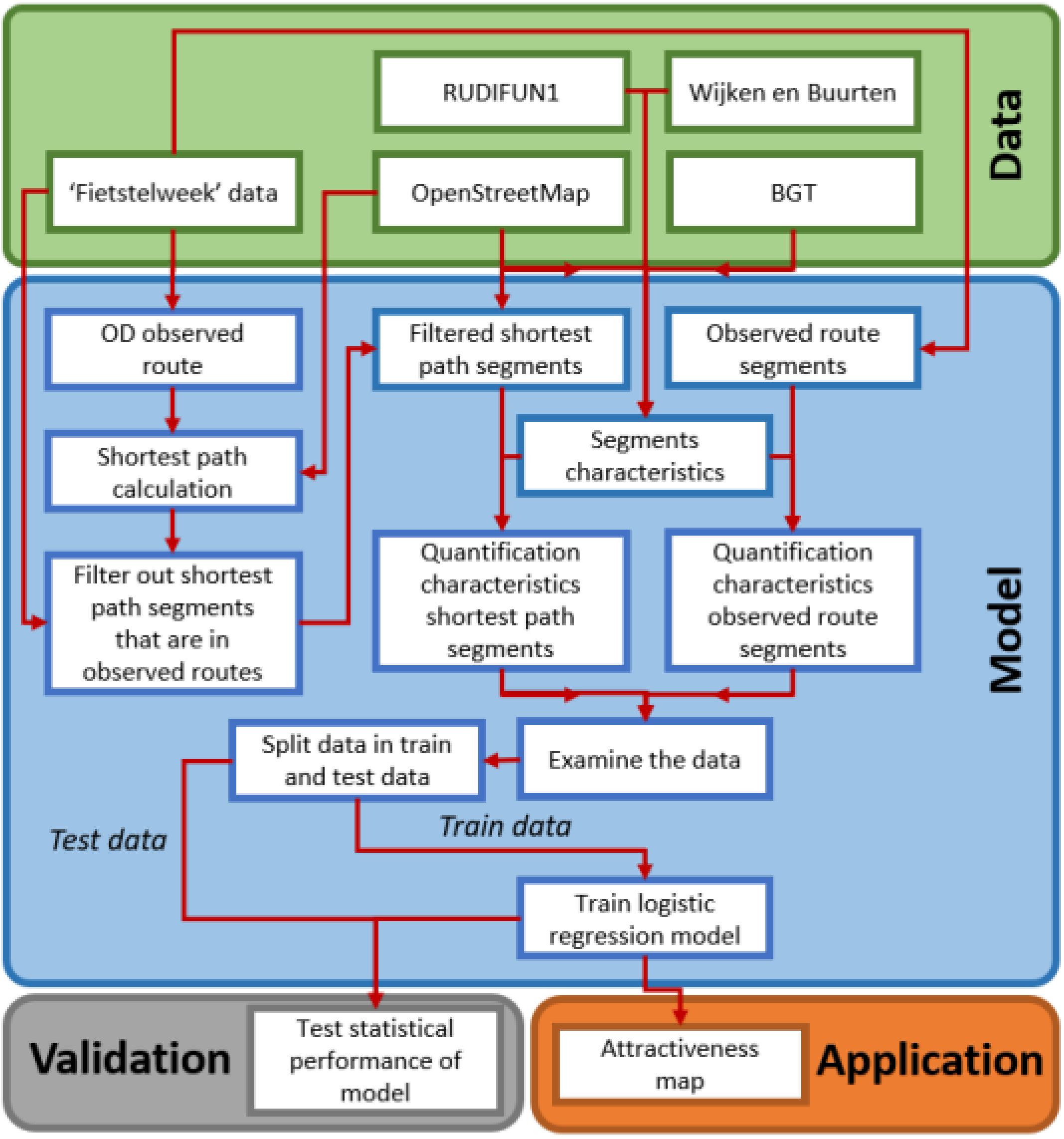


Figure 3: Conceptual framework of the logistic regression model (Nijen, 2022)

3. Methodology

The methodology section of a research paper is essential in outlining the strategies and techniques that were used in the study in order to obtain the results. This section will provide a clear and detailed explanation of the data collection methods, data analysis techniques that will be used in this study or were previously used in related studies. The purpose of this section is to provide sufficient information on how the study is going to be conducted. The methodology flowchart of this study is illustrated in Figure 4.

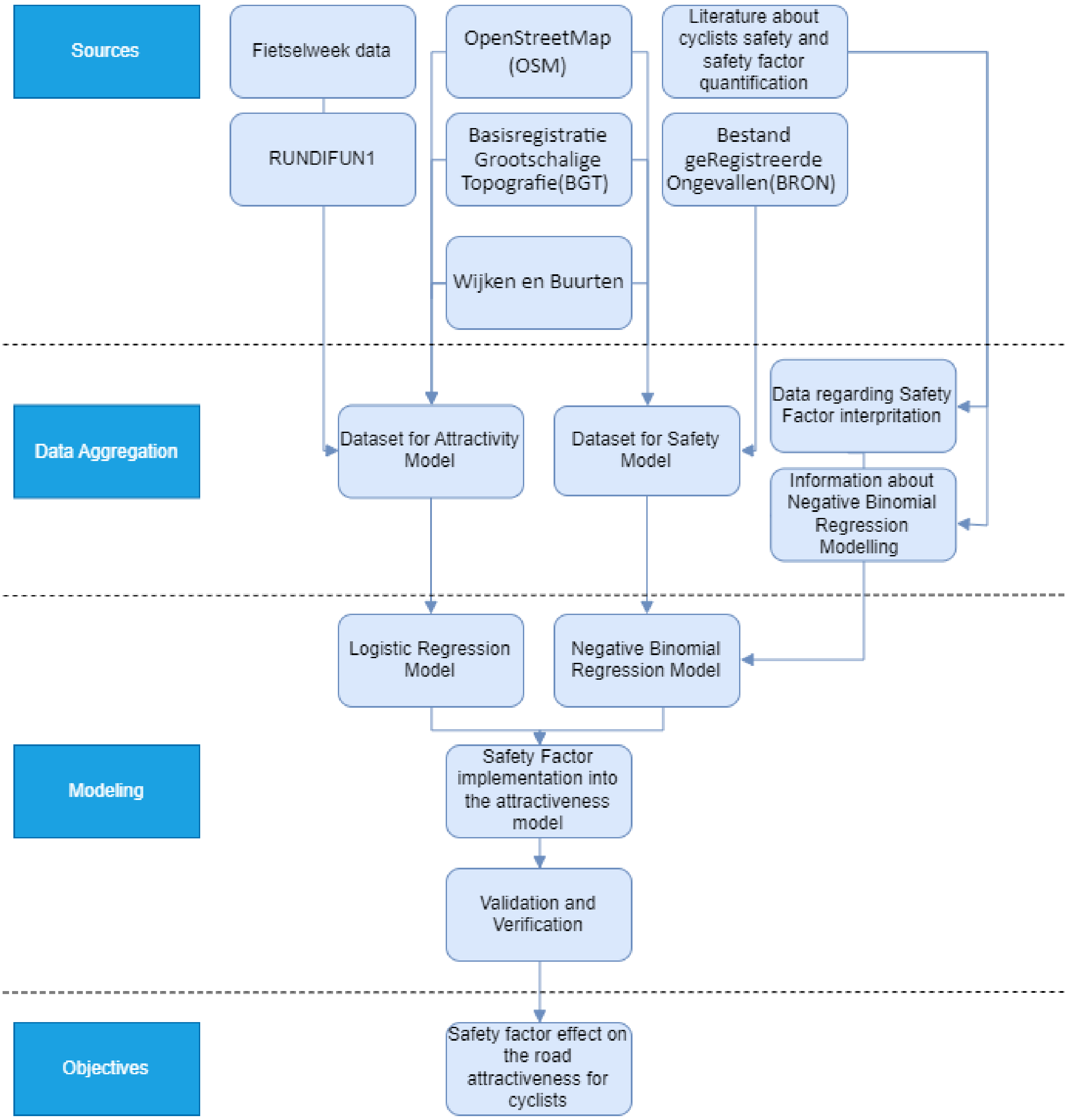


Figure 4: Methodology Flowchart

3.1. Data Collection

The success of any research project relies heavily on the quality and accuracy of the data collected. In the present study, data were collected to investigate the relationship between risk factors and road safety. The data collection process involved the use of various sources, including official reports from Dutch government. The following section describes the data collection methods and its further categorisation.

3.1.1. Available Databases

The scope of this research involves modelling road safety based on the environmental characteristics of the road segments. The use of various data sources in this study is crucial for generating accurate and reliable results. Therefore, different open datasets that will be used within the scope of this study are going to be discussed. Table 3 Illustrates the overview of the databases that are going to be used in this study.

Dataset name	Provided Data	Data Required For The Study	Source
OpenStreetMap	Detailed infrastructural information	Route characteristics clarification	(OpenStreetMap, 2022)
Fietselweek data	Cycling counts collected in September 2016, with data on origins and destinations of cyclists	Cycling counts	(Breda University of Applied Sciences, n.d.)
(BRON)	Information on road accidents: Location, Facts, Victim characteristics (gender, age, etc.), Vehicle information, Data subjects (information about directors)	Accidents involving cyclists for modelling road safety	(Rijkswaterstaat, 2022)
BGT (Basisregistratie Grootschalige Topografie)	Detailed information on all physical objects in the built environment	Greenery distribution, safety islands location	(PDOK, 2022)
RUDIFUN1	Information on land use of buildings	Land use coefficients	(RUDIFUN1, 2022)
Wijken en Buurten	Information on all neighborhoods, districts, and municipalities in the Netherlands	Boundaries of the study area, degree of urbanisation	(Centraal Bureau voor de Statistiek., 2022)

Table 3: Available databases overview

In this project, cycling counts will be obtained from the “Fietselweek data” database. This database provides access to the cycling counts that were collected during September 2016, for the period of two weeks data about origins and destinations of the cyclists was collected (Breda University of Applied Sciences, n.d.). This data was collected through the mobile app by using GPS, therefore it has some accuracy limitations (Koch & Dugundji, 2021). For example, the accuracy of GPS may not always be precise enough to determine the exact location of a cyclist, particularly when cycling on a parallel cycle path. Furthermore, the dataset is anonymised, and information on demographics, such as age and gender, is not available, even though studies suggest that these factors influence cycling behaviour and route selection. Another limitation is that the participants in the fietselweek survey were mainly experienced cyclists, which may have resulted in longer trips being overrepresented in the data. This may affect the results, as the average distance per trip in the fietselweek data was higher than that reported by the Centraal Bureau voor de Statistiek in 2017. Nevertheless, despite these limitations, the fietselweek dataset remains a valuable resource for this research.

In order to get data about road accidents Bestand geRegistreerde Ongevallen Nederland (BRON) will be used. This data set collects all the information about accidents that were registered by police (Rijkswaterstaat, 2022). In the context of this study, only accidents where cyclists were involved will be considered. The BRON database, which records all road crashes in the Netherlands, is incomplete due to underreporting, especially for less severe crashes. However, despite its limitations, the dataset still provides valuable insights into relative number of crashes at a given location.

The next data source utilised is 'OpenStreetMap' data, which is a detailed and continually updated open-source project containing infrastructural information of the entire world (OpenStreetMap, 2022). While the benefits of this data are vast, it should be noted that its open-source nature means that anyone can contribute to it, resulting in some misidentified elements. Despite this disadvantage, the level of detail provided makes it a valuable resource for this study.

'Basisregistratie Grootschalige Topografie' data, is managed by the Dutch government and provides detailed information on all physical objects in the built environment (PDOK, 2022). The data is updated regularly, and changes are validated by the government. In this study, this dataset will be mainly used in order to gather information about greenery distribution.

The 'Wijken en Buurten' data source, which is a part of the 'Centraal Bureau voor de Statistiek,' is another valuable data source utilized in this study (Centraal Bureau voor de Statistiek., 2022). The data contains information on all neighborhoods, districts, and municipalities in the Netherlands and is obligated by Dutch law to provide relevant statistical data. Therefore, this data will provide information regarding boundaries of the study area.

Finally, the 'RUDIFUN1' data source, managed by the 'Planbureau voor de Leefomgeving,' contains data on the built environment and specifically provides information on which land use the buildings fall under (RUDIFUN1, 2022).

