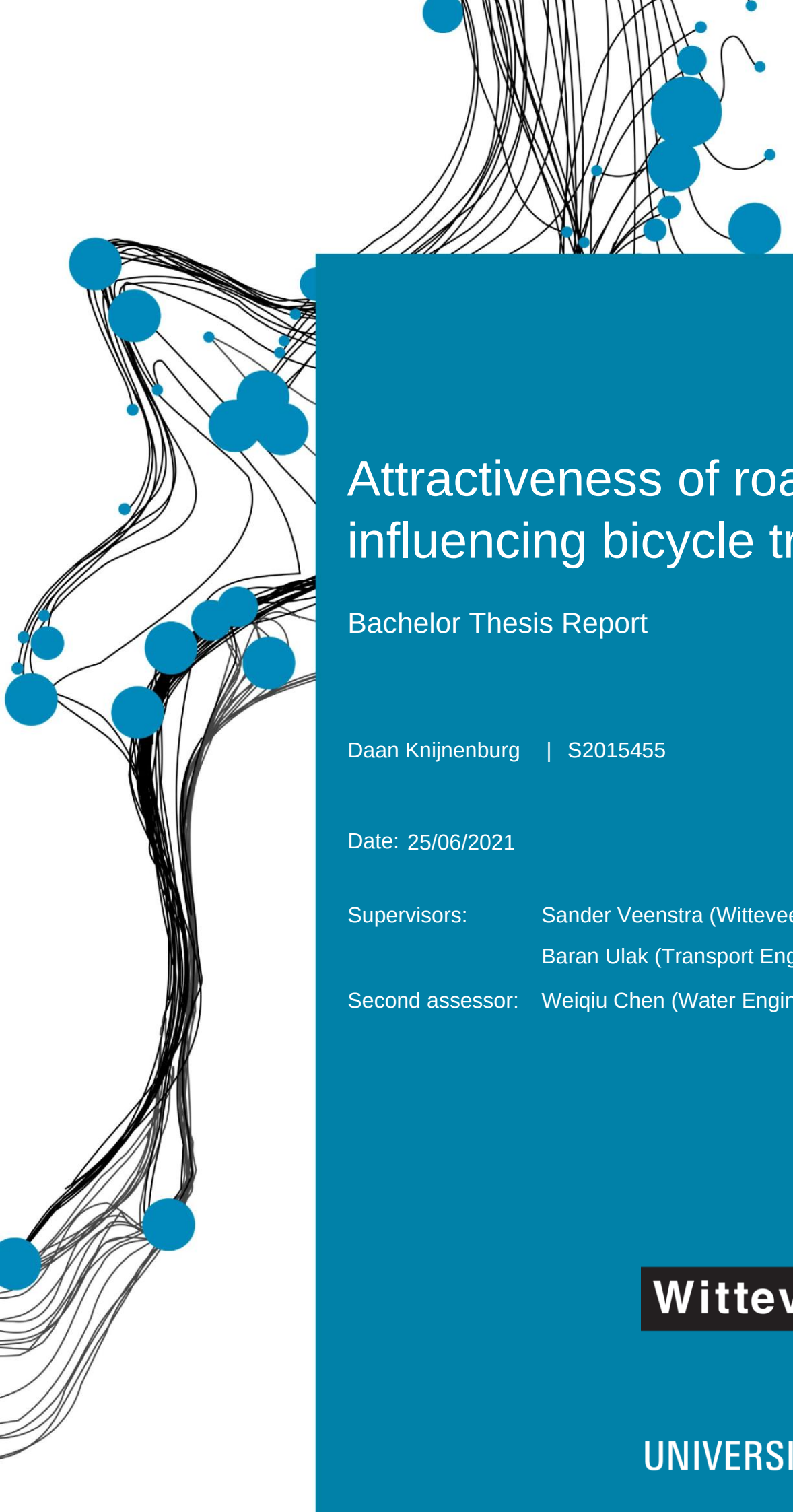

Attractiveness of roads influencing bicycle traffic

Bachelor Thesis Report

Daan Knijnenburg | S2015455

Date: 25/06/2021

Supervisors: Sander Veenstra (Witteveen+Bos)
Baran Ulak (Transport Engineering and Management)
Second assessor: Weiqiu Chen (Water Engineering and Management)

Witteveen Bos

UNIVERSITY OF TWENTE.

Preface

In this report I present the results of my Bachelor thesis for the completion of my Bachelor Civil Engineering at the University of Twente. The topic of this Bachelor thesis is 'Attractiveness of roads influencing bicycle traffic'. In the last decade modelling bicycle traffic has become more important than ever before. To do this appropriately a good insight in the factors that influence bicycle traffic is key. This report contributes to that goal by looking at spatial and demographic characteristics influencing bicycle traffic and seeing how and if these characteristics influence bicycle traffic.

This research was carried out from April to June 2021 and was commissioned by Witteveen+Bos. During these months I have unfortunately mostly worked from home due to the global pandemic of COVID-19. Therefore, I have missed the experience of carrying out a research within a company a bit. Luckily in the last three weeks, the restrictions were lifted a bit in the Netherlands and I was able to work at the office of Witteveen+Bos in Deventer for one day a week.

I would like to thank Witteveen+Bos for giving me the chance to develop myself personally and to work on this interesting project. I would like to thank my supervisor, Sander Veenstra, in particular for guiding me through the process. I have learned a lot during this short period of time which is partly thanks to him. Despite working from two different places, he was always available to help me, which has helped me well every time I ran into some issues or needed anything from him.

I would also like to thank my supervisor from the University of Twente, Baran Ulak, for his help in conducting this research. You were also always available when I needed it and your feedback, enthusiasm and expertise were really helpful in shaping this assignment and writing this final report.

Table of Content

Table of Figures

Table of Tables

Summary

Over a quarter of all the trips made in the Netherlands are made using a bicycle (Kennissinstituut voor Mobiliteit, 2020) and this will keep increasing in the coming years. Therefore, it is important to have infrastructure that can cope with this increasing demand. To make sure that the infrastructure can coop with this new demand it is important to have an insight in the bicycle volumes at every road and this can be done by setting up a prediction model.

Witteveen+Bos has started developing a model to predict the bicycle volumes based on the shortest path between an origin and a destination, but literature has proven that also other factors influence bicycle traffic. In this research the influence of spatial and demographic characteristics is examined. At first characteristics are identified that could influence bicycle traffic. From this list, a selection, based on previous research, is made by selecting the significantly proven characteristics. Nine characteristic are proven to be a significant influence, which are the bicycle infrastructure, residential density, employment density, number and size of shops, greenness, on-street car parking, car speed and land use mix.

In previously conducted research on the influence of spatial characteristics on bicycle traffic multiple modelling approaches have been used. All of them have advantages and disadvantages, but for this research linear and loglinear regression are suited best. In the literature it was also deemed important to use a stepwise regression and to look into the correlation between independent variables properly to prevent multicollinearity.

Eventually the nine characteristics proved to be of a significant influence in the literature and three additional demographic variables, which are address density, percentage of residents between the age of 15 and 24 and the percentage of non western immigrant, are used in this research. With the city of Arnhem as study area together with the city of Apeldoorn, two loglinear regression models were developed. In the first model one characteristic was deemed to be significant on its own and together with three other characteristics a loglinear model was setup to predict the rate between observed counts and modelled bicycle intensities by the Fietsmonitor, a bicycle prediction model of Witteveen+Bos. This model was set up by using the count data of the city of Arnhem and validating the model using the count data of Apeldoorn.

The second model was a model that is setup by taking five different test and train sets of the data (splits) of Apeldoorn and Arnhem and conduct a stepwise regression with all these sets. Two characteristics appeared in all or almost all of the different splits, which were residential density and land use mix. To determine the coefficients k-fold cross validation is used for this model.

Both models had their advantages and disadvantages, but the model that uses the data of Arnhem for setting up the model was eventually chosen as the best model. This model is built for use as an addition to the shortest path method that is currently in use by the Fietsmonitor to predict the number of cyclists.

This research therefore contributes to creating a better understanding of which factors cause people to choose certain routes and travel at certain roads and in how to adjust the Fietsmonitor to better predict the number of cyclists. Further research with more detailed count data and with other cities is advised to get an even better overview of the attractiveness of roads influencing bicycle traffic.

1. Introduction

Cycling is a frequently used mode of transport, 27% of all trips are done by bike as shown in Figure 1, and with the government planning to reduce the use of cars over the coming years, it is expected that bicycle traffic will increase even more in the coming five years (Kennissinstituut voor Mobiliteit, 2020). Therefore, more cities and provinces are adjusting their infrastructure to coop with this increasing demand in bicycle traffic.

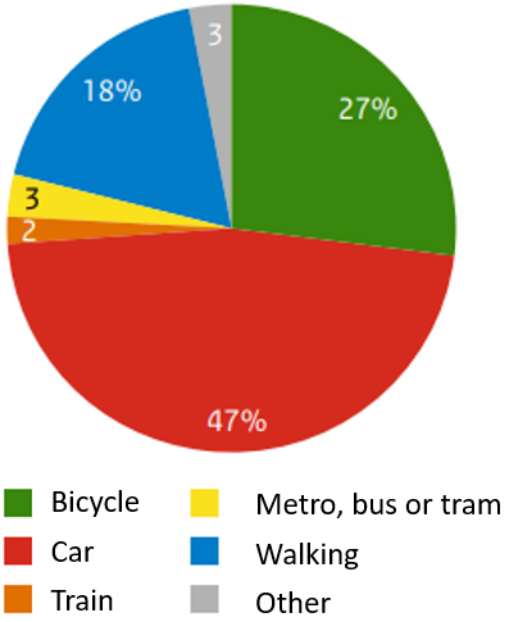

Figure 1 - Distribution of trips by modes of transport in 2016 (Harms & Kansen, 2017)

To adjust the infrastructure accordingly, it is necessary to have an overview of where bicycle traffic can be expected. Therefore, it is important to have accurate models that predict the bicycle traffic. There are currently some models being used, however these are far from optimal yet. Therefore, it is important to keep doing research on this topic, to make the models more accurate.

A model to predict bicycle traffic is build on people moving from point A to point B by using a bicycle as their mode of transport. This sounds easy, but it is rather complex since there are multiple routes that can be taken to reach point B. There are always some routes taken more often than others, but every cyclists has the choice to take their own preferred route. One of the different incentives for taking a specific route is that it is the shortest route (in distance or in time) between the origin and destination. Another could be, because the cyclists is familiar with just one route and therefore always tends to stick to this route, although it could take longer to reach their destination. Thirdly, the attractiveness of the route could influence a cyclists route choice. There are a lot more incentives for why a certain route is chosen, but during this research the main focus will be on the third mentioned incentive, the attractiveness of a route. In this research the understanding of attractiveness is spatial and demographic characteristics improving or diminishing the incentive to take a certain route.

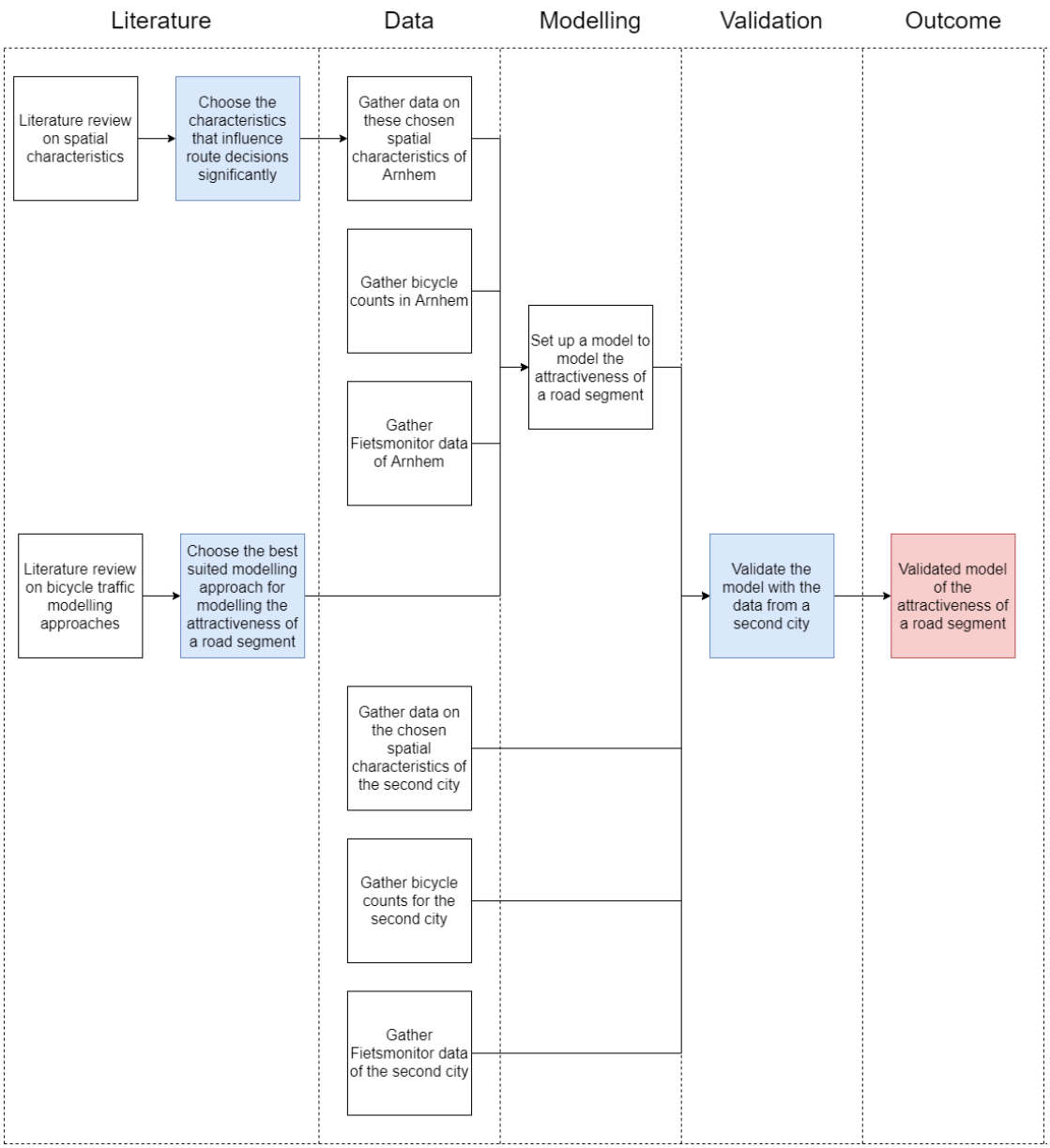

Figure 2 - Schematic overview of the methodology

In Figure 2 a schematic overview of the methodology of this research is given. The structure of this report follows this methodology. In Chapter 2 and 3 this project and research is introduced by describing the project context, the research objective and the research questions. In Chapter 4 the literature on spatial characteristics influencing bicycle traffic and the literature on modelling approaches is described. After the existing literature is described, the report continues in Chapter 5, with the data that is used in this research followed by the methodology of the research, which is described in Chapter 6. In Chapter 7, the results of this research and the validation of the models are discussed. In this validation the model is tested for a second city. Finally in Chapter 8 and 9, the findings in this research are concluded and the interpretations of the results, the shortcomings of this research and the suggestions for further research are mentioned.