3.1.2. Crashes Data Collection

As it was previously discussed in chapter 3.1.1. Available Databases, information regarding cycle crashes will be retrieved from Bestand geRegistreerde Ongevallen Nederland (BRON). However, first it needs to be filtered since it contains records about all traffic crashes that were registered all over the Netherlands.

Therefore, only crashes where cyclists were involved, only in the Enschede regions were filtered out. To have a better representation about cycle crashes distribution, the heat map was constructed.

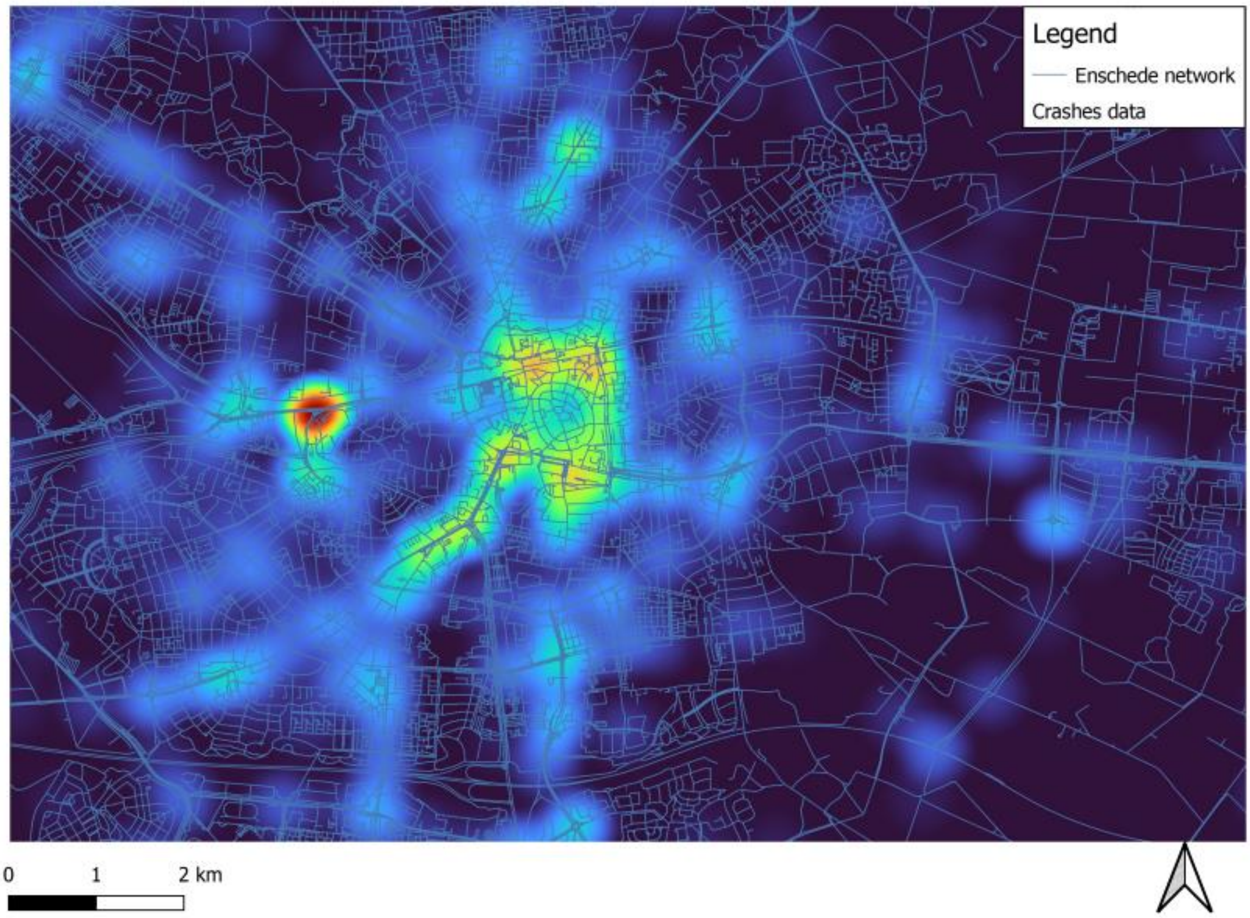


Figure 5: Heatmap of the cycle crashes in the Enschede

Figure 5 depicts a conspicuous area that deviates from the overall trend. Subsequently, it was discovered that this discrepancy was attributed to the auto-allocation of crash records prior to 2014 to the municipal geographical centre when no specific location was specified. Consequently, data obtained post-2014 was considered more reliable, and therefore, the former data was excluded from analysis.

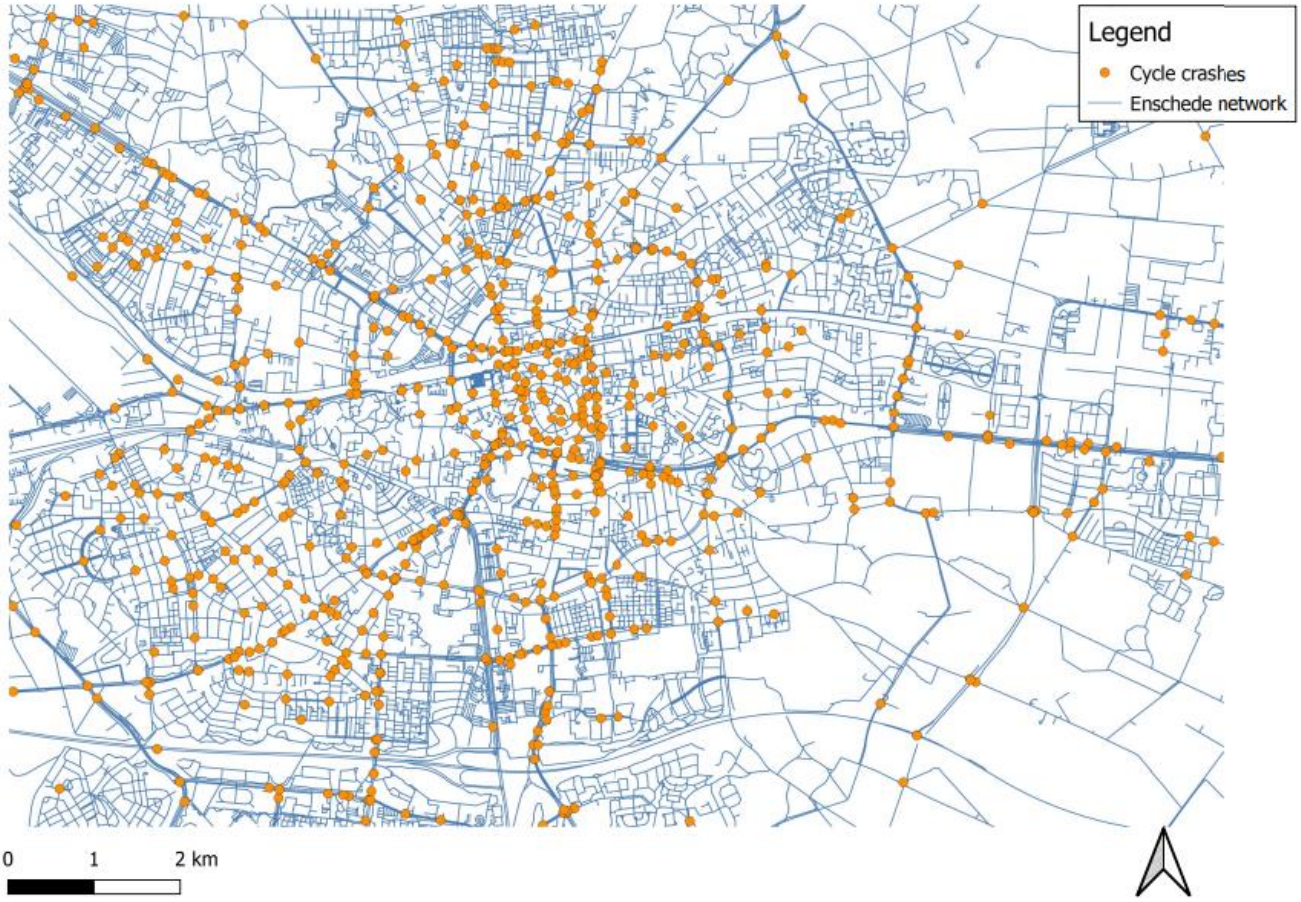


Figure 6:Crashes with cyclists being involved in Enschede

Figure 6 represents the filtered cyclists crash data and its distribution over the Enschede network.

3.1.3. Risk Factors Collection and Categorisation

In order to build negative binomial regression model, it is important to gather required data and reasonably process it. As it was illustrated in the Table 1, the list of factors that influence the road safety was outlined. Therefore, it is necessary to categorise this data in order to reduce the number of degrees of freedom for the model. However, some data will be used as a numerical data without any further categorisation.

Traffic volumes will be categorised while road segment length will be used as a numerical values for the regression analysis. Nevertheless, other factors require categorisation. Road function class factor will have 6 categories namely: residential, primary, secondary, cycleway, tertiary and unclassified function classes. Therefore, every road segment is going have its unique function class. Roads where the function class was not indicated or the road function class was not on the list of this 6 outlined categories will be categorised as unclassified. This will be done due to the fact that in the Enschede region exists around 25 road function classes that are represented on the Figure 7 and all of them could not be considered.

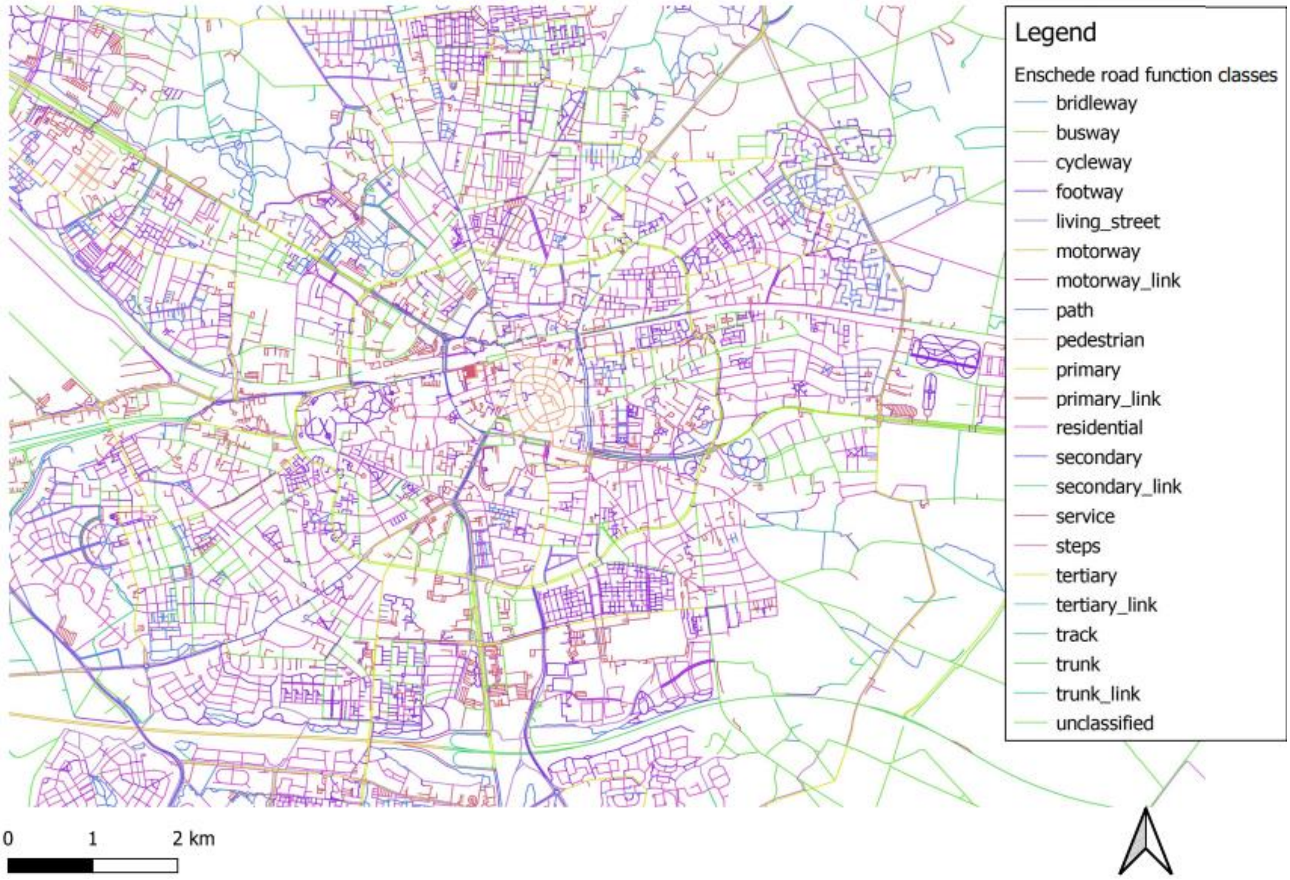


Figure 7: Road function class diversity in the Enschede

This six classes were selected based on the fact, that the highest number of crashes per 100 metres occur on the road segment associated with these classes, which is illustrated on the Figure 8. Therefore, these six function classes can be considered as a risk factors and included in the further analysis. Roads that fall under the unclassified functional class are typically shared by various road users and do not have a clear designated function. All the remaining function classes, excluding the six most popular ones identified in the graph, will not be considered in the analysis due to their varied functional purposes and comparatively low number of recorded accidents which makes them unsignificant risk factors for the analysis. As a result, the occurrence of cycle crash accidents on these road segments may be perceived as relatively sporadic and unpredictable, which makes it not relevant to further consider their function class characteristic in the analysis.

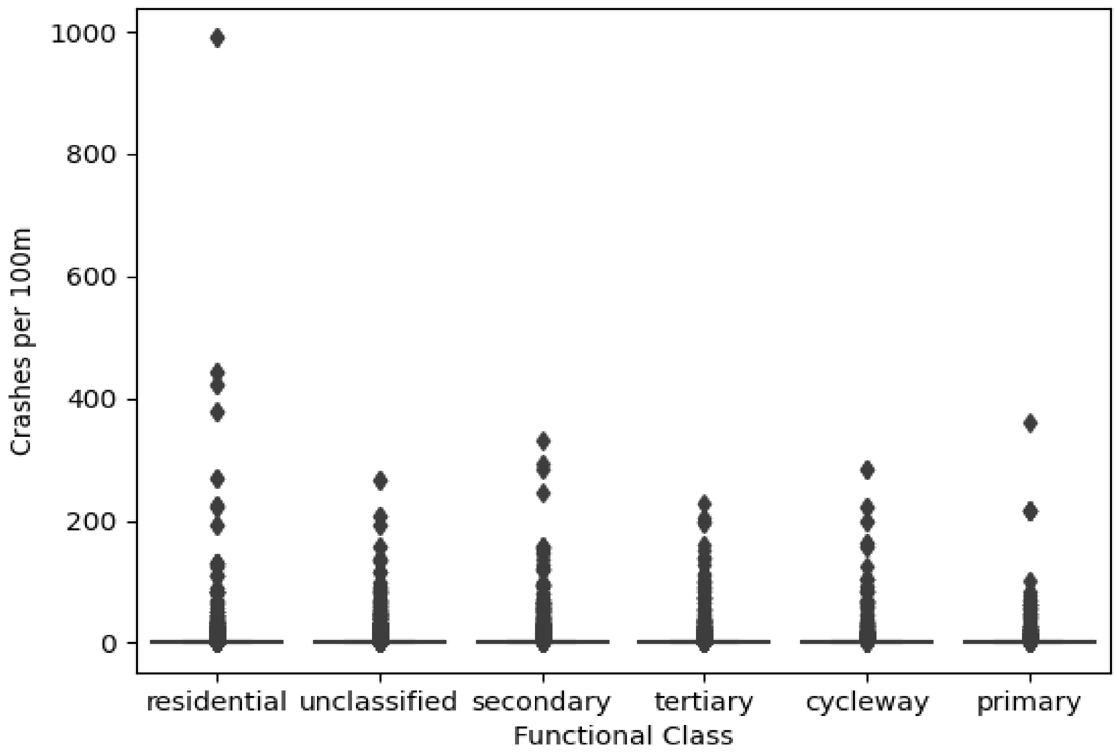


Figure 8: Boxplot(Function classes with the highest number of crashes per 100 metres)

According to the previous studies, three surface pavement types were selected. Asphalt, paving stones and gravel will be included in the analysis since these three types of pavement are the most popular in the study area. Intersections and safety islands will be categorised as present or not present on a certain road segment. Speed limits will be grouped in three categorised since there are around 10 different speed limitations all over the Enschede network. Road segments where speed limit was not indicated will be assigned to a 0 – 15 km/h speed limit group due to the fact that this type of issues were common for residential zones were speed limitation is not usually stated. Traffic direction will be categorised as one way or two way, therefore two categories will be created. In this study, cycling volume and motorised vehicle volume will be categorised in four groups, different scaling will be applied since motorised vehicle volume and cycling volume have different value ranges. Route segment length will be considered as a numerical values and no categorisation will be performed for these values. However, route segment length will be considered as an offset variable. That means that road segment length will be used in regression with a fixed coefficient of one. In other words, increase of this variable will correspond to an increase in the predicted number of cycle crashes on the road segment. This is logical since longer road segments have a statistically higher chance of an accident to happen. Figure 3 represents an overview of how the environmental factors included in a road safety model will be categorised.

Environmental factor	Category
Road function class	1. Residential 2. Primary 3. Secondary 4. Cycleway 5. Unclassified 6. Tertiary
Surface pavement	1. Asphalt 2. Paving stones 3. Gravel
Intersection	1. Present 2. Not present
Speed limits	1. 0 – 15 km/h 2. 15 – 30 km/h 3. 30 – 50 km/h 4. 50 – 130 km/h
Traffic direction	1. One way 2. Two way
Safety island	1. Present 2. Not present
Cyclists volume	1. 0 – 250 2. 250 -500 3. 500 – 1000 4. 1000 – 2000
Motorised vehicle volume	1. 0 – 1000 2. 1000 -2000 3. 2000 – 4000 4. 4000 - 8000
Road segment length	Numerical value

Table 4: Table of Environmental factors involved in the analysis

3.2. Safety Parameter Estimation

Road safety estimation requires negative binomial Regression model construction. This section will explain the main concepts that will be used during regression modelling, how the gathered data and theory will be applied.

Firstly, all the categorised data will be processed as “dummy variables” in order to be included in a regression modelling (Fox, 2010). Afterwards, the formula for road safety modelling was presented in chapter 2.3.1. Negative Binomial Regression Modelling will be implemented in the model using “patsy” syntax. This method is usually used for the regression type model and provides wide range of opportunities (Mummert, 2023).

The next step in the modelling will be to find appropriate distribution for the model. In statistical analysis, the choice of distribution for generalized linear regression models depends on whether the count data exhibits over-dispersion or not. Over-dispersion is when the variance of the count data is greater than the mean, which is commonly observed in count data in practice. One commonly used distribution for count data is the Poisson distribution, which assumes that the variance is equal to the mean. However, if the count data is over-dispersed, a negative binomial distribution may be a better fit, as it allows for greater variability (Reurings & Jassen, 2006).

To determine which distribution is most appropriate, the level of over-dispersion needs to be approximated. The dispersion parameter, denoted by ϕ, is used to estimate the over-dispersion in the data. For a Poisson distribution, this can be approximated by calculating the ratio between the Pearson's chi-squared and the residual degrees of freedom of the Poisson model. If this ratio is (sufficiently close to) 1, it is assumed that the data is approximately Poisson distributed, and the assumption that the variance is larger than the mean cannot be rejected. However, if the ratio is significantly larger than 1, the data is found to be over-dispersed, and the negative binomial distribution may be a better fit.

One way to estimate ϕ in a negative binomial model is to first develop a model based on the Poisson distribution and estimate ϕ using the residuals of this model. The derived value of ϕ can then be used to fit a model based on the negative binomial distribution (Bentem, 2022).

Finally, when all the steps are complete, it becomes possible to estimate coefficients for every factor that was included in the analysis. Nevertheless, couple of assumptions need to be taken. First of all, dataset should not contain any outliers and secondly there should be no collinearity between factor. Outliers were already eliminated on the stage of data gathering, therefore, multicollinearity check has to be performed.

After the predictions will be made the Empirical Bayes method will be applied to the observed data in order to improve the precision of estimates by incorporating prior information. The method is particularly useful when the sample size is small or the data is limited, as it allows for more accurate parameter estimation. Additionally, the method is known to correct for the regression-to-mean bias, which is a common issue when analysing data with extreme values and leads to artificially high number of predicted values. By accounting for the prior information and correcting for the regression-to-mean bias, the Empirical Bayes method produces more reliable estimates that are less likely to be affected by outliers or other sources of variability in the data (Hauer & Harwood, 2002).

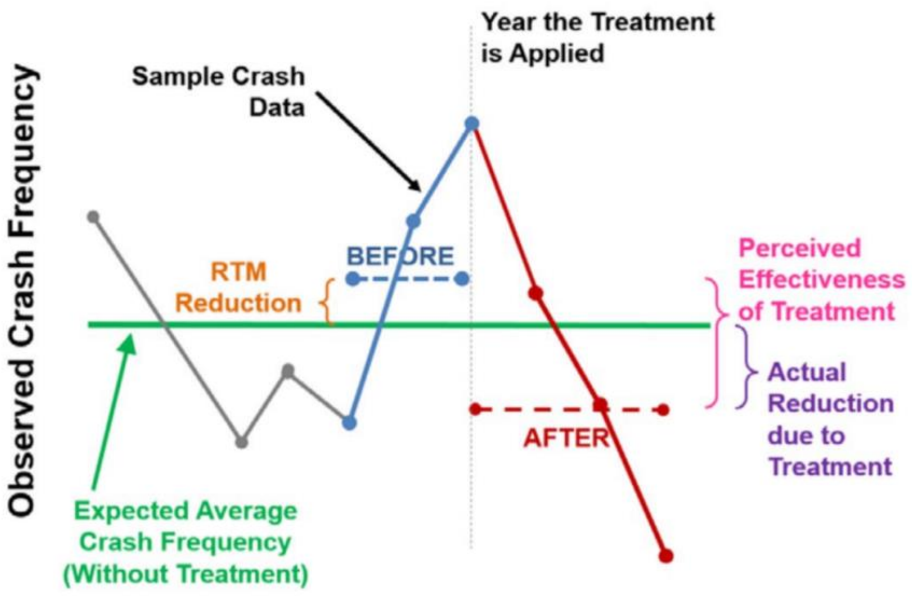


Figure 9: Empirical Bayes method concept illustration

This will be done in a two steps, first the weight will be calculated for every prediction using the following formula:

Equation 4: Weight calculation for overdispersed predictions

$$weight = \frac{1}{1 + \frac{\mu \times Y}{\phi}}$$

Where μ stands for the expected number of crashes based on the regression model, Y stands for the number of years when the data was collected and the ϕ corresponds to the overdispersion parameter.

After, the predicted values will be recalculated using the formula:

Equation 5: Calculation of the predicted number of crashes using Empirical Bayes method

$$N_{predicted} = weight \times \mu + (1 - weight) \times N_{crashes}$$

In this formula $N_{crashes}$ represent the actual number of crashes that was registered on the road segment.

3.3. Road safety Influence on the Route Attractiveness

When the road safety is defined for each road segment, the next step is to analyse it influences the overall network attractiveness. First, the main concepts of the logistic regression model will be briefly covered. Afterwards, the procedure of road safety implementation will be described.

3.3.1. Logistic Regression Modelling

In this study, a previously developed model that was used to analyse environmental factor influence on the route attractiveness will be adopted. Generally it was based on the four main steps. Firstly, the shortest paths and observed routes were described, using data from both the "fietstelweek" and "OpenStreetMap". Secondly, infrastructural and land use allocation characteristics were connected to the segments of the shortest paths and observed routes. The data from different sources such as the "Basisregistratie Grootschalige Topografie", the "Wijken en Buurten" data, and the "RUDIFUN1" data were combined with the segments of the observed routes and shortest path. The third step involved making a comparison between shortest paths and observed routes, which is referred to as the segment approach. Finally, a regression model was constructed (Nijen, 2022).

Segment approach applied in this model required describing the observed routes and shortest path by calculating the shortest direct route using the Dijkstra graph theory algorithm. Next, all segments were described based on infrastructural and land use allocation factors using the data from different sources, involved factors are listed on the Table 2. After all segments were characterized, the third step involved a comparison between the segments, comparing those chosen by cyclists with those not chosen but recommended when using the shortest path. Finally, a regression model was constructed based on the results from the previous step.