2. Project Context

Witteveen+Bos is a consultancy firm that is often asked by municipalities and provinces to consult on projects that concern urban redevelopment with an explicit role for bicyclists and projects that concern the expansion and construction of bicycle highways. For most of these projects it is important to know the amount of cyclists in the area at any moment. Since these intensities differ every day, models are used to get an expectation of the number of cyclists that can be in the area. Witteveen+Bos devoloped a bicycle model that predicts traffic by looking at the origins and destination and the optimal routes between these two points. Cycle and walking trips are also influenced by spatial characteristics in the area (Saelens, Sallis, & Frank, 2016). Therefore, Witteveen+Bos has started making a model to predict the presence of pedestrians by looking at the attractiveness of an area for pedestrians by looking at the spatial characteristics in the area.

This model is called the LoopMonitor and has been set up for the city of Rotterdam, in addition the aim is to make an universal model to predict the pedestrian traffic in other cities as well. The input for this model is a street network, pedestrian counts and multiple thematic layers. An example of these thematic layers were the proximity, connectivity and facilities. These thematic layers have been analysed by a visual interpretation and a geographically weighted regression. A machine learning model was built that is dependent on these input layers. This model gave a good overview of the possibilities and trustworthiness of these types of models and therefore Witteveen+Bos also wants that make such a model for bicycle traffic. Witteveen+Bos is already working on a model, which is called the Fietsmonitor. However, at the moment this model is fully relying on the shortest path (distance) between origins and destinations, but the aim is to eventually also take the attractiveness of the roads into account. When this is also taken into account in the model, this model will be used to give an insight on the bicycle traffic flow and quality of the infrastructure.

These new models will be used to further improve the route choice algorithm that is already being used in the old models and will therefore be useful in predicting bicycle traffic more accurately. With this research I want to contribute to this new bicycle traffic prediction model by creating a better understanding of which factors cause people to make a different bicycle routing choice.

3. Research objective and research questions

The research objective of this research project is to get a better understanding of which factors cause people to make a different bicycle routing choice. This research aims to contribute to this objective by investigating the different route choices of cyclists by looking at the attractiveness of a road segment.

Three research questions follow from the research objective. These questions concern the theory on the influence of spatial characteristics, type of modelling choices that are appropriate for modelling the attractiveness and the validation of the model.

1. Which spatial characteristics influence cyclists to take different routes from the shortest path?

There are multiple spatial characteristics that can influence bicycle traffic taking different routes than the shortest route. However, not all characteristics have a significant influence, therefore it is important to see which characteristics have a significant influence, to eventually predict bicycle traffic. It is not only important to see which characteristics influence bicycle traffic, but also how these characteristics can be measured and implemented in a model predicting the attractiveness, so that will also be researched in this part of the thesis.

2. What type of modelling approaches are appropriate to model the attractiveness of a road segment?

After answering the first research question, it is useful to look into which modelling approaches are appropriate to model the attractiveness of a road segment. It is important to find a model that best takes all the different parameters, the spatial characteristics, into account and will give a clear outcome on the attractiveness of a road segment. This can be done more easily after assessing all the possible approaches.

3. How does the model perform for a different city than Arnhem?

Next to answering the first two research questions and setting up a model to assess the attractiveness of a road segment, it also important to research how well the model is performing for a different city than the one that is used for setting up the variables. This is important to assure the accuracy of the model outcome and therefore it is important to test the different variables of the model in a different city as well.

4. Literature

In this chapter the already conducted researches on spatial characteristics influencing bicycle traffic are described. A list of the characteristics mentioned in these researches, with the corresponding researches in which they are mentioned, is shown in Table 1. Furthermore, the characteristics that were deemed to be of a significant influence on bicycle traffic are bundled in Table 2. Besides the literature on the characteristics influencing bicycle traffic also the literature on modelling the attractiveness of bicycle infrastructure and the surroundings is described in this chapter. This entails researches modelling spatial characteristics, but also modelling approaches that could be used in this research.

4.1 Spatial characteristics influencing bicycle traffic

In this chapter the research on spatial characteristics influencing bicycle trips that is already conducted is reviewed. In multiple researches different characteristics that could influence bicycle trips are discussed. The literature discussed in this chapter did not all focus on the same things and did not always draw the same conclusions. However, based on the literature review a list is made on characteristics that could influence bicycle trips, see Table 1, and then after combining the conclusions of the literature review the characteristics that are deemed significant in most of these researches are chosen to be used in this research.

A lot of studies have already been conducted on the influence of spatial characteristics on route choice of cyclists. First of all, Winters et al. (2010) showed that not always the shortest route is chosen when travelling from an origin to a destination, which you would expect to happen. This research showed that spatial characteristics influenced the route choice of cyclists. These spatial characteristics are objects and properties that define a certain area.

Winters et al. (2010) questioned over 1000 people in Vancouver on their bicycle trips and then looked at the demographic variables and built environment measures along those routes. The built environment measures that were taken into account were grouped in four classes, physical environment, road network, bicycle facilities and land use, with all having multiple subclasses, which are mentioned in Table 1, along with all the other findings on possibly influencing characteristics. From all of these measures and variables not all were significantly associated with a higher likelihood of cycling. The ones that were are less hilliness, higher intersection density (good route connectivity), local roads instead of highways and arterial roads, higher population density and neighbourhood commercial, educational and industrial land uses.

Saelens et al. (2016) mentions that there are a lot of characteristics that could be relevant for bicycle traffic. They mention that connectivity, residential density and land use mix could influence the choice to travel by bicycle. After conducting a correlational analyses of the relations between these characteristics and bicycling, it was concluded that a higher density, commercial and non residential buildings within 100 meter, supermarkets between 100 meter and 1.5 kilometres, access to public transport, employment density, mixed land use, separate bicycle paths and a good accessibility, significantly influence bicycle trips.

Hunt and Abraham (2007) have also done research on the influence of spatial characteristics. Along with some of the already mentioned characteristics, they mention in their research that also the speeds and the volume of motorised vehicles on the roads could influence bicycle trips. After conducting a survey with over a thousand responses, a logit model was set up to review the results. The results of this research show that having separated bicycle paths from the road and having enough secure parking places have a significant positive effect on cyclists.

Griswold et al. (2011) has presented multiple models that can be used to estimate bicycle intersection volumes and in those models they have taken spatial characteristics into account as

well. In their literature research multiple characteristics associated with cycling are mentioned and these are added to Table 1. In this research factors like commercial land use and properties, connectivity, distance to public transport, bicycle infrastructure and slope are modelled. Some of these ways of modelling are not useful for this research since the research of Griswold et al. is only focussed on intersections and this research looks upon all roads, but it shows the influence of characteristics that can be used in this research.

Furthermore, Sener et al. (2009) conducted a survey with 1621 responders as input for their empirical model of bicycle route choice. In their research the focus was on on-street parking, bicycle facilities and road characteristics. This research eventually concluded that on street parking and higher speed limit roads significantly influence bicycle trips negatively. Furthermore, the research concluded that hilliness does not significantly influence bicycle trips, which contradicts the study of Winters et al.

Pucher et al. (2010) has done research towards interventions that can encourage bicycling, which showed some measures that influenced cyclists and are therefore interesting for this research. The measures are almost all related to bicycle infrastructure and Pucher et al. shows that reduced speed limits, bicycle parking, separated bicycle paths increase the amount of cyclists on a specific road or destination.

Heinen et al. (2010) has conducted a literature research on factors that influence cycling behaviour. In this research natural environment, built environment characteristics, psychological factors and socio-economic factors were addressed. For this research, only the first two are interesting. Heinen et al. concluded that higher residential density, separated bicycle paths, no on street parking and less hills increase the number of cyclists in an area or road significantly.

Snizek et al. (2013) conducted a map-based questionnaire and got 4700 respondents. This survey contained routes that were drawn on a map by the respondents and there positive and negative experiences. Eventually, these responses were than modelled by using a multinomial logistic regression model to see what the probability of a certain phenomena occurring as negative or positive response. This research eventually showed that cycling on primary or secondary roads and the presence of intersections effect bicycle trips negatively significantly. Additional this research shows that greenness of the environment negatively correlates with a negative experience and is therefore a significant positive effect on bicycle trips. This research also showed that having separated bicycle paths increased a positive experience. This research also showed that high traffic density has negative effects.

Winters et al. (2011) conducted a survey in Vancouver with just over 1400 respondents. This survey was focussed on items that could influence people to cycle. This could be useful for this research since reason for why people are cycling are also influencing the places where people are cycling. A lot of different factors (73) were taken into account in this research. From these factors only a part was deemed significant and also not all of these significant factors are applicable for this research. To conclude, a flat route, a scenic route, a route with less traffic volumes, low speed limits, paved roads, separated bicycle paths are all preferred significantly by cyclists.

Table 1 - Spatial characteristics that could influence bicycle route choosing

Spatial characteristics	Source
Connectivity	(Saelens, Sallis, & Frank, 2016), Witteveen+Bos, (Winters, Brauer, Setton, & Teschke, 2010) (Griswold, Medury, & Schneider, 2011)
Urban design	(Saelens, Sallis, & Frank, 2016), (Heinen, Wee, & Maat, 2010)
Residential density	(Saelens, Sallis, & Frank, 2016), Witteveen+Bos, (Heinen, Wee, & Maat, 2010), (Winters, Brauer, Setton, & Teschke, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013) (Griswold, Medury, & Schneider, 2011)
Land use mix	(Saelens, Sallis, & Frank, 2016), (Winters, Brauer, Setton, & Teschke, 2010), (Heinen, Wee, & Maat, 2010), (Fraser & Lock, 2011) (Griswold, Medury, & Schneider, 2011)
Bicycle infrastructure	(Hunt & Abraham, 2007), Witteveen+Bos, (Heinen, Wee, & Maat, 2010), (Winters, Brauer, Setton, & Teschke, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013), (Sener, Eluru, & Bhat, 2009), (Winters, Teschke, Grant, Setton, & Brauer, 2010), (Winters, Davidson, Kao, & Teschke, 2011), (Pucher, Dill, & Handy, 2010), (Saelens, Sallis, & Frank, 2016), (Moudon, et al., 2005) (Griswold, Medury, & Schneider, 2011)
Hilliness	(Winters, Brauer, Setton, & Teschke, 2010), (Heinen, Wee, & Maat, 2010), (Winters, Davidson, Kao, & Teschke, 2011), (Sener, Eluru, & Bhat, 2009) (Griswold, Medury, & Schneider, 2011)
Public Transport	Witteveen+Bos, (Hunt & Abraham, 2007), (Winters, Davidson, Kao, & Teschke, 2011), (Rijsman, et al., 2019), (Saelens, Sallis, & Frank, 2016), (Snizek, Nielsen, & Skov-Petersen, 2013)
Nature	Witteveen+Bos, (Snizek, Nielsen, & Skov-Petersen, 2013), (Winters, Davidson, Kao, & Teschke, 2011), (Fraser & Lock, 2011) (Winters, Brauer, Setton, & Teschke, 2010), (Hunt & Abraham, 2007), (Heinen, Wee, & Maat, 2010)
Employment density	(Saelens, Sallis, & Frank, 2016), (Heinen, Wee, & Maat, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013) (Griswold, Medury, & Schneider, 2011)
Facilities	Witteveen+Bos, (Snizek, Nielsen, & Skov-Petersen, 2013), (Winters, Davidson, Kao, & Teschke, 2011), (Moudon, et al., 2005) (Saelens, Sallis, & Frank, 2016)
Bicycle Parking facilities	(Heinen, Wee, & Maat, 2010), (Winters, Brauer, Setton, & Teschke, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013), (Pucher, Dill, & Handy, 2010), (Winters, Davidson, Kao, & Teschke, 2011)
On-street car parking	(Sener, Eluru, & Bhat, 2009), (Hunt & Abraham, 2007), (Heinen, Wee, & Maat, 2010) (Griswold, Medury, & Schneider, 2011)
Air pollution	(Winters, Brauer, Setton, & Teschke, 2010), (Winters, Davidson, Kao, & Teschke, 2011)
Road types	(Winters, Brauer, Setton, & Teschke, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013)
Motor vehicle speeds	(Hunt & Abraham, 2007), (Sener, Eluru, & Bhat, 2009), (Pucher, Dill, & Handy, 2010), (Snizek, Nielsen, & Skov-Petersen, 2013), (Winters, Davidson, Kao, & Teschke, 2011) (Griswold, Medury, & Schneider, 2011)
Motor vehicle volume	(Hunt & Abraham, 2007), (Snizek, Nielsen, & Skov-Petersen, 2013), (Winters, Davidson, Kao, & Teschke, 2011) (Griswold, Medury, & Schneider, 2011)
Stop signs and traffic lights	(Heinen, Wee, & Maat, 2010) (Rietveld & Daniel, 2004)
Demographic variables	(Griswold, Medury, & Schneider, 2011), (Hochmair, Bardin, & Ahmouda, 2019), (Taylor & Mahmassani, 1996), (Strauss & Miranda-Moreno, 2013), (Moudon, et al., 2005)

Moudon et al. (2005) has conducted a survey and has used GIS environmental variables to see which presence of certain social demographic or built environmental characteristics do to the odds of cycling. Since in this research the focus is not on social demographic characteristics, only the built environmental characteristics are interesting. Moudon et al. (2005) concluded that again having bicycle lanes increases the use, but also that small areas with convenience stores are often good environments to cycle around. The research also includes areas with many offices, fast-food

restaurants, hospital and multifamily residential setting, but this was not deemed as a significant factor.

Although some of these studies concluded that traffic volumes have a significant influence, some other studies showed that these factors do not have a significant influence (Hood, Sall, & Charlton, 2011) (Prato, Halldórsdóttir, & Nielsen, 2018). Therefore, this characteristic is not deemed significant in this research.

The different researches show that a lot of different factors influence bicycle trips. Not all show a significant influence and therefore a distinction will be made between the important, read most significant, characteristics and the less important characteristics. After combining the conclusions from the different researches a clear similarity can be seen, which is the characteristic bicycle infrastructure (separated bicycle paths, clearly indicated bicycle lanes or just a small distinction with a dotted line on the road), which is present in all researches. The presence of good bicycle infrastructure has a positive influence on cycling according to the literature.

Besides the road characteristics, a conclusion can be drawn from all these researches that a higher residential and employment density increases the number of cyclists in an area according to most of these studies. Thirdly, a good mix of land use is mentioned as a significant influence in most of the researches. Fourthly, the presence of supermarkets have shown to significantly influence bicycle traffic. Fifthly, water and green areas alongside the road has shown to be an influence and sixthly, on-street car parking has shown to have a negative influence on the attractiveness of a road. Finally, motor vehicle speed is a characteristic that is deemed to have a significant influence on bicycle trips.