While the segment approach was useful in comparing characteristics on a segment level, it had its limitations, including the fact that route characteristics were hard to include. Moreover, trip length could not be directly included, although it is a significant factor influencing the route choice of cyclists. However, it could be included indirectly using ratios between the length of the observed route and the shortest path in a multinomial regression model. Figure 10 illustrates the segment approach concept.

Before regression modelling could be implemented, the Pearson r coefficient was used to examine correlations between all variables included in the study. When the Pearson r coefficient exceeded ±0.8, one of the two independent variables was eliminated before executing regression. Furthermore, since the dependent variable was binary and the independent variables were non-continuous, no outliers were expected to occur.

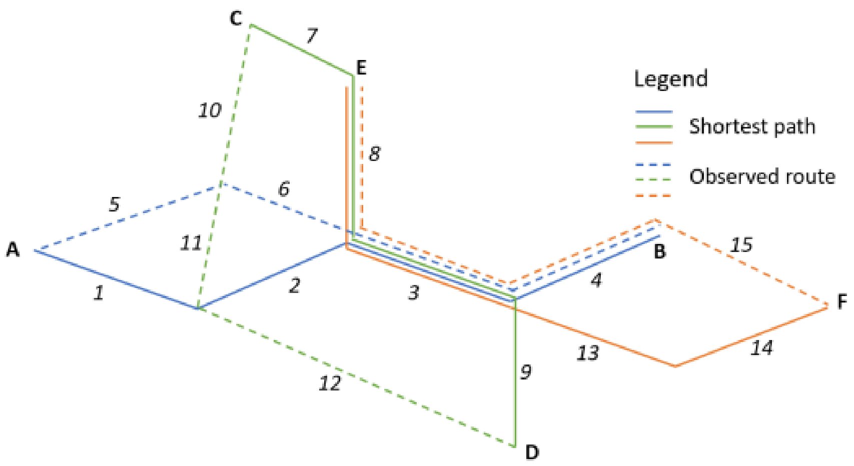


Figure 10: Segment approach illustration (Nijen, 2022)

At the end of the regression modelling, in order to get a better understanding about different factors effect, regression coefficients are standardised. Because the factors do not use the same scales or measurement unit, their coefficients does not directly represent their importance toward the road attractiveness predictions. To solve this problem, standardized regression coefficients are utilized to measure the importance of the influential factors. By using units of standard deviations, standardized regression coefficients overcome this issue (Siegel & Wagner, 2022). The calculation of standardized regression coefficients is demonstrated in the following equation:

Equation 6: Logistic regression coefficients standardisation

$$\beta_{st,i} = \beta_i \frac{\sigma_{x,i}}{\sigma_y}$$

Where $\sigma_{x,i}$ stands for the standard deviation of the data corresponding to the coefficient β_i; $\sigma_y$refers to the standard deviation of the variable that was threated in the regression analysis as dependent.

3.3.2. Road Safety Implementation

Upon completion of road safety estimation for each road segment in the network, an analysis of its impact on route attractiveness is required. To achieve this objective, the logistic regression model, which was previously described, will be utilized. The road safety as a predicted number of crashes will be assigned to each route segment of the network. Additionally, an assessment of multicollinearity with other factors will be conducted. Subsequently, the results will be visually compared, model performance indicators will be evaluated, and validation and verification procedures will be executed.

3.4. Validation and Verification

Validation and verification are essential processes in scientific research. This study will require to visually compare results. Moreover, data statistical tests will be performed. At the end the obtained results will be checked with previously performed results in terms of the model performance indicators. This will help to ensure the validity and reliability of the research, and ensure that any conclusions drawn from the data are accurate and trustworthy. Moreover, the conclusion regarding chosen methodology and assumptions will be made. Overall, validation and verification are critical steps in the research process that help to ensure the quality and accuracy of the results.

Visual comparison will be done between two attractiveness maps, the first one where the road safety was considered and the second one no road safety was includes in the logistic regression, Therefore, it would be possible to see if road safety has any significant effect on the road attractiveness and if it does, how exactly it effects cyclists behaviour.

Data statistical tests application on a dataset can help to answer important research questions, such as whether there is a significant difference between two groups or whether a treatment has a significant effect on an outcome. Therefore, tests such as Kolmogorov-Smirnov and Chi-square tests will be applied to see if the road safety implementation has any significant influence on the road attractiveness.

Furthermore, model key performance indicators will be checked to indicate if the road safety influence the accuracy of the model. Statistical performance of the model can be evaluated using a test dataset. This can be done by creating a confusion matrix, which shows the number of true positives, true negatives, false positives, and false negatives. Four key performance indicators can be extracted from the confusion matrix: accuracy, precision, sensitivity, and F1 score (Behesthi, 2023). Accuracy represents the proportion of correct predictions made by the model, while precision indicates the proportion of positive predictions that are true. Sensitivity refers to the proportion of negative predictions that are true. The F1 score provides a balanced measure of precision and sensitivity. It's important to note that the evaluation should be done on a separate dataset from the training dataset to ensure that the model is able to generalise well. Furthermore, the specific evaluation metrics used should be tailored to the particular task. Achieving high scores for all key performance indicators is an indication of a well-performing model. However, the level of performance required can vary depending on the task and dataset. This study will compare the performance test results to assess how the implementation of the road safety in the analysis impacts the model's performance.

4. Results

This section will be dedicated to results that were retrieved after required modelling was completed.

4.1. Predicted Number of Crashes on Road Segments

In this chapter, the results of a negative binomial regression analysis that aimed to predict the number of crashes on road segments will be presented. All factors were included in the analysis, as the predefined threshold for correlation in this study was set to be between +0.7 and -0.7. Therefore, any factor with a correlation coefficient falling within this range was considered acceptable and included in the study. Correlation heat map of the risk factors is presented in the Appendix – A. The Poisson ratio previously indicated as ϕ was estimated as 59.99. Based on this value, it was concluded that data is over-dispersed and negative binomial distribution will be used.

The presented table showcases the results of the Negative Binomial Regression modelling as well as the coefficients of the risk factors obtained from the regression modelling. The obtained coefficients provide insight into the factors that have a significant influence on cyclists' safety. Pavement type, road function classes and the location of the road intersection next to the route segment were found to have a significant impact on cyclist safety.

Generalized Linear Model Regression Results			
Dep. Variable	crashcount	No. Observations	168896
Model	GLM	Df Residuals	168878
Model Family	NegativeBinomial	Df Model	18
Link Function	Log	Scale	1.0000
Method	IRLS	Log-Likelihood	-64319
No. Iterations	42	Deviance	98120
Covariance Type	nonrobust	Pearson chi2	8.12e+14
		Pseudo R-squ. (CS)	0.7180

Variable	Coefficient	Std Err	Z-Score	P-Value	95% Conf. Int.
Intercept	-6.1328	0.031	-200.629	0.000	[-6.193, -6.073]
0-15km/h	0.6790	0.033	25.673	0.000	[0.598, 0.778
15-30 km/h	0.8650	0.029	30.293	0.000	[0.809, 0.921]
30-50 km/h	0.7508	0.040	18.548	0.000	[0.671, 0.830]
50-130 km/h	-0.6874	0.045	-15.258	0.000	[-0.776, -0.599]
motorised volume	0.3244	0.007	46.626	0.000	[0.311, 0.338]
cycle volume	0.9096	0.021	42.370	0.000	[0.868, 0.952]
asphalt	0.4331	0.022	19.455	0.000	[0.389, 0.477]
paving stones	0.7328	0.023	31.450	0.000	[0.687, 0.778]
residential	-0.2128	0.082	-2.583	0.010	[-0.374, -0.051]
gravel	0.8778	0.031	28.210	0.000	[0.817, 0.939]
tertiary	1.3754	0.044	30.912	0.000	[1.288, 1.463]
cycleway	0.2558	0.034	7.617	0.000	[0.190, 0.322]
unclassified	1.0986	0.032	34.633	0.000	[1.036, 1.161]
secondary	1.8344	0.054	34.198	0.000	[1.729, 1.940]
primary	0.4879	0.063	7.705	0.000	[0.364, 0.612]
Island	0.5006	0.044	11.278	0.000	[0.414, 0.588]
oneway	0.8747	0.027	32.138	0.000	[0.821, 0.928]
intersection	2.9074	0.017	174.206	0.000	[2.875, 2.940]

Table 5: Negative Binomial Regression coefficients

Table 5 represents the results of completed modelling. The regression analysis utilised a Generalized Linear Model (GLM) with a negative binomial distribution and log link function to estimate the number of crashes. The model has 18 degrees of freedom and a scale of 1.0. The maximum likelihood method was employed, using iteratively reweighted least squares (IRLS), to estimate the model parameters. The log-likelihood was found to be -64318, and the deviance was estimated as 98120. The covariance type was non-robust, and the Pearson chi2 test statistic was 8.12+14, which can be considered as high an effect in a poor modelling predictions quality. The pseudo R-squared (CS) value was calculated to be 0.7180, indicating that the model explains 71.80% of the variance in the dependent variable. These findings suggest that the GLM with negative binomial distribution and log link function is a suitable fit for the data and has significant predictive power for predicting the number of cycle crashes. However, the features of analysed dataset can explain models inaccuracies. Namely, proportionally low number of road segments with occurred crashes to a high number of road segments where the crashes occurred could negatively affect the models prediction power.

The coefficients indicate the degree to which each factor affects the risk of crashes. Among the variables with positive coefficients, the highest impact on crash risk is associated with the "intersection" variable, with a coefficient of 2.9074. This indicates that areas with junctions have a much higher risk of crashes than road segments without adjacent road intersections. Other variables with significant positive coefficients include "gravel", "15 – 30 km/h", "tertiary" and "secondary" road function classes indicating that areas with gravel roads and with speed limits of 15 – 30 km/h are the areas with the highest probability of experiencing a bicycle accident.

The representation of the road safety map is illustrated on the Figure 11.

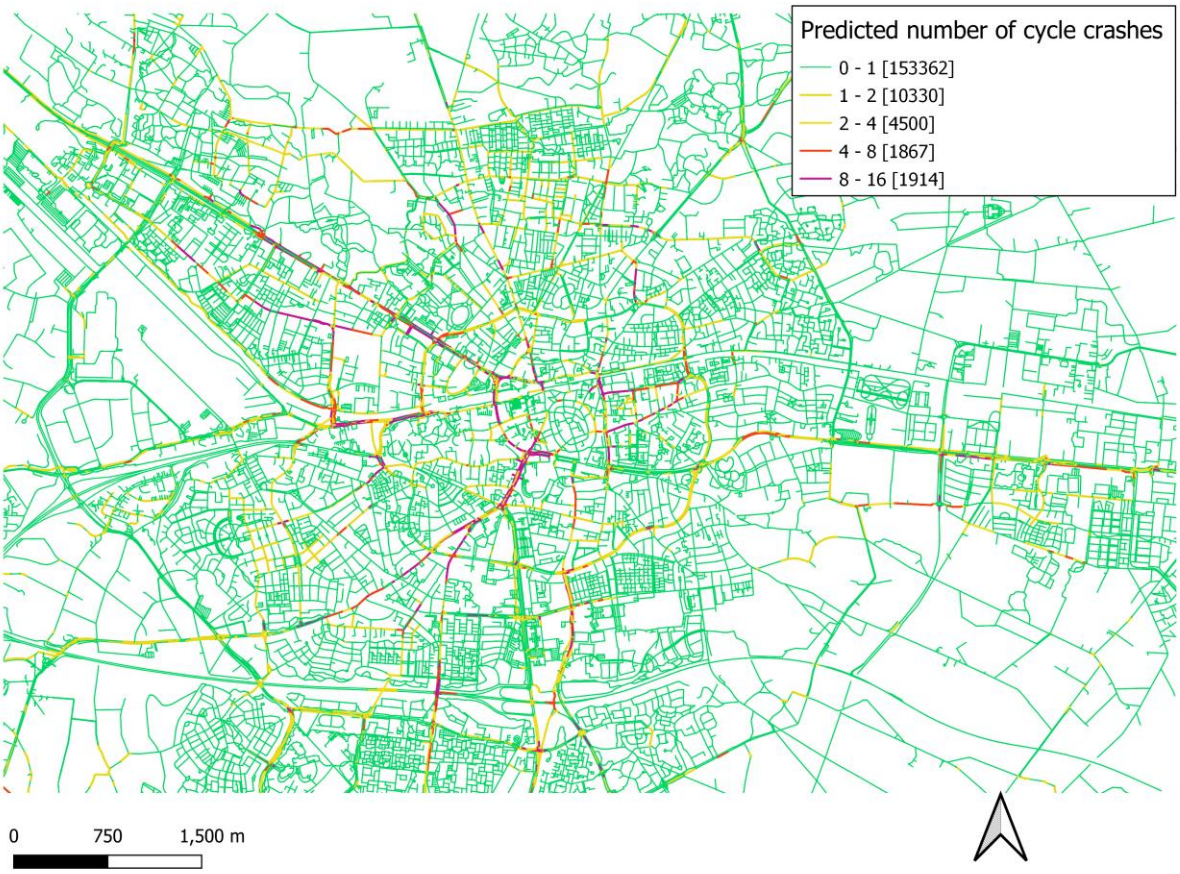


Figure 11:Cycle crashes prediction map of Enschede

Upon visual analysis of the map, it is evident that the major road links and the ring road surrounding the city of Enschede have the highest predicted number of crashes. Conversely, minor links exhibit a significantly lower number of predicted crashes. Specifically, the Hengelostraat, Gronausestraat, and Haaksbergerstraat, as well as the ring road, are highlighted as having higher predicted crash rates in comparison to other links with looks natural.

As it can be concluded, roads with higher predicted crash numbers correspond to roads where with a high traffic volumes. From the map, it can be seen that areas of road intersections are also highlighted as a road segments with a predicted possible cycle crash. Residential and cycleway road function classes correspond to lowest regression coefficients, and it can signalise that these road function classes correspond to a lower number of predicted crashes on the road segments.

After the application of the Empirical Bayes method, the number of predicted crashes significantly decreased, however the general trend remained the same. This could meant that the number of outliers has decreased and extreme values were eliminated from the dataset. Major roads have a higher number of predicted crashes while minor roads have 0-0.1 predicted crashes per road segment. Figure 12 illustrates the road safety map of Enschede after Empirical Bayes method application.

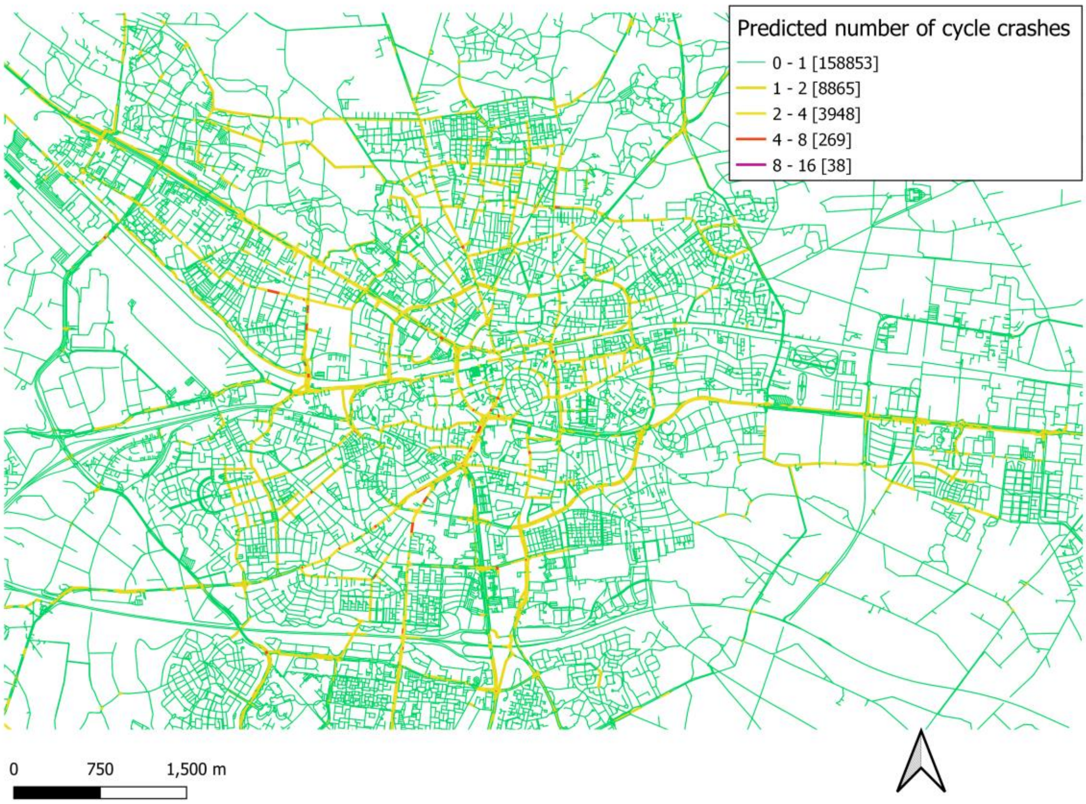


Figure 12: Cycle crashes prediction map of Enschede

4.2. Road safety Influence on Route Segments Attractiveness

When the road safety was calculated for each road segment it becomes possible to check how it influences road attractiveness. Multicollinearity check of the road safety with environmental factors was performed and no collinearity was detected. Correlation heat map is presented in Appendix – A.2.

After the logistic regression was performed, the new coefficients for the factors were obtained.

id	Factor	Standardised coefficients(Road safety included)	Standardised coefficients(Road safety not included)
0	Distance to traffic control installation	-0.133	-0.124
1	Road safety	0.157	-
2	Cycle lane	0.070	0.095
3	Separate cycle path	1.178	1.072
4	Artificial lighting	-0.246	-0.246
5	Asphalt	0.443	0.699
6	Motorised vehicle intensities	-0.532	-0.675
7	Bicycle intensities	1.203	1.108
8	Residential land use zone	0.645	0.689
9	Commercial land use zone	-0.051	-0.446
10	Industrial land use zone	-0.448	-1.478
11	Green land use zone	-0.361	-0.227
12	Land use mix	0.033	0.034
13	Degree of urbanisation	0.418	0.365
14	Constant	0.000	0.000

Table 6: Standardised coefficients with and without road safety being included

The findings from Table 6 suggest that the inclusion of road safety did not have a significant impact on the majority of coefficients of the regression factors that were already used in the analysis. However, a more comprehensive understanding of the impact of road safety on the model can be obtained by examining Figure 13. The figure reveals that the implementation of road safety measures had a minor effect on some environmental factors, such as cycle lane separation, distance to the traffic control installation or the presence of a cycle lane. However, the coefficient for road safety itself was found to be relatively small comparing to the coefficients of other factors. Further discussion on the possible reasons behind this finding will be presented in the subsequent chapter.

Although, the road safety addition to the road attractiveness analysis did not significantly effect most of the coefficients, 3 factors has a notable change. Asphalt pavement factor decreased by 0.256, commercial land use zone had an increase of 0.395 and industrial zone had an increase of 1.03 in their coefficients.

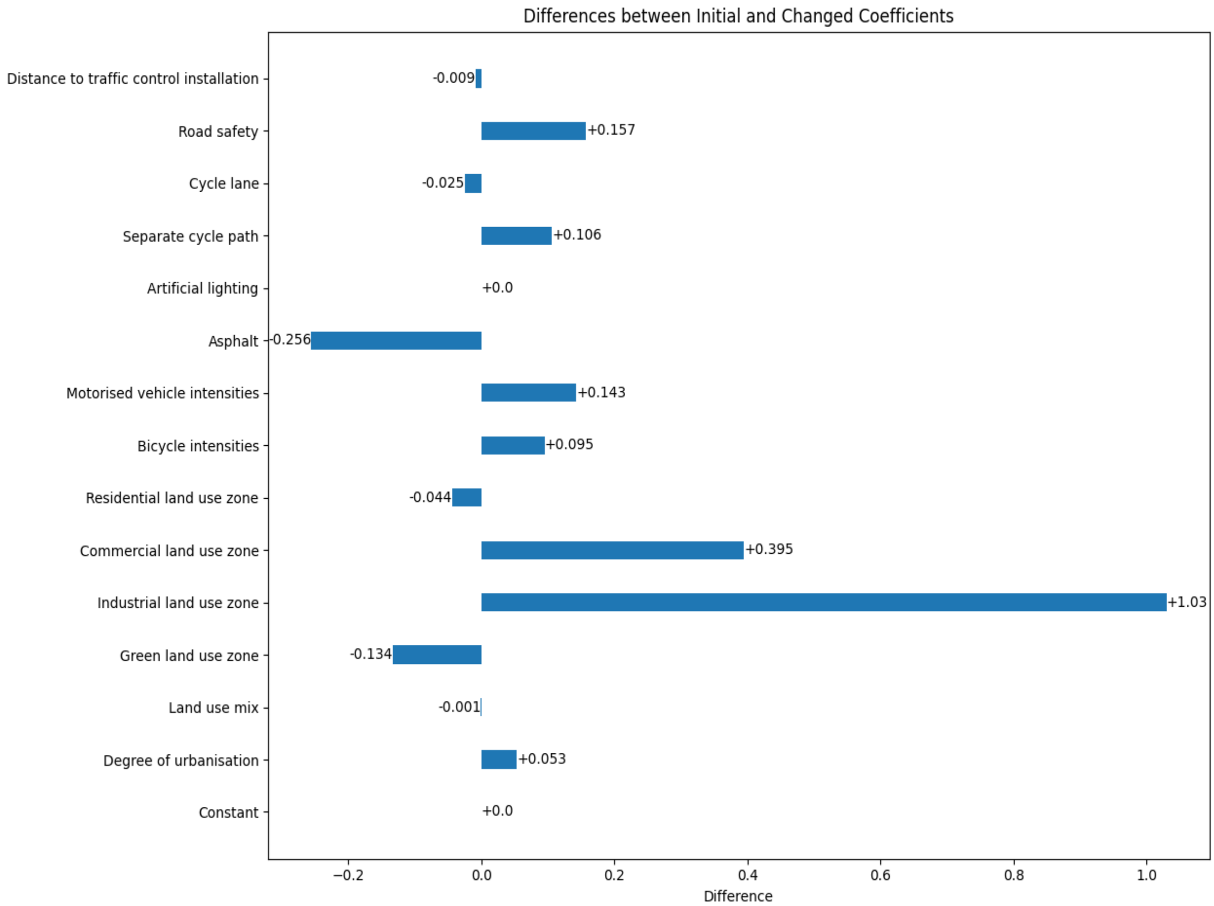


Figure 13: Coefficient difference

As a result of logistic regression modelling, road attractiveness map was obtained. Figure 14 illustrates the road segment attractiveness in the Enschede. In other word, this map represents a possibility of each road segment to be by cyclists. First, the visual comparison was performed, based on it, no significant differences was identified between attractiveness maps with and without road safety implementation in the analysis. However, based on the counted number of road segments with different attractiveness level, the addition of road safety to the analysis did not significantly change the overall attractiveness level of the road segments. The difference between the two maps is minor, indicating that road safety did not have a significant impact on the attractiveness level of the road segments.

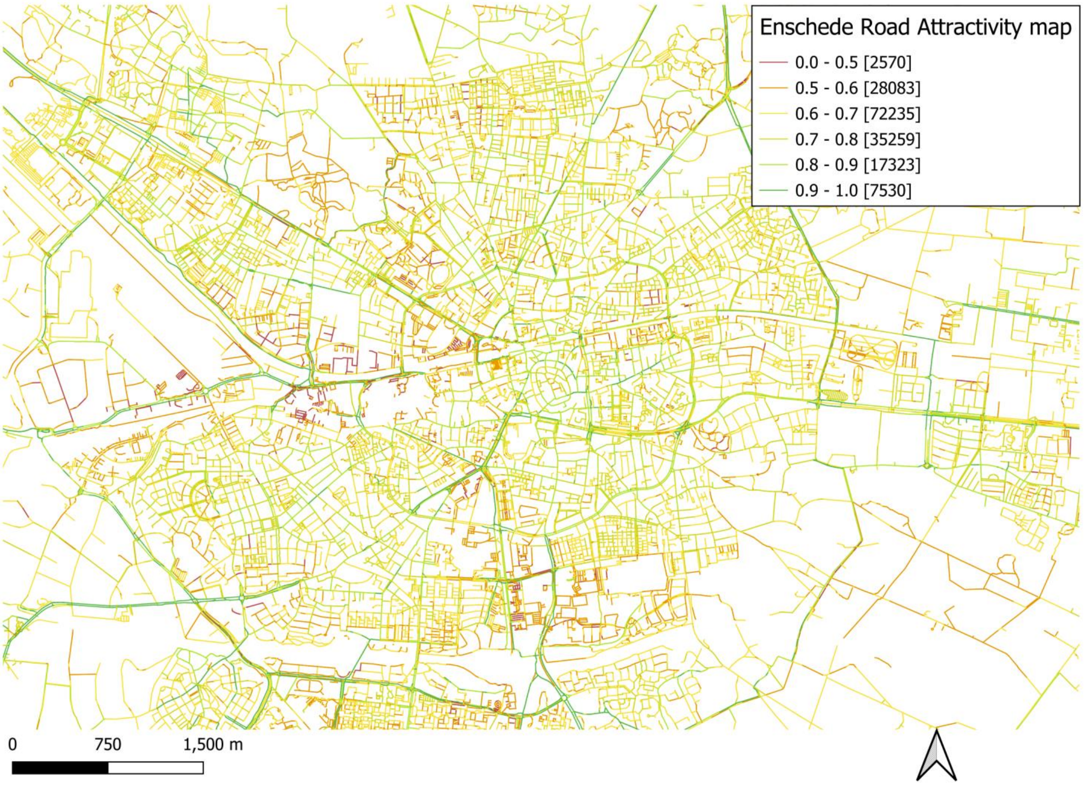


Figure 14: Road attractiveness map(road safety is included)

Based on this illustration, it clear that the most preferred roads by cyclists are the major links and cycle paths. Comparing it with the attractiveness map that does not include road safety it becomes clear that addition of a road safety does not make any significant difference to the overall network attractiveness.