4.2 Measuring influential characteristics

After concluding on the characteristics that are going to be implemented in the model in this research, it is important to set how the different characteristics are measured. Some of the previously mentioned researches have already shown how certain characteristics could be measured and this is helpful in determining the measures that will be used.

Literature shows that bicycle infrastructure could be measured by seeing if there is bicycle infrastructure present at the road and if so what type it is. Cyclists prefer separated bicycle paths over bicycle lanes, which they prefer over roads with curb lanes (extra space on the road for cyclists but without road signage) (Taylor & Mahmassani, 1996) (Heinen, Wee, & Maat, 2010).

For residential density and employment density previous researches have used the number of persons per squared kilometre whilst including population density into their bicycle traffic models (Hankey, et al., 2012) (Saelens, Sallis, & Frank, 2016).

Multiple studies have used the same equation to calculate land use mix (LUM) ((Winters, Teschke, Grant, Setton, & Brauer, 2010), (Chen, Zhou, & Sun, 2017), (Winters, Brauer, Setton, & Teschke, 2010) and (Hankey, et al., 2012)). This equation is shown in Equation 1. This is a diversity index called the Shannon index. This index is used often to calculated land use mixture since it takes both the number of different land uses into account and their relative abundances.

Other studies have opted to use another entropy index, which is shown in Equation 2 (Strauss & Miranda-Moreno, 2013). This second equation is almost the same as the first equation, but it excludes water bodies and open spaces in the buffer.

$$LUM = -\Sigma_k[(p_i)\ ln\ (p_i)]/ln\ (k) \qquad \text{Equation 1}$$

$$LUM = -\Sigma_{i=1}^{k} \frac{\frac{A_{ij}}{D_j} ln \frac{A_{ij}}{D_j}}{\ln (k)}$$

Equation 2

Within Equation 1 the p_i is the proportion of each of the land uses classes and k is the number of those classes present. Within Equation 2 the A_{ij} is the area of land use i in buffer j, D_j is the area of buffer j excluding open space and water body and k is the same as in Equation 1. LUM eventually ranges from 0 to 1, with 0 being homogeneous land use and 1 the most mixed land use.

To calculate the land use mix surrounding a road segment a buffer needs to be setup around the road segment. The studies that used land use mix used different buffers ranging from 50 meters to a 1 mile (approximately 1.6 kilometres) buffer (Zhao, Lin, Ke, & Yu, 2020) (Chen, Zhou, & Sun, 2017) (Winters, Brauer, Setton, & Teschke, 2010) (Strauss & Miranda-Moreno, 2013). Since the bicycle network that is used in this research is detailed and to make a clear distinction in the land use mix between different road segments near each other, a small buffer of 100 meters is chosen. Furthermore, Equation 1 will be used in this research, since water bodies are linked to recreational land use in this research and will therefore not be excluded in the land use mix.

For supermarkets Moudon et al. (2005) and Saelens et al. (2016) concluded it to have a significant influence on bicycle traffic. Moudon et al. (2005) measured the association with cycling by looking at the parcel area of the supermarkets and seeing a negative impact on cycling when the parcel area became larger. Saelens et al. (2016) looked at supermarkets within a certain distance of the road and saw a positive effect. These two ways of measuring are different from one another, but can both be used in this research. Therefore, this research will use the presence of a supermarket within 250 meters, same buffer size as Saelens et al. (2016) used, as a positive influence and the size of that same supermarket parcel as a negative influence.

The greenness of an area is measured in most of the researches by calculating the percentage of land area with green cover in a certain buffer around the road (Winters, Teschke, Grant, Setton, & Brauer, 2010) (Snizek, Nielsen, & Skov-Petersen, 2013) (Winters, Brauer, Setton, & Teschke, 2010). The buffer sizes of these percentages vary from 50 meters to 250 meters. Since some researches also include distance to nearest water body as a factor (Hankey, et al., 2012) (Snizek, Nielsen, & Skov-Petersen, 2013), this will be included in the greenness of the area. A 250 meters buffer size is chosen to measure this characteristic based on the research of Winters et al. (2010).

The final characteristic that is measured is on-street car parking. This characteristic can be measured by looking at if parallel parking on the street is permitted or not (Sener, Eluru, & Bhat, 2009).

The significantly influencing characteristics are shown in Table 2. In this table also the ways of measuring the characteristics and how they influence bicycle traffic is mentioned.

Table 2 - Influential characteristics, how can they be measured and the type of influence on bicycle traffic

Influential characteristics	**How is the characteristic measured**	**Influence on bicycle traffic**
Bicycle infrastructure	Four categories (0. No bicycle infrastructure, 1. Bicycle curb lanes, 2. Bicycle lanes, 3. Bicycle paths)	Positive
Residential density	Persons/km^2	Positive
Employment density	Persons/km^2	Positive
Land use mix	0 to 1 calculated with Equation 1 in a buffer of 100 meters	Positive
Supermarkets	Number of supermarkets within 250 meters	Positive
	Size of supermarket parcel	Negative
Greenness	Percentage of land area with green cover in buffer of 250 meters around the road	Positive
On-street car parking	0 or 1 (parking not permitted or permitted)	Negative
Motor vehicle speeds	Km/h	Negative

4.3 Modelling approaches

To see the importance of these significant characteristics a regression model is used. There are several regression models that could be useful in modelling bicycle counts and the attractiveness of a road segment. In this part of the report the different regression models and their advantages and disadvantages for this research are described.

First of all, the simplest form of regression, which is multiple linear regression. The equation for multiple linear regression is shown in Equation 3. In this equation, the y is the variable that one want to predict, the X are the independent variables, the β are the regression coefficients and ε the error term. This form of regression model is used when there is a linear relationship between independent and dependent variables. An advantage of using this type of linear regression is that it can easily involve multiple independent variables, which is useful when modelling different characteristics. A weak point of this type of regression models is that it is sensitive to outliers.

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \qquad \text{Equation 3}$$

Another common form of regression is a polynomial regression. The equation for polynomial regression is shown in Equation 4. This type of regression models is used when a relation is not linear. However, a disadvantage of using this type of regression models is that it can lead to overfitting more by adding more quadratic terms.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon \qquad \text{Equation 4}$$

A third regression method that is used often in modelling bicycle intensities is log linear ordinary least squares regression, in short loglinear regression. In Equation 5, the equation for this type of regression is shown. This is used often for counts since there are no negative values. The ln is used in this equation to transform the dependent variable to make the data less skewed. Besides this transformation, the entire regression is similar to linear regression.

$$\ln(y) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k \qquad \text{Equation 5}$$

Fourthly, Poisson regression is often used when dealing with count data. The Poisson regression equation is shown in Equation 6 (Kim & Susilo, 2011). Where μ is the expected value of the dependent variable. This expected value can be calculated with Equation 3, by having the μ instead of the y.

$$f(y_i|x_i, \beta) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad \text{Equation 6}$$

An advantage of this type of regression modelling is that the estimates are adapted to the actual data, which gives a small error, and that the change of over or underfitting is very low when using this type of regression model. A disadvantage of using Poisson regression can be that the assumption of the skewness of a Poisson model are mostly lower than the skewness of the actual distributions (Lord, Park, & Levine, 2013).

A fifth type of regression model that is used in modelling spatial characteristics is the negative binomial regression. This type of regression model is similar to the Poisson regression, but it does not assume the variance to be equal to the mean (Meyer, sd).

A sixth type of regression that is used in using built environmental and demographic characteristics in modelling route choice and bicycle traffic is logistic regression. This type of regression is used when the variables that are modelled are binary.

4.4 Use of regression models with spatial characteristics in literature

Multiple studies were conducted that used regression models involving built environmental or sociodemographic variables. All of the abovementioned regression types appear in these researches. The Poisson regression and negative binomial regression are used in studies of Wang et al. (2014), Fagnant & Kockelman (2016), Chen et al. (2017), Strauss & Miranda-Moreno (2013) and Hankey et al. (2012). These studies focussed on modelling bicycle counts and that is especially where these types of regression methods are best suited for. Since bicycle counts change according to hours, days and seasons and Poisson and negative binomial models are capable of identifying correlations and heterogeneity over time (Chen, Zhou, & Sun, 2017). In most of these studies weather variables and travel demand patterns are included, but in this research these variables are not included.

Logistic regression is also used by multiple studies in modelling the involvement of built environment in bicycle route choice and bicycle traffic in general. In the studies of Winters et al. (2010), Snizek et al. (2013) and McBain & Caulfield (2017) logistic regression is used. In these studies logistics regression is used for different purposes. Winters et al. (2010) and McBain & Caulfield (2017) used it to see what the likelihood was of detouring because of a certain characteristic, which is not the aim of this research. Snizek et al. (2013) used logistic regression to explain the probability of positive or negative experiences against the other experiences of a cycle route with a certain characteristic, this is not what will be done in this research and therefore logistic regression will not be used in this form.

Also loglinear regression is frequently used in literature on using built envormint to model bicycle route choice. In the studies of Griswold et al. (2011) and Strauss & Miranda-Moreno (2013) loglinear regression is used. Griswold et al. (2011) and Strauss & Miranda-Moreno (2013) used it to examine the relationships between an intersection volume and built environment surrounding the intersection, which is similar to what will be used in this research and therefore this way of using loglinear regression is useful for this research.

Finally also linear regression is used in some relevant studies, which are Hochmair et al. (2019), Hankey & Lindsey (2016) and Zhao et al. (2020). Hochmair et al. (2019) used linear regression

models in determining bicycle volumes on Strava BKT by using built environment and sociodemographic variables. Hankey & Lindsey (2016) used stepwise linear regression to model bicycle volumes in an area by using spatial characteristics as independent variables. In this study the use of stepwise linear regression was concluded to be very important, since independent variables could be selected at multiple scales and it only included statistically significant factors that are not correlated with each other.

The last type of regression that was mentioned that it could be of use in this research was the polynomial regression. This type of regression is not used often in previous researches. A study of Holmgren et al. (2018) could be found in which a polynomial regression is used, which showed that polynomial regression is useful for a trend curve, but that is not the aim of this research.

From all of these researches it can be concluded that linear regression and loglinear regression (a linear regression, but with a transformation of the bicycle counts) are the best modelling approaches for this research. The use in previous researches of these types corresponds the most with the objective of this research. In almost all of the researches that used Poisson and negative binomial models, weather and travel demand patterns are taken into account, but as this is not the case in this study, these regression types can overcomplicate the model and therefore these regression types will not be used.

5. Data

In order to predict the number of cyclists at a road segment by looking at certain built environment and demographic characteristics, it is important to define a study area of which data is collected. With this data a machine learning model can be setup. In the literature review, Chapter 4, multiple characteristics that could influence bicycle traffic were discussed and also significant characteristics were mentioned. In this chapter the study area and also the characteristics that are taken into account in this research are described. Not only are the characteristics that are used mentioned in this chapter, but also how the data is gathered is discussed.

5.1 Study area

For this research project, the city of Arnhem, which is shown in Figure 3, will be the study area. This city is chosen as the study area, because the province of Gelderland, in which Arnhem is located, has conducted a sufficient amount of bicycle traffic counts. Furthermore, Arnhem is a big city in the Netherlands with a lot of bicycle traffic, this can also be seen in the ambition of the province, that wants to be the bicycle province of the Netherlands (Provincie Gelderland, sd). Besides this ambition and the traffic counts, Arnhem also is a city with a lot of places of interest and nature, which attracts a lot of cyclists.

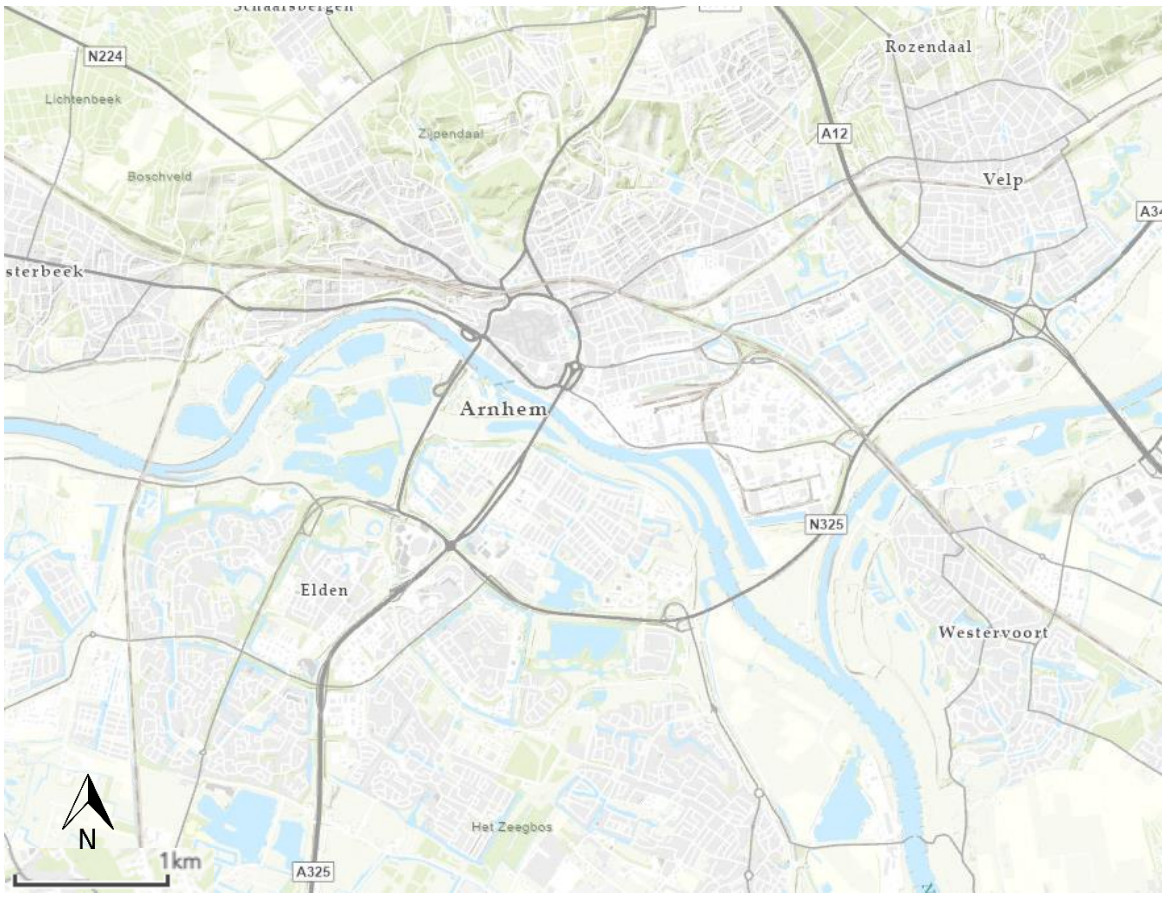

Figure 3 - The city of Arnhem, which is the study area (Provincie Gelderland, 2020)

5.2 Dependent variable

The eventual aim of this research is to contribute to a better bicycle traffic prediction model by creating a better understanding of which factors cause people to make a different bicycle routing choice. In achieving this aim it is important to use the already existing bicycle traffic prediction model of Witteveen+Bos. Besides the use of this existing model being useful to achieve the research objective, it is also useful in seeing the relation between the origin and destination traffic (as predicted in the model of Witteveen+Bos) and observed bicycle counts. Therefore, the dependent variable of this research is the rate between the observed counts at a certain point in the city and the predicted number of cyclists by the prediction model of Witteveen+Bos. This model is called the Fietsmonitor and is a model that uses shortest path between origins and destinations to predict the presence of cyclists.

The rate is transformed by using the natural logarithm to linearize the relationship with the independent characteristics, which is non-linear without transformation.

For the observed bicycle counts multiple sources can be used, for example data from the Fietstelweek or from count stations. In this research both of these examples have bee proposed as observed counts. Fietstelweek was proposed as a source, since this data source gives data per road segment, which is similar to what the Fietsmonitor predicts and therefore could give more data points to observe than the counts stations. However, since the Fietstelweek data is provided by mostly recreational cyclists that make trips that do not represent the trips made by the common cyclists, which is a way larger group of cyclists, it was decided to use counts stations as the source for the observed counts. These count stations are from the province of Gelderland and the municipality of Arnhem (Provincie Gelderland, 2020). The count points used in this research are shown in Figure 4.

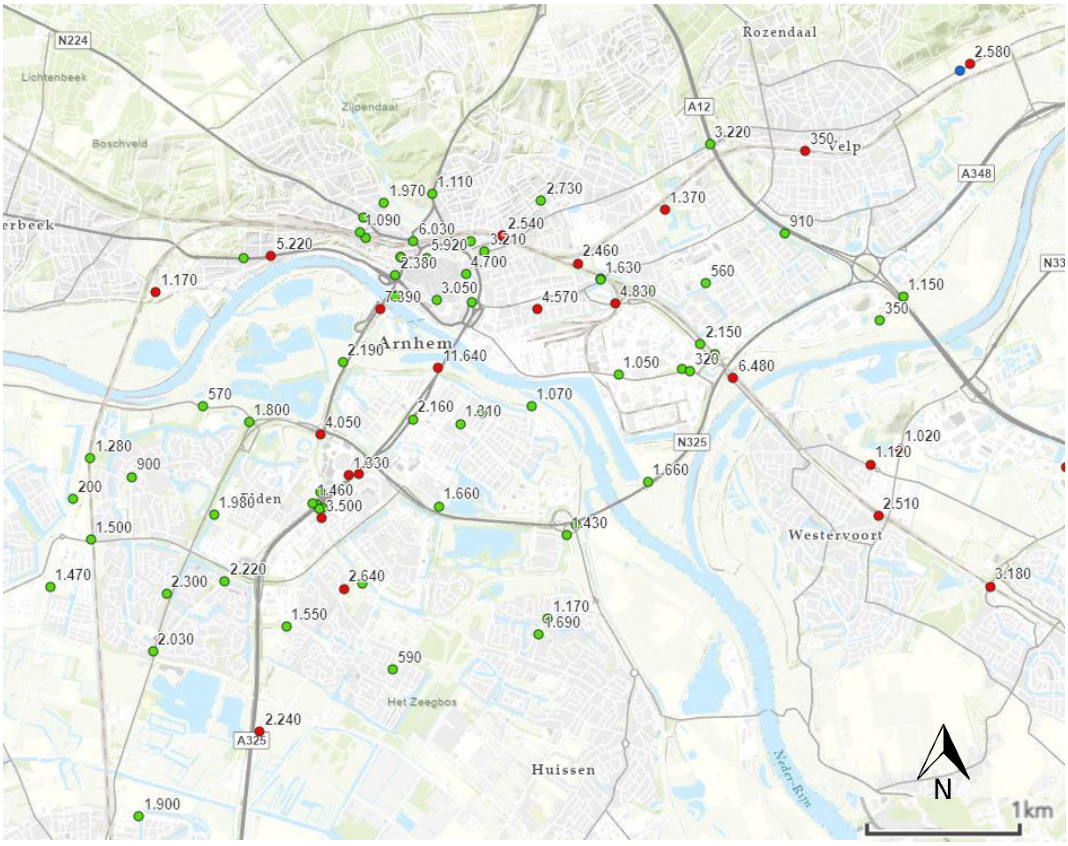

Figure 4 - Bicycle count points in Arnhem (Provincie Gelderland, 2020)

The counts of the 'snelle fietsroutes' in 2020 (shown with the red dots in Figure 4), which were conducted by an external company commissioned by the province of Gelderland, counts of the municipality of Arnhem in 2020 (shown with the green dots in Figure 4) and counts of the province of Gelderland itself in 2020 (shown with the blue dots in Figure 4) have been used. The counts of the 'snelle fietsroutes' were mechanical counts that were conducted for three weeks in a row. The counts of the municipality of Arnhem were counted for a month between October and November. The counts of the province of Gelderland were again counted for a month, but this time in June. For all of these counts the averaged weekday averages have been used.

All these counts have been used as points on the map. Furthermore, some of the counts of the 'snelle fietsroutes' have been combined since these counts were along the same road, but the two directions had separated bicycle paths on two different sides of the road.

5.3 Identifying spatial characteristics of count stations

After defining the study area, which is Arnhem, the processing of the data is described. In this research Open Street Map, BGT data, BAG data, CBS data (CBS, 2021) and Google Maps are used to process the data. In this research all the characteristics that are deemed to be significant in literature are taken into account and additionally some demographic characteristics are taken into account as well. In Table 2, the ideal way of measuring the significant characteristics is described. However the data did not always allow the characteristics to be measured in that way. Therefore some of the ways of measuring have been altered to still take that characteristic into account.

5.3.1 Bicycle infrastructure

Bicycle infrastructure was deemed to be of a significant influence in the literature review and will be measured by using four different infrastructure categories being no infrastructure, curb lanes, bicycle lanes and bicycle paths. The data of this independent variable is processed by using Google Maps. For every count point, Google Maps is used to assign the corresponding infrastructure type manually as seen on street view and on the satellite pictures on Google Maps.

5.3.2 Residential and employment density

The residential density and employment density are two independent variables that will be measured similarly. Both will be processed by using BAG data. In Table 2 it was mentioned that in literature these two variables are often measured in persons per squared kilometres, but in this research it will be measured by looking at the number of residences (for residential density) and the number of addresses with the function industry or services (for employment density) in a based on an assumption radius of 250 meters around the count point. This is different from Table 2 since the data did not allow it to be measured in persons per squared kilometres. This radius is chosen to have a good distinction in the values of these characteristics between different count points.

The values for residential density have been standardised by dividing the number of residences around a count point by the maximum number of residences at one count point, which was 1540 residences.

The values for employment density were standardised by dividing the number of addresses with the function industry or services by 100.

5.3.3 Land use mix

For the land use mix five different land uses are defined which are industry, services, residences, shops and nature (which includes green area and water bodies). For this independent variable Equation 1 is used to calculate the land use mix by using the area proportions of the different land

use classes as stated in BGT and BAG data within a 100 meter radius of the bicycle count point. The radius of 100 meter is a small buffer compared to other researches like Zhao et al. (2020), Chen et al. (2017) and Winters et al. (2010), however since a detailed bicycle network is used and a distinction in values for this characteristic between count points is desired, this buffer is chosen. This buffer size is similar to a buffer size used by Strauss & Miranda-Moreno (2013).

5.3.4 Supermarkets

The presence of supermarkets is also an independent variable in this research, but it has two different influences and ways of measuring the variable. At first the number of supermarkets will be measured by looking at the number of addresses with the BAG function shops within a 250 meter radius. The second way of measuring the influence of supermarkets on bicycle traffic is by the size of the supermarket parcel. For this the sum of the area of the addresses with the BAG function shops within a 250 meter radius is taken to take this variable into account. The buffer sizes of these characteristics are based on the research of Saelens et al. (2016).

The number of shops has been standardised by dividing the number of shops by 100. Furthermore, the size of the supermarket parcel has also been standardized by dividing the are by 10000.

5.3.5 Greenness, on-street car parking and motor vehicle speeds

The greenness data will be gathered by calculating the percentage of the total area that is overgrown or have a water body in the BGT data within a 250 meter radius around the bicycle count point. This radius is based on the research of Winters et al. (2010).

The independent variable on-street car parking will be measured by looking at the presence of parking spaces in BGT data within a 25 meter radius around the bicycle count point. When a parking space is present the value 1 will be given and otherwise the value 0 is given.

The motor vehicle speeds of the roads near the bicycle count point are measured using data of Open Street Map. The data is gathered by looking at the maximum speed that is allowed on the roads according to Open Street Map within a 25 meter radius around the bicycle count point. The motor vehicle speeds have been standardised by dividing the maximum speeds by the maximum speed allowed at a count point, which was 100 kilometres per hour.

5.3.6 Demographic variables

Three characteristics are measured by using CBS data. These characteristics are the percentage of non western immigrants, address density and the percentage of residents between 15 and 24 years old. For all three characteristics the average of the values as mentioned by ‘CBS wijken en buurten’ in a radius of 50 meters around the count point is taken as the corresponding value for the characteristic.

The address density has been standardised by dividing the density by 4000 for every count point.

6. Methodology

With the data, as described in Chapter 5, a regression model can be setup. In the literature review, conducted in Chapter 4, multiple characteristics and modelling approaches have been discussed. In this chapter the final modelling approach is described.

After gathering and processing this data, the data is used in a multiple loglinear regression model. The equation for a multiple loglinear regression is shown in Equation 7.

$$\ln(y) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k \qquad \text{Equation 7}$$

The aim of this loglinear regression model is to find the independent variables (the spatial and demographic characteristics) that have a significant influence on the dependent variable (rate between observed counts and Fietsmonitor predicted number of cyclists) and after finding these variables to calculate coefficients to eventually have an equation that predicts bicycle traffic. For finding the significantly influencing variables, a stepwise regression method is used. This method processes each variable and removes the variables that are not significantly influencing bicycle traffic.

6.1 Collinearity

Before doing the stepwise regression it is important to first determine if there is collinearity between the independent variables. If this is the case, the stepwise regression will not work as well as it should, so therefore if there is collinearity between independent variables, one or more independent variables need to be dropped from the stepwise regression. To determine collinearity the Pearson correlation coefficient is calculated for the relation between every independent variable. The Pearson coefficient is the covariance of both variables divided by the multiplication of the standard deviation of both variables. This coefficient is frequently used in statistical analysis and has a result of a value between -1 and 1. Figure 5 shows the meaning of the different correlation values. In this research when a correlation is strongly negative or positive, so either a correlation coefficient of -0.8 or 0.8, one of the two correlated independent will be dropped. When there is collinearity between independent variables the correlation with the dependent variable will be calculated and the independent variable with the worst correlation will be dropped.

Correlation Coefficient Value (r)	Direction and Strength of Correlation
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

Figure 5 - Interpretation Pearson correlation coefficient (Ratnasari, Nazir, Toresano, & Pawiro, 2016)

6.2 Stepwise regression

After determining if there is collinearity between the independent variables, the variables that are left are used in the stepwise regression. In this stepwise regression the significance of every variable is first tested in a loglinear regression model with just the one variable on its own. The significance of a variable is tested by calculating the P-value. A P-value can range between 0 and 1. Stepwise regression is a combination of forward selection and backward elimination and therefore uses two thresholds in determining variables that need to be added or removed. In this research a 90% confidence interval is used for entering the model and a 85% confidence interval for being excluded from the model. These threshold values differ to stabilise the choosing of the characteristics entering or leaving the model. In the forward selection part the variable with the best P-value is added to the model if the P-value is lower than the threshold that has been set to enter the model, so lower than 0.1.

After the forward selection, the backward elimination starts, which calculates the P-values again for added variables from the forward selection. The P-values change for every variable once more variables are added to the model and therefore some combination of variables in the model can make individual independent variables become a less significant influence on the dependent variable. When a variable has a P-value in the backward elimination that is lower than the threshold, which is a P-value that is higher than 0.15, it will be excluded from the model.

This combination of multiple loglinear regression approaches ensures that the eventual model has the combination of independent variables included that has the lowest error sum of squares. This way of modelling also has a small chance of overfitting.

7. Results

In this chapter a descriptive overview of the count data and variables and the results of the multiple stepwise regressions and collinearity tests are discussed. Two different stepwise regressions are conducted. The first being a stepwise regression that only uses the data of Arnhem to setup the model and that uses the data of Apeldoorn as data to validate the model. The second is a loglinear model that uses a dataset which contains the count points in Apeldoorn and in Arnhem and is validated by dividing the data into a test and train set randomly.

7.1 Descriptive statistics of the count data and the variables

In this section a descriptive overview of the count data and the variables is given. Also the reason of transforming the dependent variable is discussed.

7.1.1 Overview of the observed counts

In this research the observed counts at count stations of the province of Gelderland and the municipalities of Arnhem and Apeldoorn have been used together with the expected counts at these locations of the Fietsmonitor as dependent variable. In Figure 6 a histogram is shown of the observed counts. From this figure it can be seen that the most counts have counted less than 2100 cyclists at there count stations. This histogram also shows that the counts are right skewed.

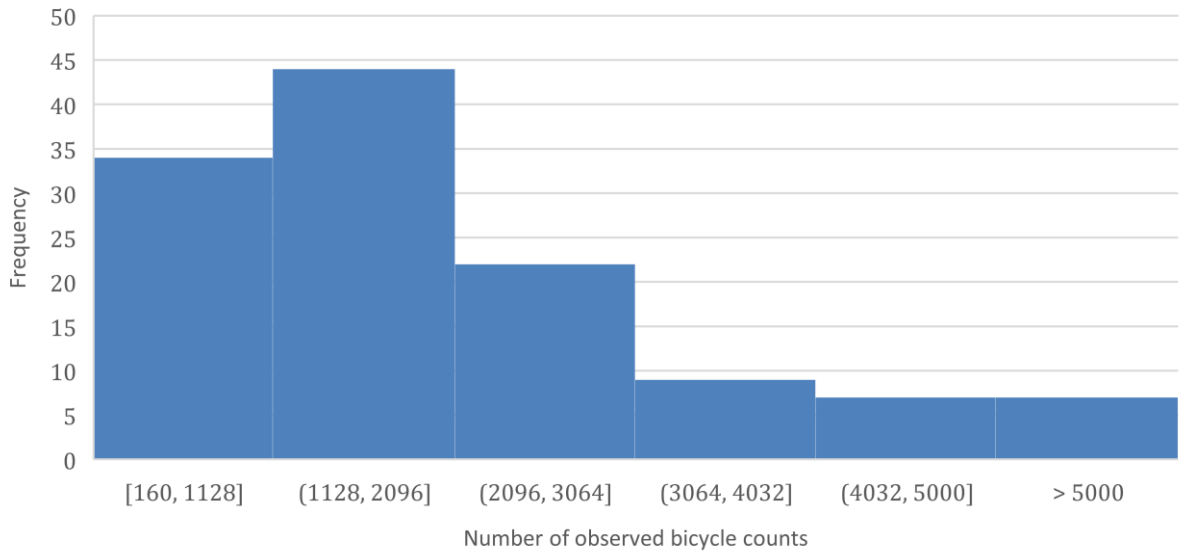

Figure 6 - Histogram of the observed counts

7.1.2 Overview of the variable and count data

During this research 123 count points have been used and at all these count points values have been retrieved for 12 different spatial or demographic variables. In Table 3 an overview is given of the summarized statistics for every (in)dependent variable and the observed and expected bicycle counts.