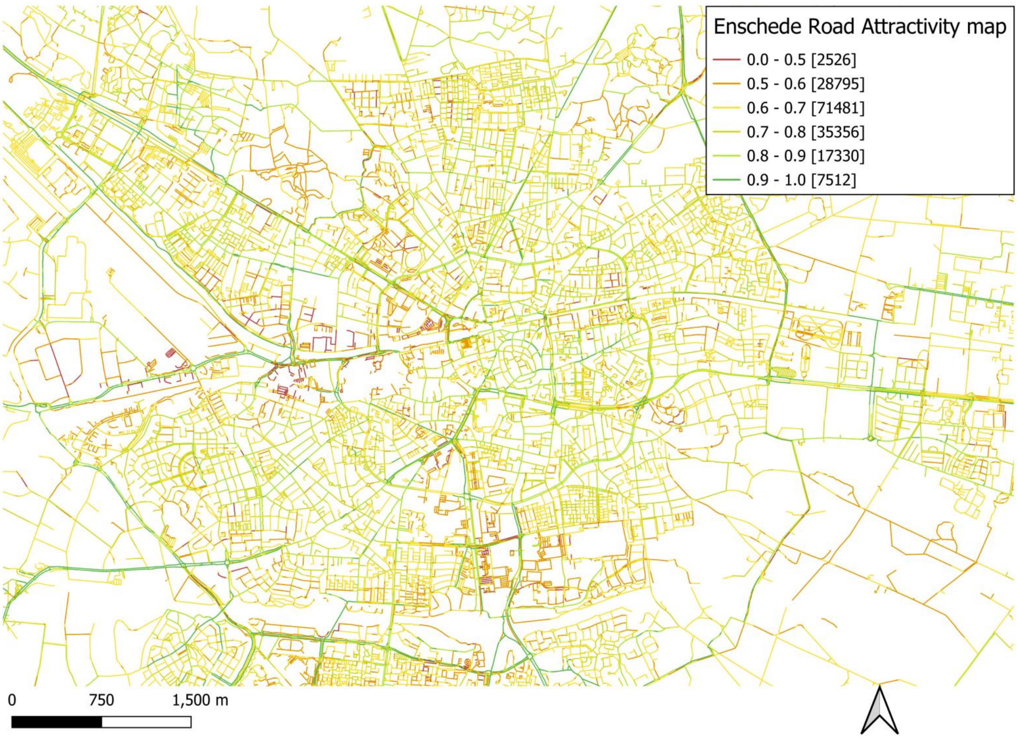


Figure 15: Road attractiveness map(road safety is not included)

Both maps can be found in the appendix Appendix – B.

4.3. Results Validation and Verification

After observing the results and finding a minor difference between the presence or absence of previously modelled road safety in the analysis, it is crucial to validate and verify the outcomes to determine the usefulness and validity of the road safety addition to the analysis. This will provide a better understanding of whether the road safety data contributed to the analysis and whether it was a reasonable addition. It is essential to ensure the reliability and accuracy of the results to prevent any misinterpretation or incorrect conclusions. Further investigation and validation are necessary to confirm the validity of the findings.

4.3.1. Model Performance Comparison

In order to check how the inclusion of the road safety effected the model, the performance indicators that were previously discussed in 3.4. Validation and Verification chapter, will be checked.

	accuracy	precision	sensitivity	F1 Score
Model without road safety	0.817528	0.817866	0.999434	0.899580
Model with road safety	0.809868	0.812449	0.994597	0.894343

Table 7:Performance indicators comparison

As it can be read from the Table 7, the road safety implementation in the model did not significantly influence its performance. Therefore, road safety did not considerably decreased or increased the performance of the model. However, this can be explained by the insignificant overall effect of the road safety on the logistic model predictions.

4.3.2. Data Statistical Analyses

In order to assess the significance of implementing a road safety in the model, statistical tests were conducted. These tests serve the purpose of evaluating whether the observed data could have arisen by chance, or if there is a statistically significant difference between two sets of data. By performing such tests, it is possible to gain insight into the degree to which the implementation of a road safety has an impact on the overall results of the model.

Test name	Result
Kolmogorov-Smirnov test statistic	0.011
Kolmogorov-Smirnov test p-value	0.0
Chi-square test statistic	5427.672
Chi-square test p-value	0.0

Table 8: Results of statistical tests

Results of the statistical tests are represented in the Table 8. The Kolmogorov-Smirnov test yielded a test statistic of 0.011 and a p-value of 0.0, indicating a statistically significant difference between the two datasets. These results suggest that the datasets do not come from the same distribution. The chi-square test similarly indicated a significant association between the two datasets, with a test statistic of 5427.672 and a p-value of 0.0. Overall, the results of all tests suggest a difference between the two datasets.

For better understanding, the graph of difference between road attractiveness with and without road safety included versus the values road attractiveness with road safety was plotted.

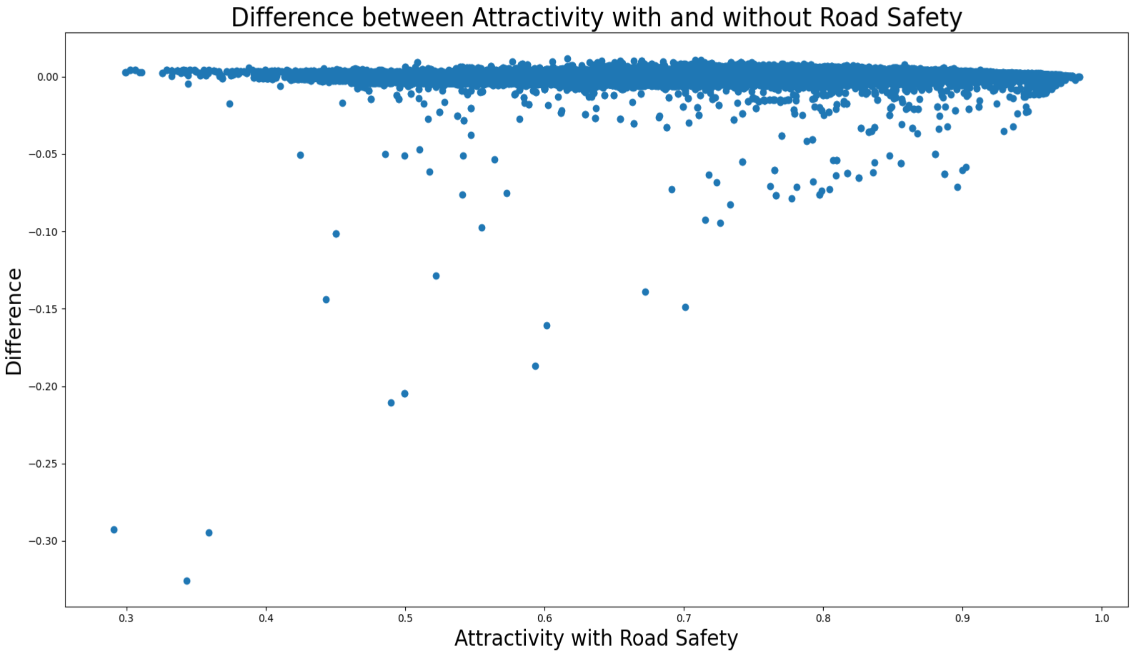


Figure 16: Difference between Attractivity with and without Road Safety

As it can be seen on the Figure 16 and concluded from statistical tests, predicted attractiveness data is different, but not significantly. Road safety implementation in the analysis definitely has an effect on the predicts. However, as it can be seen from the attractiveness maps , performance indicators and statistical tests, this effect I small comparatively to other environmental factors and addition of the road safety in the form as it was added in this research is debatable and will be discussed in the further report sections.

5. Discussion

5.1. Modelling Approach

The modelling approach used in this study involved the use of negative binomial and logistic regression models to investigate the factors influencing the safety of cyclists on road segments and their route choice decisions. Although these models have been commonly used in transportation safety research, there are limitations to their application.

One of the limitations of the modelling approach is that the data used to develop the models may not be entirely accurate or representative of the actual conditions on the road. In this study, the BROM data was based on a different road network geometry. The geometry was less accurate comparing to the network that was forming the data frame for the further analysis. This could potentially lead to inaccuracies in the results. The same issue was faced with motorised vehicle volumes data.

Regarding the adopted methodology for interpreting the road safety as the predicted number of crashes, it can be inferred that the approach provided pragmatic and authentic outcomes, emphasising major thoroughfares and the most favoured cycling routes as having the highest risk of crashes. Therefore, within the scope of this study, the methodology proved to be a fitting choice. Nevertheless, additional factors may be integrated into the negative binomial regression analysis, and diverse intersections may be categorized differently to strive for more precise outcomes. By doing so, it may be possible to achieve a higher degree of accuracy in assessing the influence of safety considerations on cyclists' route preferences.

It should be noted that the number of road link with registered crashes in the dataset is relatively small comparing to the total number of links in the network, this may affect the statistical power of the models and limit the conclusions that can be drawn from the analysis. While the negative binomial regression model is appropriate for modelling count data with over dispersion, the limited number of crashes may still result in unstable estimates and high variability in the coefficients. Additionally, the limited number of crashes make it difficult to identify all of the significant factors that affect cyclist safety.

Despite these limitations, the findings of this study provide important insights into the factors that affect cyclist safety and route choice decisions. Policymakers and transportation planners can use this information to develop targeted interventions and infrastructure improvements that improve cyclist safety and promote cycling as a sustainable mode of transportation.

5.2. Results Discussion

After conducting validation and statistical tests, the impact of implementing the road safety on the overall road network attractiveness can be considered as insignificant. The inclusion of a road safety , interpreted as the predicted number of crashes on a road segment, resulted in minor changes in the attractiveness of the road segments that was estimated not incorporating road safety as a separate regression factor.

The number of reasons can explain this. First of all, the regression coefficient of the road safety in the logistic regression has a positive value, what can be interpreted as the fact that the higher number of predicted crashes on the road segment will make the road segment more attractive for the cyclists. This does not sound logical and can be caused by the fact that road safety was modelled on the same factors that were used in the logistic regression. In other words, this can be interpreted as that the road safety was already indirectly included in the road attractiveness prediction model.

Another possible cause, is that the negative binomial model predicts the high number of crashes for the roads that are further predicted by the logistic model as the most attractive roads for the cyclists. This is logical since most of the crashes occur on the roads that are more frequently used by the cyclists and other road users. Therefore, implementation of the road safety to the logistic model as a separate factor does not bring any significant difference.

One more reason for this could be that the cycling infrastructure in the Netherlands is designed in such a way that cyclists feel secure and self-assured while traversing most of the road sections. Also, it could be explained by the mind set of Dutch cyclists, as it was previously mentioned, cycling as a transportation mode has a strong role in a Dutch culture. As a result, they base their route selection on factors other than road safety, which is a combination of environmental and external factors and does not take into account the cyclist's personal characteristics. It should be noted that the logistic regression model used to estimate route attractiveness for cyclists relied on Fietselweek data, which primarily pertains to experienced cyclists who are confident on the road. While this finding adds value to the current theory, further investigation is required to validate this observation. It is important to mention, that due to the selected scale for the road attractiveness representation on the maps, minor differences as 0.01-0.001 fluctuations between route attractiveness with and without road safety addition are not visible. But even on the intersections that were considered as the factor that mostly effects the number of predicted crashes, the predicted attractiveness of the road segments dropped maximum by 0.02.