Table 3 - Summarized statistics of the variables and count data

	Mean	Median	Standard deviation	1st Quartile (25%)	3rd Quartile (75%)
Presence Bicycle Curb Lane	0.065	0	0.248	0	0
Presence Bicycle Lane	0.057	0	0.233	0	0
Presence Bicycle Path	0.732	1	0.445	0	1
Residential density	0.251	0.170	0.266	0.020	0.388
Employment density	0.176	0.050	0.249	0	0.260
Number of shops	0.191	0.040	0.397	0	0.190
Size of shop parcel	8825	1909	17859	0	9544
Greenness	0.319	0.280	0.219	0.169	0.441
On-street car parking	0.268	0	0.445	0	1
Maximum car speed	0.442	0.500	0.222	0.300	0.500
Land use mix	0.361	0.372	0.164	0.228	0.451
Percentage not western immigrants	13.392	10.500	10.892	4.750	20.583
Address density	0.433	0.417	0.244	0.284	0.555
Percentage of people between 15 and 24 years old	13.70	12.50	4.08	11	16.25
Observed bicycle count	2089	1730	1665	1050	2465
Expected bicycle count by the Fietsmonitor	1455	995	1275	445	2073
Rate	2.244	1.578	2.377	1.069	2.417
Transformed rate	0.499	0.456	0.743	0.067	0.883

7.1.3 Relation between independent and dependent variable

In Appendix A: Comparison scatterplots with and without transforming the dependent variable multiple scatterplots are given of a single independent variable and the dependent variable. Two scatterplots are displayed of every independent variable, a scatterplot with the rate between the observed bicycle counts and the expected number of cyclists by the Fietsmonitor as dependent variable and a scatterplot with the transformed rate. From these plot it can be seen that after transforming the rate the dots get more spaced out in the plot. Also a more linear relation becomes visible for independent variables like residential density and land use mix. These plots therefore show that the transformation of the rates is necessary to linearize the relation with the independent variables.

7.2 Loglinear model trained with Arnhem data

The first model that is discussed in this chapter is the loglinear model that is made using the data of Arnhem. After collecting the data, the correlation test and stepwise regression can be done. At first 92 bicycle count points in Arnhem were taken into account and the corresponding data was collected. After analysing this data 8 count points were removed, since no expected bicycle traffic could be generated by the Fietsmonitor for these points and therefore no ratio could be calculated for these points. Furthermore, three outliers were removed. This leaves 81 count points that are taken into account in this regression.

7.2.1 Correlation test

At first a correlation test is conducted between the independent variables. The correlation between the different independent variables is shown in Figure 7. As mentioned in Chapter Methodology6, if the Pearson correlation factor exceeds 0.8 or -0.8 one of the independent variables needed to be removed before the stepwise regression. Since this is the case for the size of the supermarket parcel and the number of supermarkets, one of these needed to be removed. This correlation can be caused by how the data is gathered for both characteristics. Both characteristics are gathered by looking at the BAG function shops and is logical that if more shops are present around a count station, the parcel area is higher as well.

Figure 7 - Pearson correlation heatmap of the independent variables of the model trained with data of Arnhem

To determine which of those variables needed to be removed the correlation between the independent variables and the dependent variable is calculated. The Pearson correlation values of these relations are shown in Appendix B: Correlation between dependent and independent variables Arnhem trained model. Since the correlation value of the size of the supermarket parcel is smaller than the value of the number of supermarkets, the size of supermarkets is dropped from the regression. Therefore, 13 variables have been taken into account in the stepwise regression.

7.2.2 Stepwise regression

In Table 4 the results of the regression analysis are shown. From these results it can be seen that only residential density is significant on a 90% Confidence Interval. The dependent variable used in this loglinear regression is the natural logarithm of the rate between the observed count at a count station and the predicted number of cyclists at that location by the Fietsmonitor.

Table 4 - Loglinear regression analysis results of model trained with data from Arnhem

Characteristic	β	SE	p-value
Constant	0.969	0.572	0.095
Presence of bicycle curb lane	0.587	0.407	0.154
Presence of bicycle lane	0.110	0.513	0.831
Presence of bicycle path	-0.008	0.293	0.977
Residential density	1.583	0.844	0.065
Employment density	-0.598	0.577	0.304
Number of supermarkets	0.017	0.475	0.971
Greenness	0.263	0.582	0.653
On-street parking	-0.086	0.265	0.746
Maximum car speed	0.027	0.387	0.945
Land use mix	-1.414	0.862	0.106
Percentage non western immigrants	0.003	0.010	0.796
Address density	-0.092	0.751	0.903
Percentage residents between 15 and 24 years old	-0.022	0.028	0.437

However, after including the residential density in the loglinear regression and continuing the stepwise regression the significance of three other characteristics is lower than the threshold of 0.1 and therefore also land use mix, employment density and the presence of bicycle curb lanes are added to the loglinear regression model. After including the residential density, the p-value of the other three characteristics lowers significantly. The regression model with the corresponding coefficients is shown in Equation 8.

$$\ln(rate) = 0.8588 + 1.3422x_1 - 1.5108\,x_2 - 0.8633x_3 + 0.6350x_4 \qquad \text{Equation 8}$$

In this equation x_1 is the residential density, x_2 is the land use mix, x_3 is the employment density and x_4 is the presence of bicycle curb lanes. This equation has a R-squared value of 0.208.

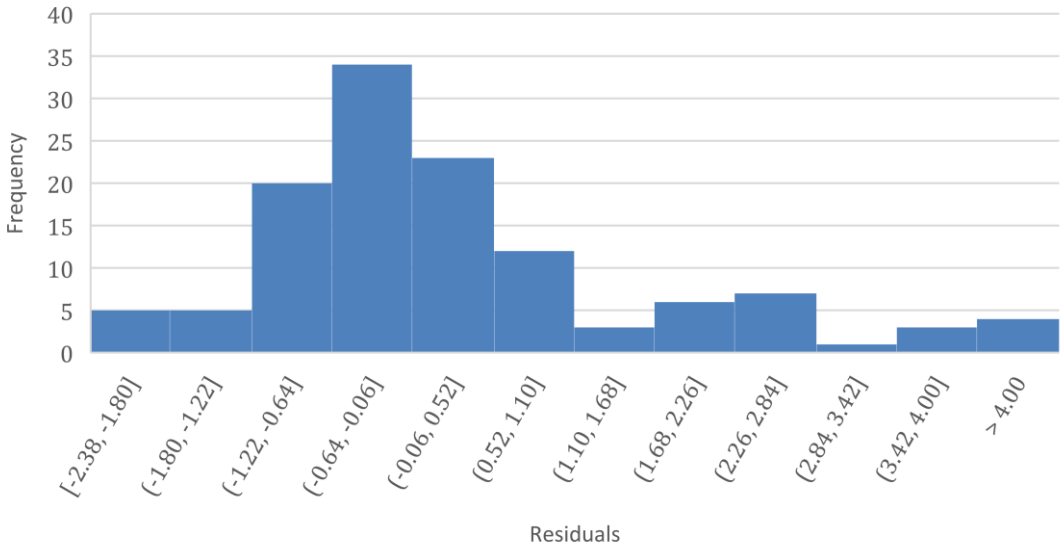

Figure 8 - Histogram of the residuals of model that is trained with Arnhem data

The residuals of this model are shown in a histogram in Figure 8 and the Q-Q plot of this model is shown in Figure 9, with the trendline shown in red. From the histogram it could be concluded that the variance is normally distributed since the histogram is symmetrically bell-shaped around zero, however the Q-Q plot is not that conclusive, since this shows a more exponential trendline.

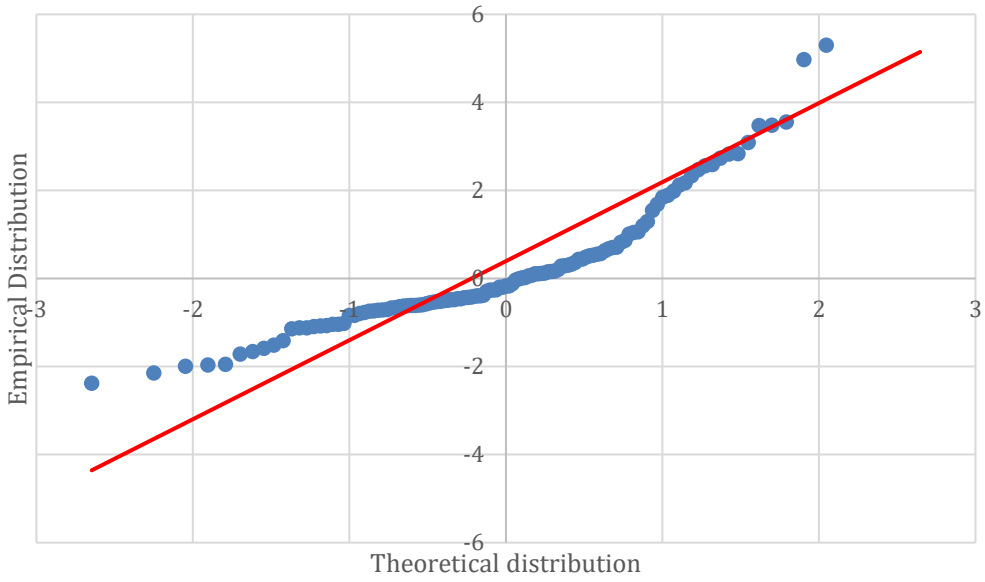

Figure 9 - Q-Q plot of model that is trained with Arnhem data

7.2.3 Validation

To see if a model does not only describe the relations between variables in one city or dataset, it is important to validate the results with data that the model has not seen before. In this research this is done by using a second city, which is Apeldoorn. This city is chosen since this city has a sufficient amount of count points and is a big city with a lot of cyclists. The bicycle count points of Apeldoorn are shown in Figure 10.

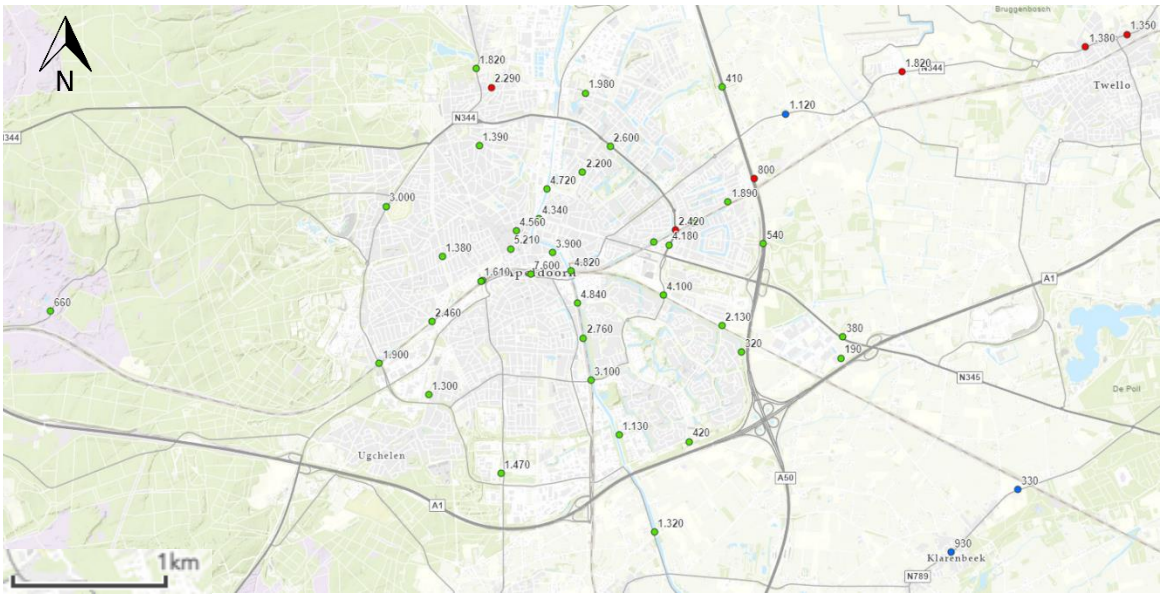

Figure 10 - Bicycle count points in Apeldoorn (Provincie Gelderland, 2020)

For 45 count points data is collected for the different characteristics, but three outliers were removed from the dataset, which leaves 42 points that are used to validate the model. The model is tested by calculating the expected rates with Equation 8 and then plotting these expected rates against the real rates for only Apeldoorn and for the entire dataset.

The plot of Apeldoorn is shown in Figure 11, together with a trend line displaying a 1 tot 1 relation between the observed rates and the modelled rates. The mean absoluter error of this regression equation in this dataset is 0.956 and the root mean square error is 1.579.

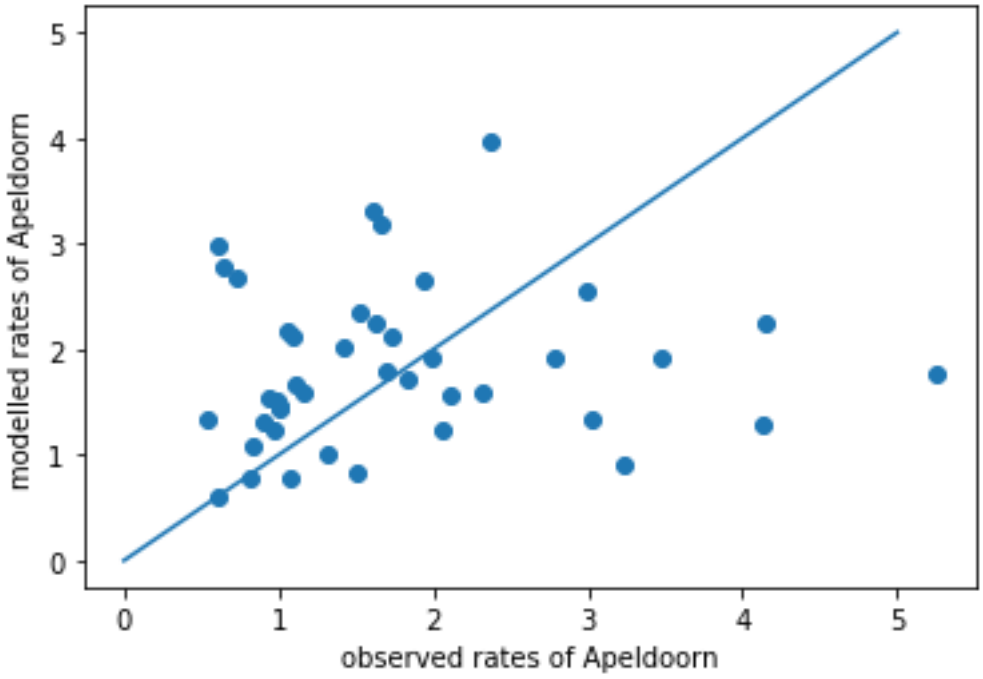

Figure 11 - Plot of the modelled and observed rates between observed counts and the Fietsmonitor data in Apeldoorn for the Arnhem trained model

To compare this model with the model trained with data from Apeldoorn and Arnhem, it is also useful to see how the model predicts the rates for the entire dataset. This plot is shown in Figure 12. The mean absolute error of this equation for this dataset is 1.220 and the root mean square error is 5.075. In red the regression trendline of the ratio is shown in the figure and in blue the 1 on 1 ratio trendline is displayed. The regression trendline shows that this model has a tendency to underpredict the observed rates.

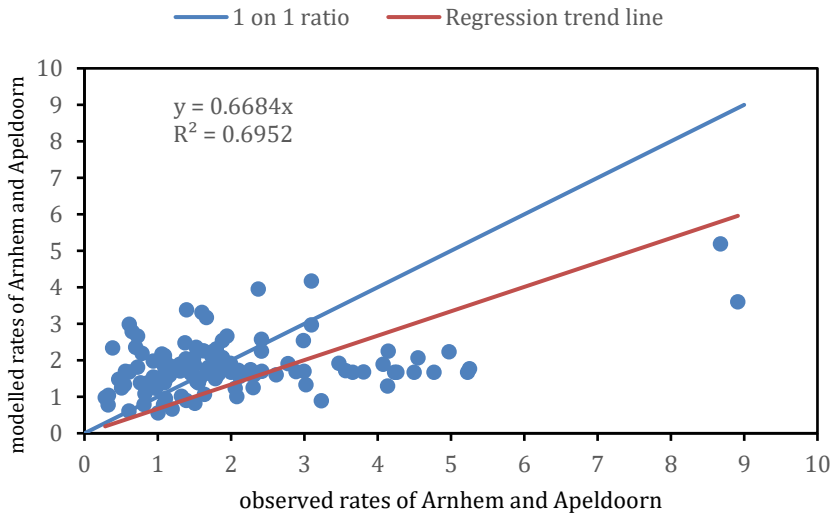

Figure 12 - Plot of the modelled and observed rates between observed counts and the Fietsmonitor data for the entire dataset for the Arnhem trained model

7.3 Loglinear model trained with data from Apeldoorn and Arnhem

The second loglinear model that is made in this research is a model based on a combined dataset of the bicycle count points in Apeldoorn and Arnhem. This model is made by dividing the dataset into two different sets. A training set, which is 67% of the data, which is used to set up the loglinear model and a test set, which is 33% of the data, to validate the model. To validate the model k-fold validation is used. In k-fold validation the data is split up multiple times in train and test sets and the outcomes of every different split are averaged to get a model that predicts the entire dataset the best and not just a specific data split.

7.3.1 Correlation test

Since the dataset is divided into training and testing sets five different times, five correlation tests are conducted. These are shown in Appendix C: Correlation results k-fold splits. In three of the five splits the size of the supermarket parcel and the number of shops was strongly correlated. To determine which of the two characteristics is allowed to enter the stepwise regression, again the correlation with the dependent variable has been the deciding factor. For every split the correlation of the size of the supermarket parcel was larger with the dependent variable than the correlation of the number of shops and therefore the size of the supermarket parcel can enter the stepwise regression, which is different from the first loglinear model.

7.3.2 Stepwise regression

In Table 5 the results of the loglinear regression analysis are shown. Although on average none of the characteristics is significant with a Confidence Interval of 90%, in single splits more characteristics were deemed to be significant. In the five different splits, the land use mix was included all five of the models, residential density in four of the five models and the percentage of residents between 15 and 24 years old in one model.

Table 5 - Loglinear regression analysis results of model trained with data from Apeldoorn and Arnhem

Characteristic	**β**	**SE**	**p-value**
Constant	1.183	0.586	0.110
Presence of bicycle curb lane	0.192	0.430	0.472
Presence of bicycle lane	-0.202	0.447	0.492
Presence of bicycle path	-0.340	0.271	0.458
Residential density	0.389	0.605	0.571
Employment density	-0.396	0.324	0.559
Size of supermarket parcel	-0.020	0.069	0.615
Greenness	0.163	0.560	0.694
On-street parking	0.083	0.238	0.739
Maximum car speed	-0.334	0.406	0.187
Land use mix	-1.018	0.771	0.245
Percentage non western immigrants	0.010	0.010	0.351
Address density	-0.204	0.649	0.721
Percentage residents between 15 and 24 years old	-0.013	0.026	0.329

Since the percentage of residents between 15 and 24 years old was only included in one model, this characteristic is not added to the final model. After taking the average of the coefficients per split the regression equation is as shown in Equation 9.

$$\ln(rate) = 1.0362 - 1.6617x_1 + 0.8041\,x_2 \qquad \text{Equation 9}$$

In this equation x_1 is the land use mix and x_2 is the residential density. This equation has an averaged R-squared value of 0.121.

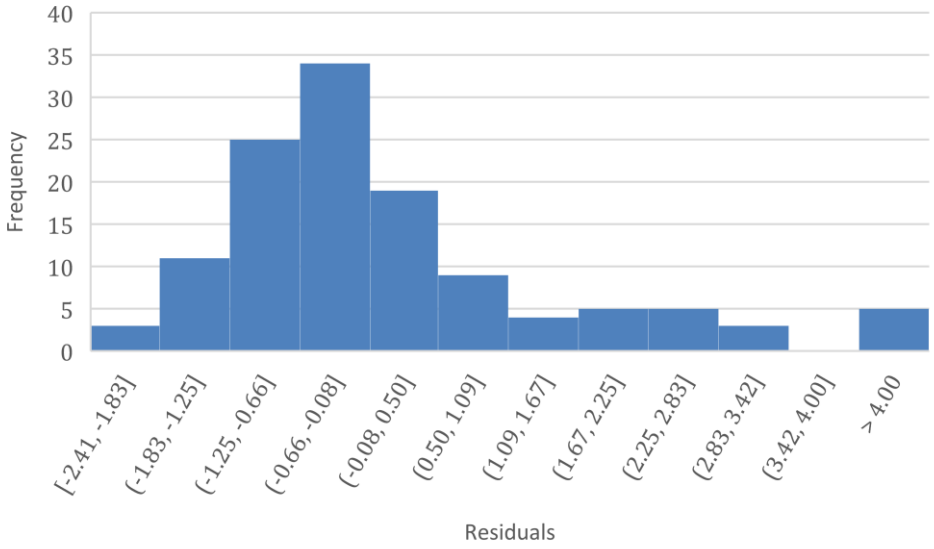

Figure 13 - Histogram of residuals of the model trained with data of Arnhem and Apeldoorn

In Figure 13 a histogram is shown of the residuals of this model and in Figure 14 the Q-Q plot of this model is shown, with a linear trendline displayed in red. From the histogram it could be concluded that the variance is normally distributed since the histogram is symmetrically bell-shaped around zero, however the Q-Q plot is not that conclusive, since this shows a more exponential trendline.

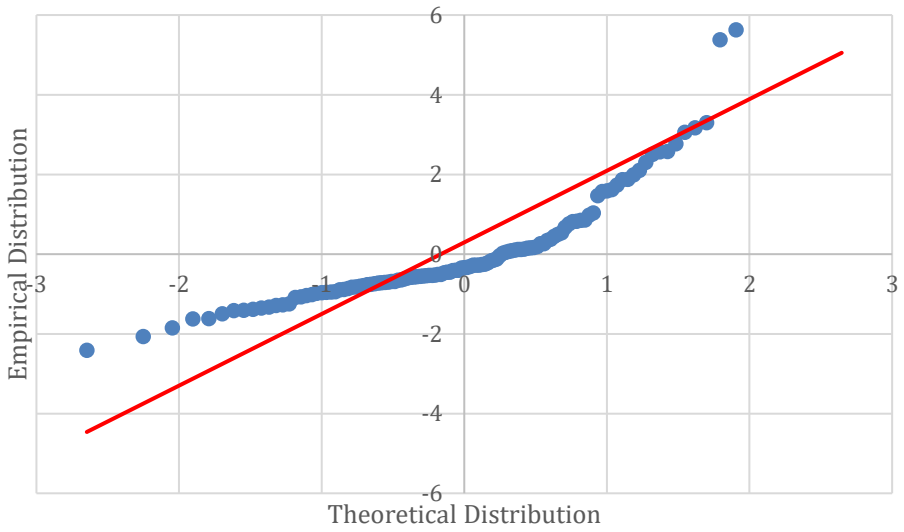

Figure 14 - Q-Q plot of the model trained with data of Arnhem and Apeldoorn

7.3.3 Validation

By using the k-fold cross validation method, the model parameters are already validated. However, how well the model is predicting can be determined by comparing the predicted rates with the observed rates. This could be done in two ways for the previous model, but not for this model, the different splits had different models as outcome than the eventual averaged model. Therefore, this model is only compared by plotting the predicted rates by the averaged model against the observed rates for the entire dataset. This plot is shown in Figure 15. In this plot five points are not displayed, since these points would have distorted the figure to not make it able to make a distinction between most of the points, but these have been taken into account in calculating the regression trendline. The mean absolute error of this equation for the entire dataset is 1.233 and the root mean square error is 5.232. In red a regression trendline is added to the figure and in blue the trendline of a 1 on 1 ratio is displayed. The regression trendline shows that the model has an tendency to underpredict the observed rates.

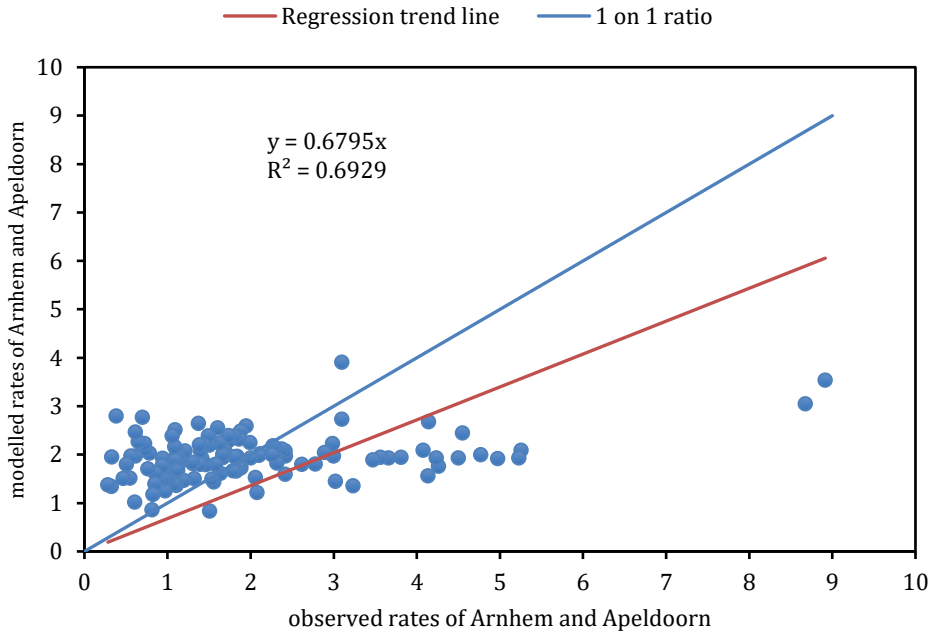

Figure 15 - Plot of the modelled and observed rates between observed counts and the Fietsmonitor data in the model with the combined data from Apeldoorn and Arnhem

8. Discussion

In this discussion first the two models will be compared and than the shortcomings and assumptions are discussed.

8.1 Comparison between the two loglinear models

In this research two different loglinear models have been defined to predict the rate between the observed counts and the predicted number of cyclists by the Fietsmonitor. The first model has used the count stations of Arnhem to train the model and the second model used all the count stations.

The first model predicted the rates of Apeldoorn well, as can be seen by the low error values. When looking at the errors for the prediction of the entire dataset the model is a bit less accurate. When a closer look is taken at the plot of the observed and modelled rates for the entire dataset, it can be seen that the first model tends to model a rate that is lower than the observed rate.

The second model predicted the rates of the observed counts and the Fietsmonitor worse than the first model, which can be seen in the higher error values. From the plot of the modelled and observed rates it can be seen that this model tends to underpredict the observed rates by a bit less than the first model. It can also be seen that there is not that much distinction in the predicted rates between the different count points.

From this comparison I conclude that the first model (the model trained with the Arnhem data) is the best model in predicting the ratio between the observed counts and the predicted number of cyclists by the Fietsmonitor. Although the second model predicts the Arnhem and Apeldoorn rates a bit better, it has a bigger error and therefore the first model is chosen. The finalised equation to model bicycle traffic is shown in Equation 10.

$$\text{bicycle traffic volume} = Fietsmonitor * e^{0.8588+1.3422x_1-1.5108x_2-0.8633x_3+0.6350x_4} \quad \text{Equation 10}$$

In this equation x_1 is the residential density, x_2 is the land use mix, x_3 is the employment density and x_4 is the presence of bicycle curb lanes.

8.2 Assumptions, recommendations and shortcomings

In this research a final model is presented. This model is subjective to the data that is used to train it and can be different when using other data sets. However, after validating the model with a different city it showed to have small error values and therefore it can be concluded that the model could still be applicable for different cities in the Netherlands as well.

In this research 12 characteristics are included in the stepwise regression. Ideally more characteristics would have been taken into account, however not for all characteristics the appropriate data is available. In further research it is advised to try and include all of the characteristics that are mentioned in the literature review in Chapter 4.

Furthermore, some datapoints were removed from the datasets. Some of these points were removed since the Fietsmonitor could not generate trips for a count point, which would make it impossible to calculate the rate between the actual counts and the Fietsmonitor. Furthermore, for some count points the data of the characteristics could not be found. Also a few data points were extremely high or low compared to the other points and were therefore removed.

Besides the removal of some data points, the way of measuring some of the characteristics could have been different to get a better result. The way of some characteristics in the eventual model differs from the proposed ideal way from previous researches. Therefore it could be the case that

certain characteristics are not taken into account properly in this research and have a bigger or smaller influence on the prediction than it should have.

Additionally the threshold values for the stepwise regression and the correlation test were assumed. If these values would have been different it could have been the case that an entirely different model was generated. Ideally, a 95% confidence interval would have been used as threshold for the stepwise regression, but this did not give any model to predict the bicycle traffic. Furthermore, the strength of the correlation coefficient was assumed at strongly correlating, which could have been set to a value less strong to prevent moderately correlated characteristics from being included together in a model.