In general, the results obtained from the negative regression analysis yielded significant predictions concerning the road safety of the Enschede road network. However, when applying this factor as an independent variable in the logistic model, the expected outcomes were not observed. Therefore, alternative approaches to incorporating road safety as a determinant in the assessment of road attractiveness and examining its impact on cyclists' route selection remain open for future research.

6. Conclusion

In conclusion, this study explored the relationship between road design factors and cyclist safety in the city of Enschede using a negative binomial regression. After, the road safety effect on the road segments attractiveness was analysed using logistic regression model. The results revealed that several factors, including traffic volumes, road pavement type, road intersections, road function classes, speed limits, and number of traffic directions on the road, significantly influenced the safety of cyclists on road segments. However, the logistic regression model showed that the road safety had a minor influence on the attractiveness of road segments for cycling

Initially this study aimed to investigate the road safety impact on the cyclists route choice. Therefore, the road safety interpretation was formulated as a predicted number of crashes per road segment. This approach resulted in getting reasonable results. Therefore, it can be concluded that the selected methodology is appropriate to model road safety and can be used in this type of studies. However, due to a number of previously listed data limitations it can be difficult to apply this methodology on a smaller study area.

In order to answer the question of how the road safety effect the cyclists route choice can be modelled, the logistic regression model was adopted for this study purposes. In general, this methodology worked and gave certain results. However, the final results did not meet the initial expectations of this study indicating that the road safety does not considerably influence the route choice of the cyclists.

Talking about the statistical significance of the results and answering the last sub question of this study it can be concluded that results of the negative binomial regression model are statistically significant. Based on the statistical tests conducted on the data, it is noteworthy that the addition of road safety as a factor had a minor effect on the overall network attractiveness. However, the effect was not statistically significant, meaning that the observed effect could have occurred by chance alone. It is possible that with a larger sample size or different statistical methods, the effect of road safety on network attractiveness could become statistically significant. Nonetheless, the current findings from the analysed literature suggest that road safety can be an important factor to consider when analysing and improving the attractiveness of cycling routes, even if the effect is not statistically significant in the present study.

It is necessary to understand all the limitations of this study and applied methodologies while drawing conclusions about how the safety affect the route attractiveness for cyclists. Nevertheless, a general conclusion of this study is that road safety that was modelled on the environmental factors and traffic volumes inclusion in cyclists route choice model that uses mostly the same factors does not give significantly valuable results.

Answering the general question of this study, which is "To what extent does road safety impact route attractiveness for cyclists?", the following can be inferred. Although the precision of the chosen approaches is arguable due to the limitations of the data discussed in earlier sections, the method adopted in this study determined that road safety does not significantly affect cyclists' route selection. However, this conclusion warrants two considerations. First, it may be contended that road safety does indeed influence cyclists' route choice, as a number of studies have highlighted the importance of road safety in determining route preferences. Second, it appears more probable that the approach utilised to project road safety onto network attractiveness did not completely capture the impact of road safety on route choice. As such, further research is necessary to investigate the relationship between road safety and route selection for cyclists, employing alternative methodologies or integrating additional factors to better capture the effect of road safety on cyclists' route decisions.

7. Future Research Recommendations

Summarising the findings of this research, a set of recommendations will be provided for future researchers to consider.

In order to improve the accuracy of the negative binomial regression modelling, future research could consider using more accurate and detailed data for the analysis. Additionally, more research could be conducted regarding the other factors that may influence cyclist safety and route choice decisions. Also, more complex modelling can be incorporated to better capture these factors. Generally, selection of a bigger study area with a significantly bigger data frames will result in the better modelling outcomes.

Moreover, another possible solution to increase a statistical power of the negative regression model could be classification and categorisation of the road sections or different regions within the selected study area. Another possible solution of enhancing the predictions accuracy of the model could be implementation of different data sampling methodologies.

Furthermore, it is recommended that future studies explore alternative approaches to interpreting the road safety, given that this study utilised predicted crash rates as the measure of safety. And this methodology needs to be validated against other possible approaches to quantify road safety.

A recommendation for future research regarding the projection of the road safety effect on the route attractiveness for the cyclists would be to investigate alternative approaches to incorporating road safety into route attractiveness models. It can be recommended not to implement the road safety to the route attractiveness model that uses almost the same factors that were used to predict the number of crashes. As this study has demonstrated, this particular methodology is not suitable for capturing the impact of road safety on cyclists' route choice. Therefore, alternative approaches need to be explored to achieve this objective.

Additionally, the findings of this study could be used as a starting point for future research examining other factors that may impact cyclists' safety. Overall, these recommendations could help to improve the understanding of the road safety effect on cyclists' route choice and provide insights into how to study the effect of the road safety on the route attractiveness for the cyclists.

Bibliography

AmohGyimah, R., & Saberi, M. (2016). Macroscopic modeling of pedestrian and bicycle. *Accident Analysis and Prevention*, 147–159. doi: https://doi.org/10.1016/j.aap.2016.05.001

Arhin, S. A., & Gatiba, A. (2020). . Predicting crash injury severity at unsignalized intersections using. *Transportation Safety and Environment,*, 120–132. doi:https://doi.org/10.1093/tse/tdaa012

Aultman-Hall, L., & Kaltenecker, M. G. (1999). Toronto bicycle commuter safety rates. . *Accident Analysis & Prevention*, 31(6), 675-686.

Behesthi, N. (2023, February 23). *Guide to Confusion Matrices & Classification Performance Metrics.* Retrieved from towardsdatascience: https://towardsdatascience.com/guide-to-confusion-matrices-classification-performance-metrics-a0ebfc08408e

Bentem, L. M. (2022). *The impact of infrastructure.* Delft : TU Delft.

Breda University of Applied Sciences. (n.d.). Retrieved from Download bestanden Nationale Fietstelweek 2015, 2016 en 2017: http://opendata.cyclingintelligence.eu/

Centraal Bureau voor de Statistiek. (2022). Retrieved from Wijken en buurten: https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/overig/wijken-en-buurten

Cordovil, R., Merce, C., Branco, M., & Lopes, F. (2022). *Learning of Cycle: A Cross-Cultural and Cross Generational Comparison. .* Frontiers in Public Health.

de Dios Ortúzar, J., & & Willumsen, L. G. (2011). *Discrete Choice Models.* Modelling Transport.

Dill, J., & Gliebe, J. (2008). *Understanding and Measuring Bicycling Behavior: a .* Portland: Portland State University.

Dozza, M., & Werneke, J. (2014). Introducing naturalistic cycling data: What factors influence bicyclists'. 24, 83–91. doi:https://doi.org/10.1016/j.trf.2014.04.001

Elsayed, M. A., Sayed, T., & Brown, L. (2016). *Modeling bicycle safety using naturalistic data: Exploratory analysis of the SHRP2 naturalistic driving study.* Transportation Research Record.

Elvik, R. (2017). Risk of bicycle accidents in single-bicycle crashes. *Transportation Research Record*, 2636(1), 59-67.

Fox, J. (2010). *Dummy-Variable Regression.* York SPIDA. Retrieved from https://socialsciences.mcmaster.ca/jfox/Courses/SPIDA/dummy-regression-notes.pdf

Garrard, J., Rose, G., & Lo, S. K. (2008). Promoting transportation cycling for women: the role of bicycle infrastructure. *Preventive medicine*, 46(1), 55-59.

Hauer, E., & Harwood, D. W. (2002). Estimating Safety by the Empirical Bayes Method. *1784. Paper No. 02-2181*.

Heesch, K. C., Garrard, J., & Sahlqvist, S. (2012). Incidence, perceptions and barriers to cycling in Australia: findings from the Cycling Promotion Study. . *Health Promotion Journal of Australia*, 23(3), 210-216.

Huber, S., Lindemann, P., & Muthmann, K. (2022). Modelling bicycle route choice in. *2021 7th International Conference on*, (pp. 1-6).

Khandelwal, R. (2020, April 7). *Quick and Easy Explanation of Logistic Regression*.

Kim, J. K., Ulfarsson, G. F., & Porrello, L. A. (2021(1)). Pedestrian and bicycle safety analysis using detailed real-time GPS data. *Transportation Research Record*, 69-77.

Koch, T., & Dugundji, E. (2021). *Taste Variation in Enviromrntal Features of Bicycle Routes.* International Workshop on Computational Transportation Science.

LaValley, M. P. (2008). Logistic Regression. In *Circulation* (pp. 2395-2399).

Lusk, A. C., Furth, P. G., & Morency, P. (2011). Risk of injury for bicycling on cycle tracks versus in the street. Injury prevention. 17(2), 131-135.

Maathuis, M. (2007). *Applied Regression and Analysis of Variance.*

Misra, A., & Watkins, K. (2018). *Modelling cyclists route choice using revealed preference data: An age and gender perspective.* doi: 10.1177/0361198118798968.

Moudon, A. V., Lee, C., Cheadle, A. D., & Collier, C. W. (2005). Cycling and the built environment, a US perspective. Transportation Research Part D: Transport and Environment. 10(3), 245-261.

Mummert, M. (2023, February). *Using Patsy for Statistical Modeling*. Retrieved from https://medium.com/: https://medium.com/@mummertm/using-patsy-for-statistical-modeling-189a9d9f5d27

Nijen, N. V. (2022). *The influence of infrastructure and land use allocation on the route choice of cyclists .* Enschede: University of Twente.

OpenStreetMap. (2022). Retrieved from OpenStreetMap: https://www.openstreetmap.org/about

PDOK. (2022). Retrieved from Basisregistratie grootschalige topografie (BGT): https://www.pdok.nl/introductie/-/article/basisregistratie-grootschalige-topografie-bgt-

Petegem, J. H., & Wegman, F. (2014). Analyzing road design risk factors for run-off-road crashes in the Netherlands with crash prediction models. *Journal of Safety Research, Volume 49*, 121.e1-127. doi:https://doi.org/10.1016/j.jsr.2014.03.003

Reid, C. (2022, Novenber 21). *Why is cycling popular in the Netherlands: infrastructure of 100+ years of history?* Retrieved from Roads were not built for cars: https://roadswerenotbuiltforcars.com/author/carltonreid/

Reurings, M. C., & Janssen, P. G. (2007). A note on over-dispersion and Poisson regression models. *Journal of applied statistics*, 34(7), 871-879.

Reurings, M., & Jassen, T. (2006). *Accident prediction models and road safety impact assessment: A state-of-art.* Brussel: Europea Comission.

Ribeiro, N. (2014). *Proactief Meten van Verkeersveiligheid ProMeV. Achtergrond, methoden en.*

Riggs, W. (2018). *Perception of Safety and Cycling Behaviour on Varying Street.* San Francisco: University of San Francisco.

Rijkswaterstaat. (2022). Retrieved from Bestand geRegistreerde Ongevallen in Nederland (BRON): https://www.rijkswaterstaat.nl/wegen/wegbeheer/brongegevens/bestand-gerigistreerde-ongevallen-in-

RUDIFUN1. (2022). Retrieved from Planbureau voor de Leefomgeving: https://www.pbl.nl/nl/producten/rudifun1

Schepers, P., & Stipdonk, H. (2017). *Bicycle fatalities: Trends in crashes with and without motor vehicles in The Netherlands. Transportation Research Part F: Traffic Psychology and Behaviour.* doi:https://doi.org/10.1016/j.trf.2016.05.007

Siegel, A. F., & Wagner, M. R. (2022). Multiple Regression: Predicting One Variable From Several Others. In A. F. Siegel, & M. R. Wagner, *Practical Business Statistics* (Eighth Edition ed., pp. 371-431). Academic Press.

Song, R., Ni, Y., & Li, K. (2017). *Inderstanding cyclists' risky route choice behavior on urban road sections.* Transportation Research Procedia.

Teschke, K., Harris, M. A., Reynolds, C. C., & Winters, M. (2012). Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. *American journal of public health,*, 102(12), 2336-2343.