During this research multiple different dependent variables have been tested. For example the bicycle counts and the difference between the actual counts and the Fietsmonitor data were tested. In this dataset these dependent variables did not give a sufficient and reliable model, however in further research it is advised to not forget to also look at the difference between the actual counts and the Fietsmonitor data as a dependent variable.

Since this research models the rate between the observed counts and the predicted number of cyclists by the Fietsmonitor, this research heavily depends on the bicycle counts of the province of Gelderland and the municipalities of Arnhem and Apeldoorn and the assumptions made by the developers of the Fietsmonitor.

The counts of the province of Gelderland and the municipalities of Arnhem and Apeldoorn are not the only counts that can be used in this research. During the span of this research it was also considered to use the count data of the Fietstelweek. The usage of the Fietstelweek could have provided a more detailed look into how spatial and demographic characteristics are influencing bicycle traffic, since the counts of the Fietstelweek are provided for almost every road segment. Although the data of Fietstelweek has its disadvantages, it can also provide a more detailed and maybe better view on the influence of the spatial and demographic characteristics. Therefore, in further research it is advised to also look into using Fietstelweek as observed counts.

Ideally the counts used in this research would have been conducted for a long period of time and in the same time period. In theory, the bicycle volume differs per month, in months with better weather and vacation periods the number of cyclists differs from months that have no vacation periods or have bad weather. This difference could have been accounted for if the counts were conducted on the same days. If they were conducted for a longer time than the few weeks that they are now, the randomness of these certain external factors could have been less and balanced.

Furthermore, most of the counts conducted by the province of Gelderland were on special bicycle paths or on other roads that have good bicycle infrastructure. Therefore, the difference between the counts in the bicycle infrastructure characteristic was really low. Which could influence the final model either including or excluding this characteristic on a few data points that were different and this made this characteristic prone to misfitting. Therefore, it is important to try and take more detailed counts into account in further research.

9. Conclusion

This research has contributed in creating a better understanding of which factors cause people to make a different bicycle routing choice than the shortest route. This is done by determining which characteristics influence cyclists to take different routes from the shortest route and by proposing a model to improve the prediction of the number of cyclists on a road.

In literature multiple characteristics were already mentioned influencing bicycle traffic and in this research these characteristics were combined in one overview. Not all of these characteristics have a significant influence in all or most of the researches, but the ones that had were separated from the complete overview. This resulted in nine different characteristics that together with three additional demographic characteristics were used in this to set up a regression model. In this research it was concluded that the use of loglinear or linear regression is best suited for this problem.

After conducting a stepwise loglinear regression only four different characteristics were deemed to have a significant influence on the rate between observed counts and the predicted number of cyclists by the Fietsmonitor. This shows that although all the characteristics used in the research were deemed to be a significant influence on bicycle traffic, only four were of additional influence on the Fietsmonitor, which solely takes the shortest path between origins and destinations into account.

In literature it was already proven that spatial and demographic characteristics influence bicycle traffic and after looking at the difference between the observed counts of the province of Gelderland and the municipalities of Apeldoorn and Arnhem and the predicted number of cyclists by the Fietsmonitor, it could be concluded that there are more motives for cyclists to take different routes from the shortest one. As the focus in the research was solely on spatial and demographic characteristics being that motive a loglinear regression model was setup with the four significant characteristics in this research to account for the differences between the observed and predicted counts by the Fietsmonitor.

This model can be used in practice by Witteveen+Bos as an addition to the shortest path prediction by the Fietsmonitor to let the Fietsmonitor predict the number of cyclists at a certain road even more precise than it is already doing.

10. References

CBS. (2021). *Kerncijfers wijken en buurten 2004-2020*. Retrieved from CBS: https://www.cbs.nl/nl-nl/reeksen/kerncijfers-wijken-en-buurten-2004-2020

Centraal Bureau voor de Statistiek. (2021). Retrieved from CBS: https://www.cbs.nl/

Chen, P., Zhou, J., & Sun, F. (2017). Built environment determinants of bicycle volumes: A longitudinal analysis. *Journal of Transport and Land Use*, 655-674.

Fagnant, D., & Kockelman, K. (2016). A direct-demand model for bicycle counts: the impacts of level of service and other factors. *Environment and Planning B: Planning and Design*, 93-107.

Fraser, S., & Lock, K. (2011). Cycling for transport and public health. *European Journal of Public Health*, 738-743.

Gemeente Arnhem. (2021). *Open data Portaal Arnhem*. Retrieved from opendata: https://opendata.arnhem.nl/

Griswold, J. B., Medury, A., & Schneider, R. J. (2011). Pilot Models for Estimating Bicycle Intersection Volumes. *Transportation Research Record*, 1-7.

Hankey, S., & Lindsey, G. (2016). Facility-demand models of peak period pedestrian and bicycle traffic: Comparison of fully specified and reduced-form models. *Transportation Research Record*, 48-58.

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

Harms, L., & Kansen, M. (2017). *Fietsfeiten.* Kennisinstituut voor Mobiliteitsbeleid.

Heinen, E., Wee, B. v., & Maat, K. (2010). Commuting by Bicycle : An Overview of the Literature. *Transport Reviews*, 59-96.

Hochmair, H. H. (2015). Assessment of Bicycle Service Areas around transit stations. *International Journal of Sustainable Transportation*, 15-29.

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami Dade county from Strava tracking data. *Journal of Transport Geography*, 58-69.

Holmgren, J., Moltubakk, G., & O'Neill, J. (2018). Regression-based evaluation of bicycle flow trend estimates. *Procedia Computer Science*, 518-525.

Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters*, 63-75.

Hunt, J., & Abraham, J. (2007). Influences on bicycle use. *Transportation*, 453-470.

Iacono, M., Krizek, K., & El-Geneidy, A. (2008). *Access to Destinations: How Close is Close Enough?* Minnesota: Minnesota Department of Transportation.

Jedlička, K., Ježek, J., Kolovský, F., Kozhukh, D., Martolos, J., Šťastný, J., . . . Beran, D. (n.d.). Retrieved from Open Transport Map: http://opentransportmap.info/download/

Kennissinstituut voor Mobiliteit. (2020). *Kerncijfers Mobiliteit 2020.* Kennisinstituut voor Mobiliteit.

Kim, N. S., & Susilo, Y. O. (2011). Compariosan of pedestrian trip generation models. *Journal of Advanced Transportation*, 399-412.

Lord, D., Park, B.-J., & Levine, N. (2013). *CrimeStat IV.*

McBain, C., & Caulfield, B. (2017). An analysis of the factors influencing journey time variation in the cork public bike system. *Sustainable Cities and Society*, 641-649.

Meyer, J. (n.d.). *Poisson or Negative Binomial? Using Count Model Diagnostics to Select a Model*. Retrieved from The analysis factor: https://www.theanalysisfactor.com/poisson-or-negative-binomial-using-count-model-diagnostics-to-select-a-model/

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

Nederlandse Overheid. (2021). *Open data van de Overheid*. Retrieved from Overheid.nl: https://data.overheid.nl/

Prato, C. G., Halldórsdóttir, K., & Nielsen, O. A. (2018). Evaluation of land-use and transport network effects on cyclists' route choices in the Copenhagen Region in value-of-distance space. *International Journal of Sustainable Transportation*, 770-781.

Provincie Gelderland. (2020). *Fietstellingen 2020*. Retrieved from https://gelderland.maps.arcgis.com/apps/webappviewer/index.html?id=74c12adc49754950aeab6f8102325845

Provincie Gelderland. (n.d.). *Fietsen*. Retrieved from Gelderland: gelderland.nl/Fietsen

Pucher, J., Dill, J., & Handy, S. (2010). Infrastructure , programs , and policies to increase bicycling : An international review. *Preventive Medicine*, 106-125.

Ratnasari, D., Nazir, F., Toresano, L. O., & Pawiro, S. A. (2016). The correlation between effective renal plasma flow (ERPF) and glomerular filtration rate (GFR) with renal scintigraphy 99m Tc-DTPA study. *Journal of Physics Conference Series*.

Rietveld, P., & Daniel, V. (2004). Determinants of bicycle use: Do municipal policies matter? *Transportation Research Part A: Policy and Practice*, 531-550.

Rijsman, L., van Oort, N., Ton, D., Hoogendoorn, S., Molin, E., & Teijl, T. (2019). *Walking and bicycle catchment areas of tram stops.* Delft: TU Delft.

Saelens, B. E., Sallis, J. F., & Frank, L. D. (2016). Environmental Correlates of Walking and Cycling. *Annals of Behavioral Medicine*, 80-91.

Sener, I., Eluru, N., & Bhat, C. (2009). An analysis of bicycle route choice preferences in Texas. *Transportation*, 511-539.

Snizek, B., Nielsen, T., & Skov-Petersen, H. (2013). Mapping bicyclists ' experiences in Copenhagen. *Journal of Transport Geography*, 227-233.

Strauss, J., & Miranda-Moreno, L. F. (2013). Spatial modeling of bicycle activity at signalized intersections. *Journal of Transport and Land Use*, 47-58.

Taylor, D., & Mahmassani, H. (1996). Analysis of stated preferences for intermodal bicycle-transit interfaces. *Transportation Research Record*, 86-95.

Wang, X., Lindsey, G., Hankey, S., & Hoff, K. (2014). Estimating Mixed-Mode Urban Trail Traffic Using Negative Binomial Regression Models. *Journal of Urban Planning and Development*.

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2010). Built environment influences on healthy transportation choices: Bicycling versus driving. *Journal of Urban Health*, 969-993.

Winters, M., Davidson, G., Kao, D., & Teschke, K. (2011). Motivators and deterrents of bicycling : comparing influences on decisions to ride. *Transportation*, 153-168.

Winters, M., Teschke, K., Grant, M., Setton, E. M., & Brauer, M. (2010). How far out of the way will we travel? *Transportation Research Record*, 1-10.

Zhao, Y., Lin, Q., Ke, S., & Yu, Y. (2020). Impact of land use on bicycle usage: A big data-based spatial. *The Journal of Transport and Land Use*, 299-316.

Appendix A: Comparison scatterplots with and without transforming the dependent variable

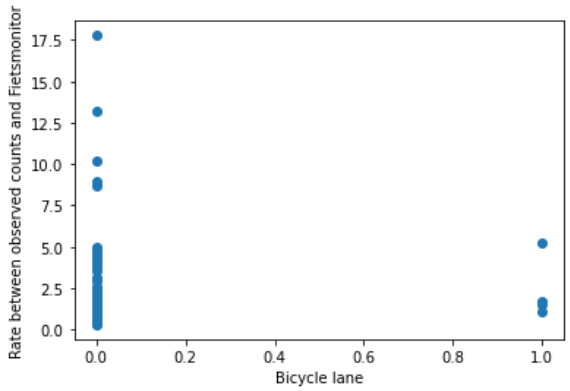

Figure 16 - Scatterplot of presence of bicycle lane and rate

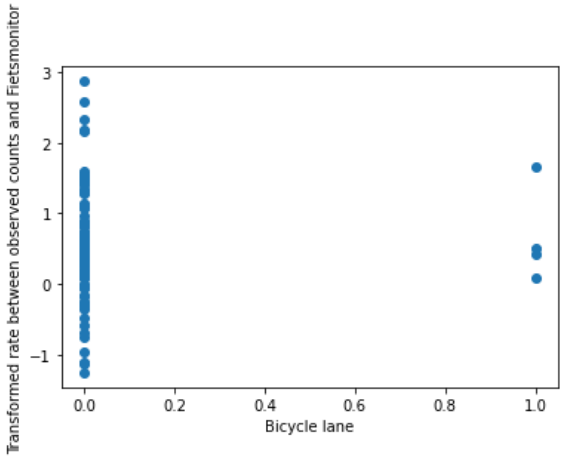

Figure 17 - Scatterplot of presence of bicycle lane and transformed rate

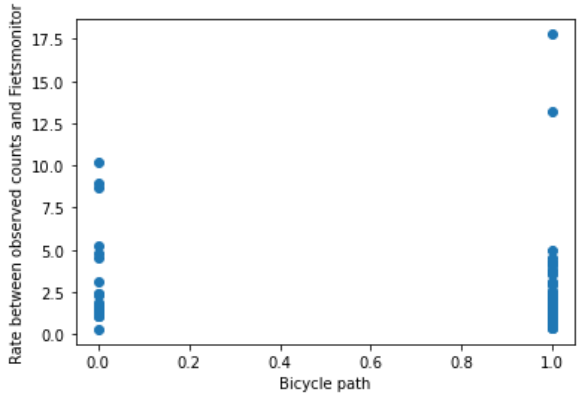

Figure 18 - Scatterplot of presence of bicycle path and rate

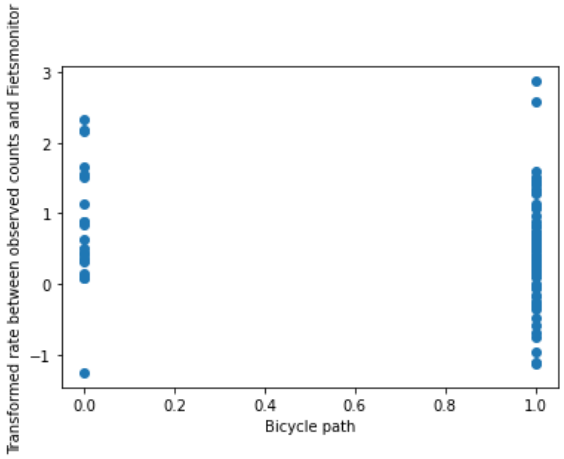

Figure 19 - Scatterplot of presence of bicycle path and transformed rate

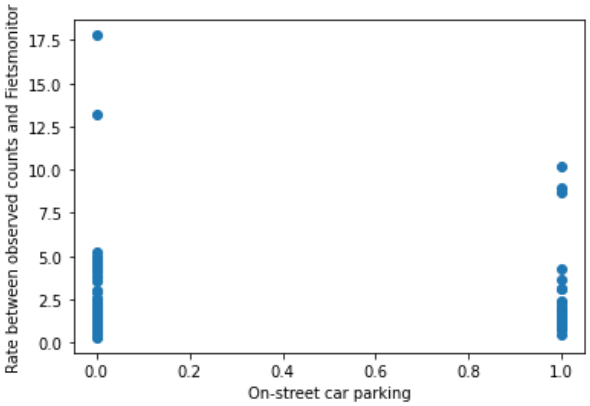

Figure 20 - Scatterplot of presence of on-street car parking and rate

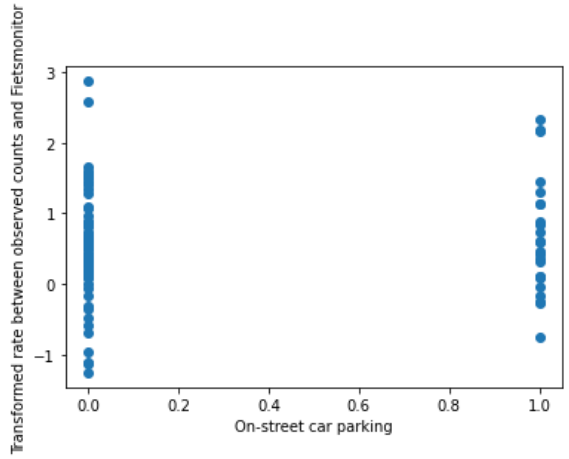

Figure 21 - Scatterplot of presence of on-street car parking and transformed rate