Vandenbulcke, G., Thomas, I., & Int Panis, L. (2014). *Predicting cycling accident risk in Brussels.* Accident Analysis and Prevention. doi:https://doi.org/10.1016/j.aap.2013.07.00

Veenstra, S. (2022, November 10). Bachelor assignment discussion.

von Stülpnagel, R. (2020). *Crash risk and subjective risk perception during urban cycling: Evidence for congruent and incongruent sources. Accident.* Analysis and Prevention. doi:105584. doi:10.1016/j.aap.2020.105584

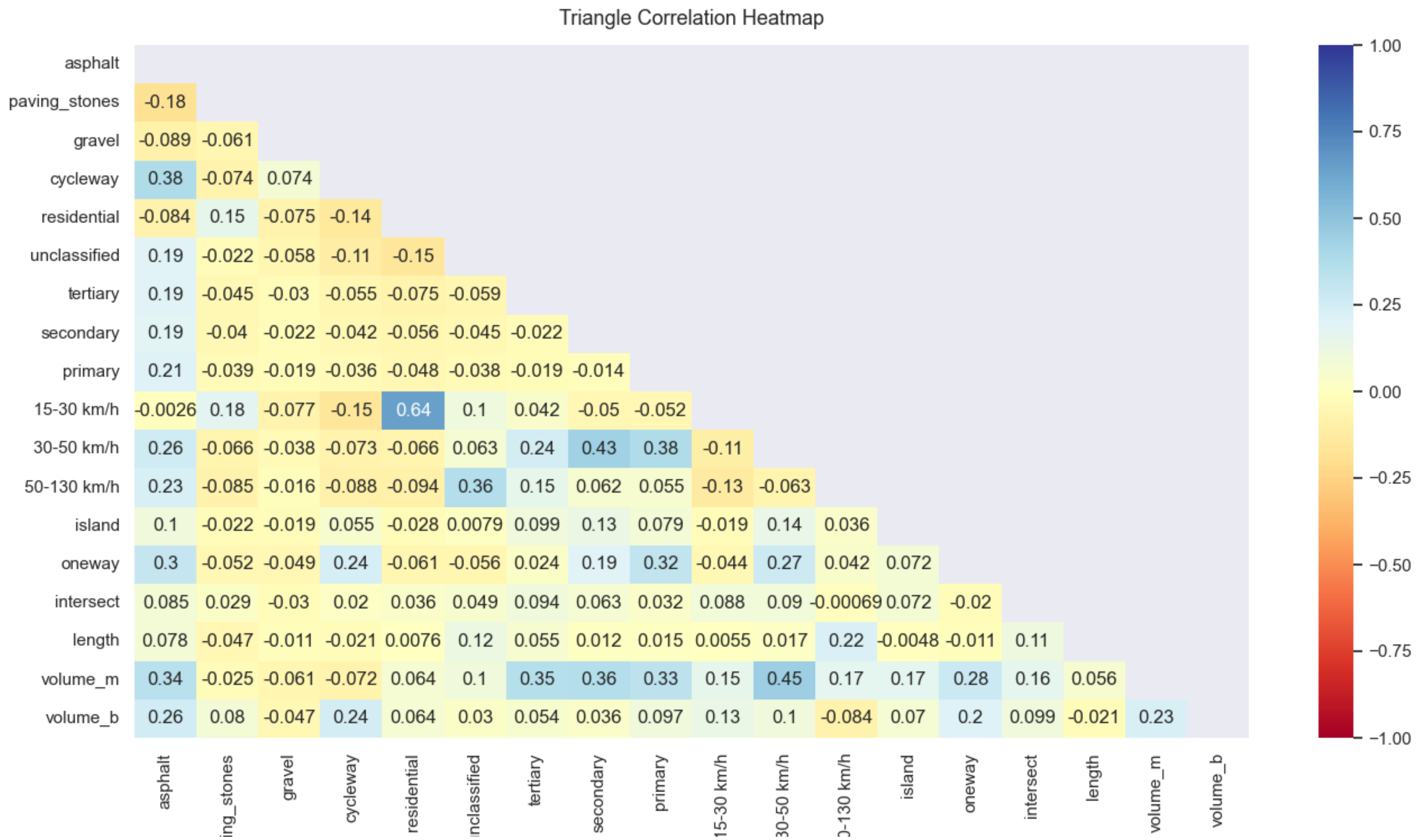
Wei, F., & Lovegrove, G. (2013). *An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression.* Accident Analysis & Prevention. doi:https://doi.org/10.1016/j.aap.2012.05.018

Wendel-Vos, W., Berg, S. v., & Giesbers, H. (2022). *Cycling in the Netherlands* . Bilthoven: National Institute for Health and Environmnet, Ministry of Health, Weltfare and Sport.

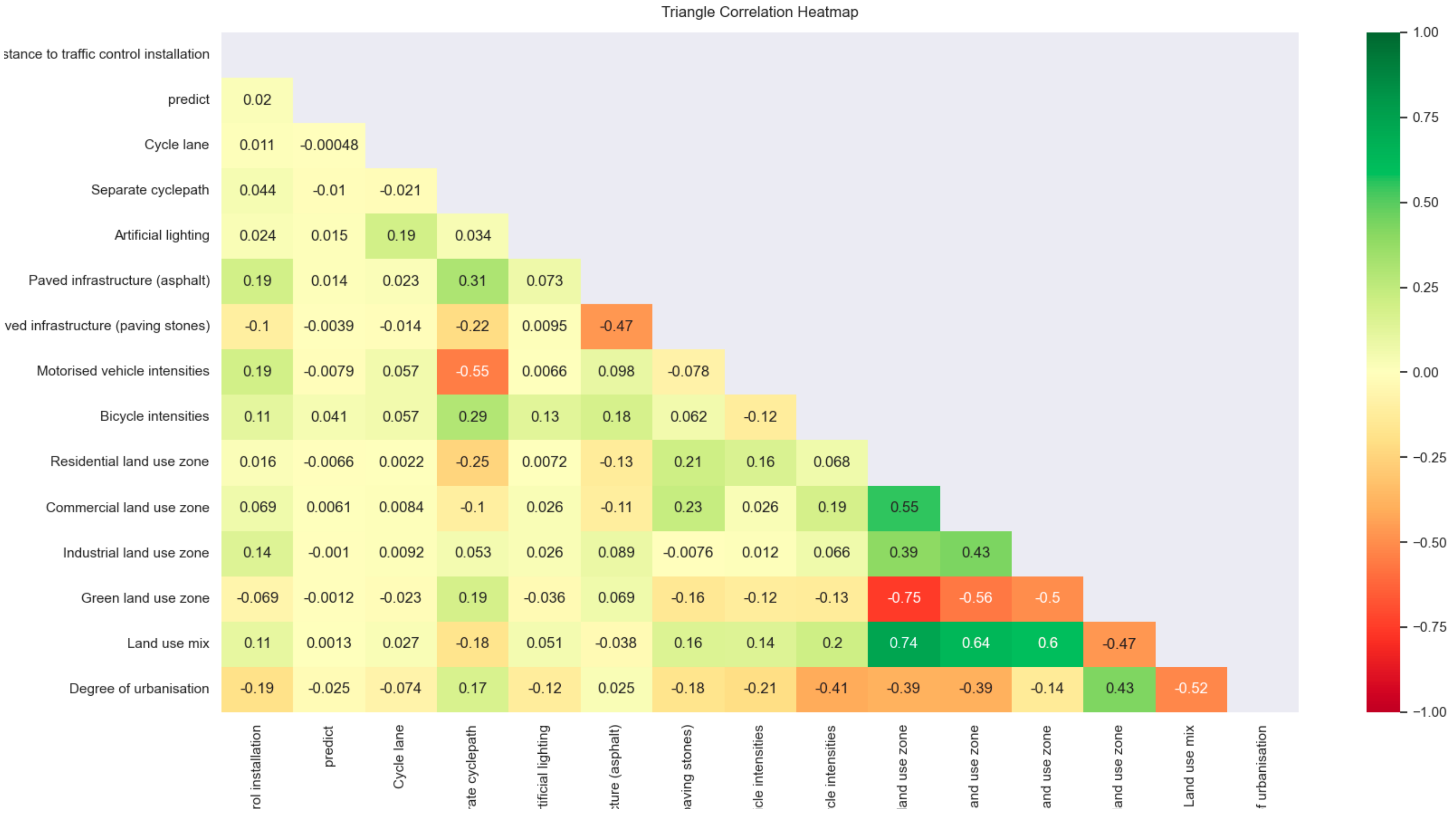
Appendices

Appendix – A

Appendix – A.1. Risk factors correlation heat map

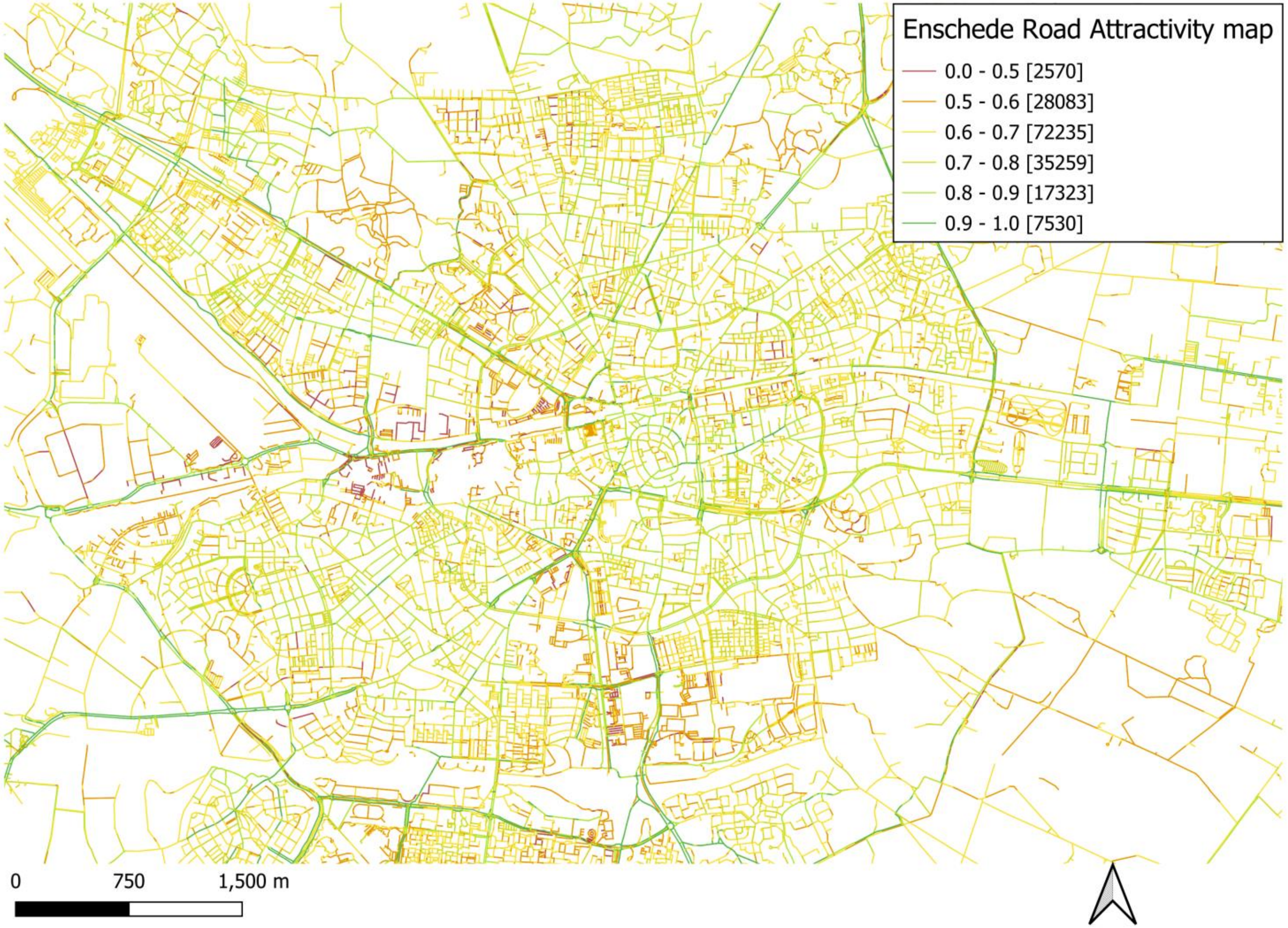


Appendix – A.2. Road safety correlation with environmental factors



Appendix – B

Road Safety is included



Road safety is not included