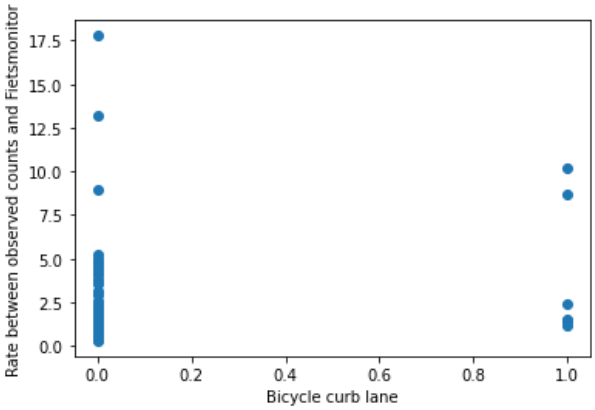

Figure 22 - Scatterplot of presence of bicycle curb lane and rate

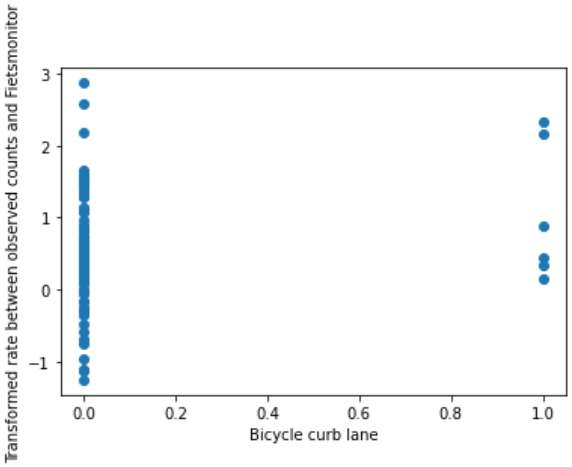

Figure 23 - Scatterplot of presence of bicycle curb lane and transformed rate

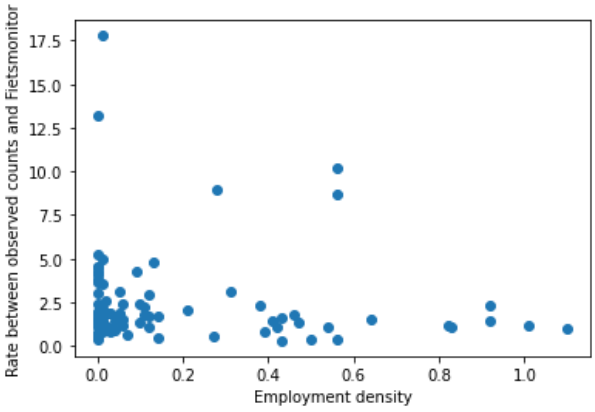

Figure 24 - Scatterplot of employment density and rate

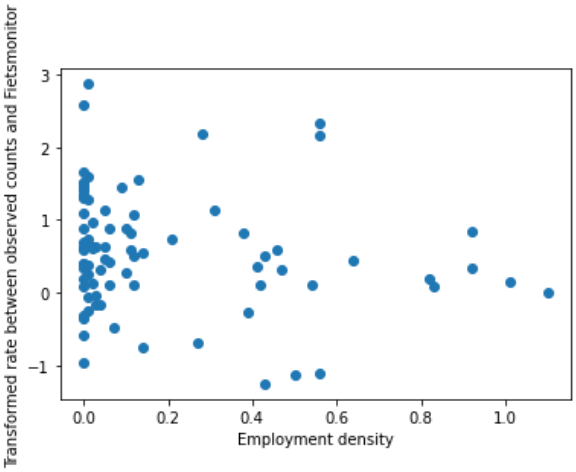

Figure 25 - Scatterplot of employment density and transformed rate

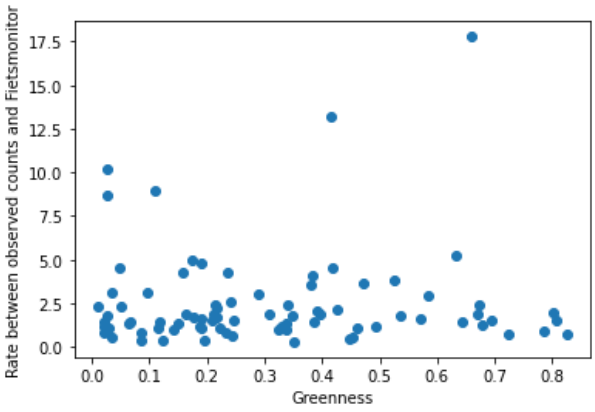

Figure 26 - Scatterplot of greenness and rate

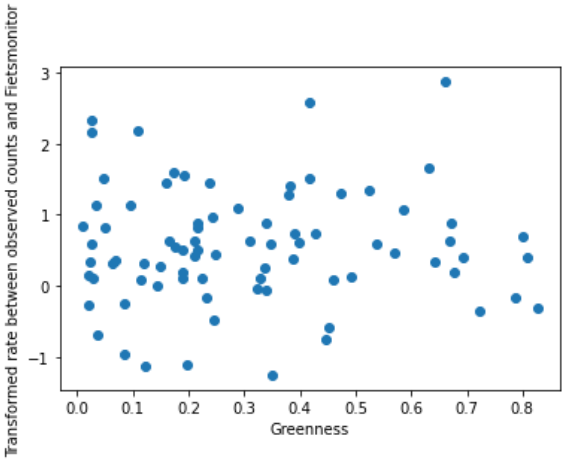

Figure 27 - Scatterplot of greenness and transformed rate

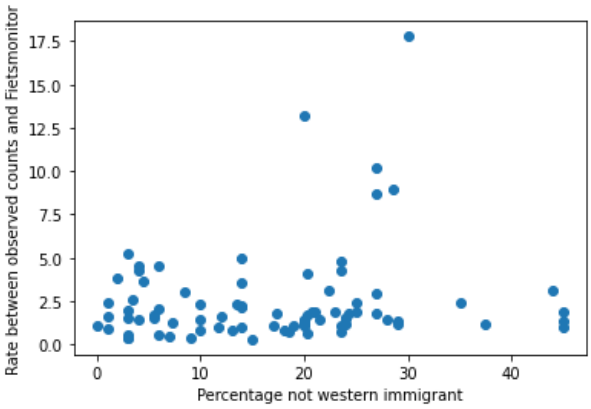

Figure 28 - Scatterplot of percentage not western immigrants and rate

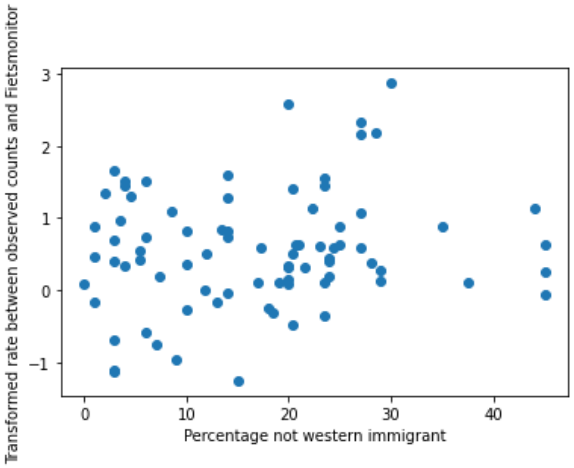

Figure 29 - Scatterplot of percentage not western immigrants and transformed rate

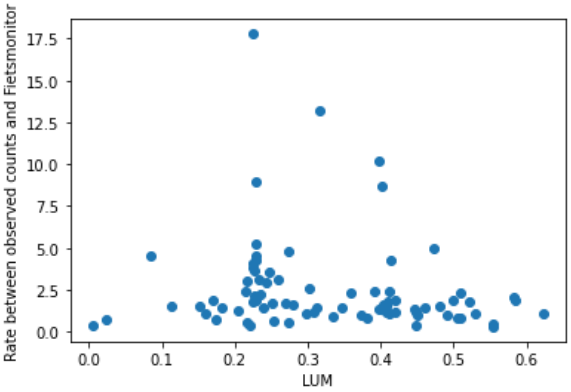

Figure 30 - Scatterplot of land use mix and rate

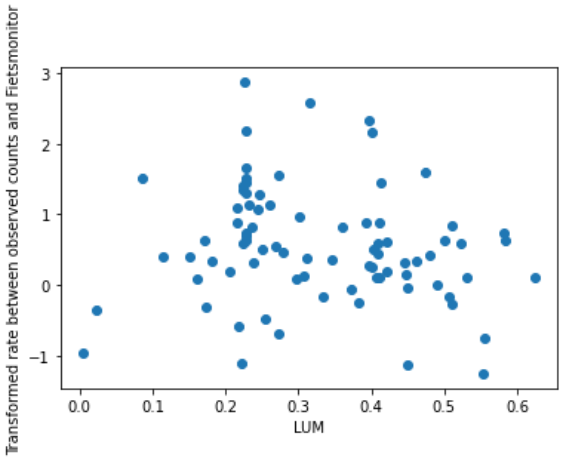

Figure 31 - Scatterplot of land use mix and transformed rate

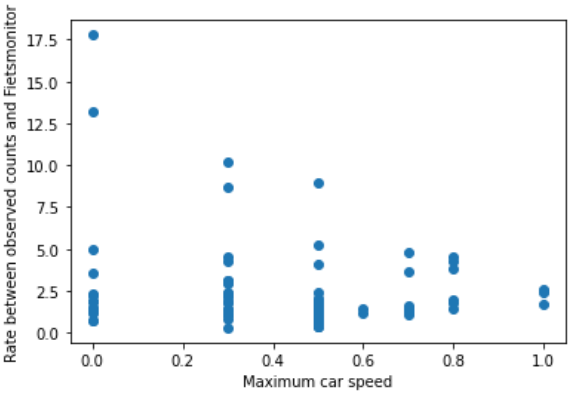

Figure 32 - Scatterplot of maximum car speed and rate

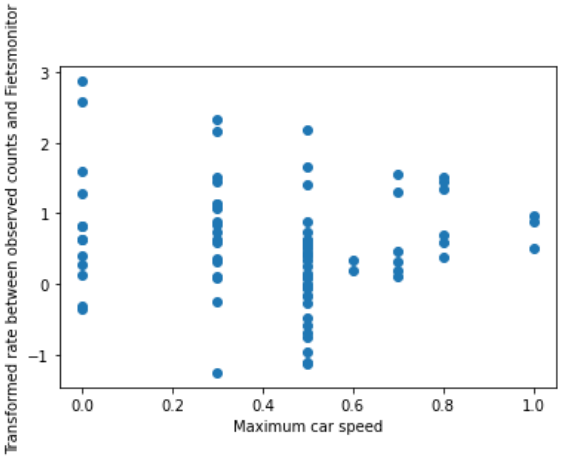

Figure 33 - Scatterplot of maximum car speed and transformed rate

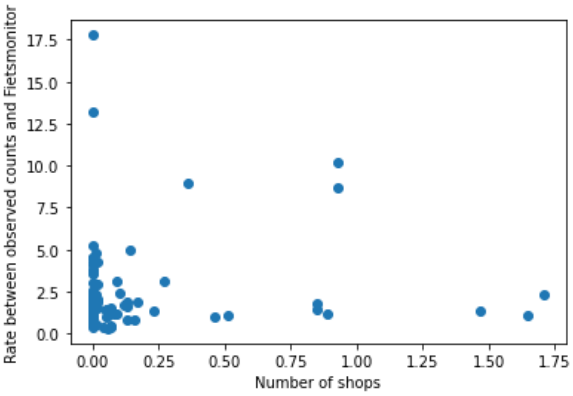

Figure 34 - Scatterplot of number of shops and rate

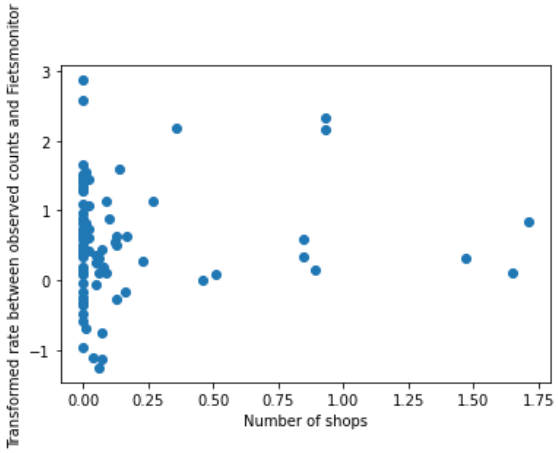

Figure 35 - Scatterplot of number of shops and transformed rate

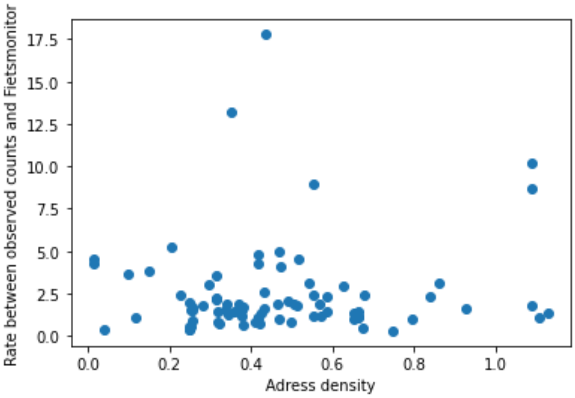

Figure 36 - Scatterplot of address density and rate

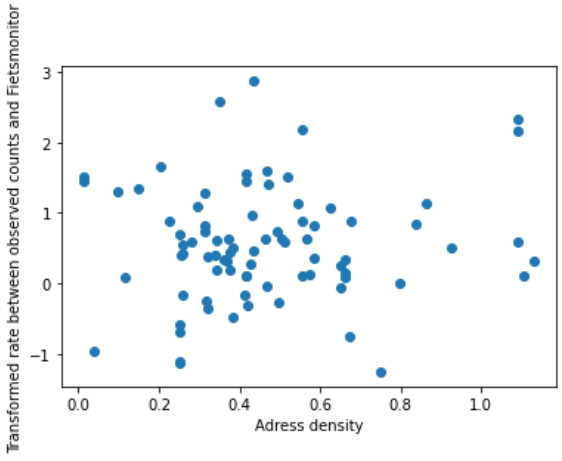

Figure 37 - Scatterplot of address density and transformed rate

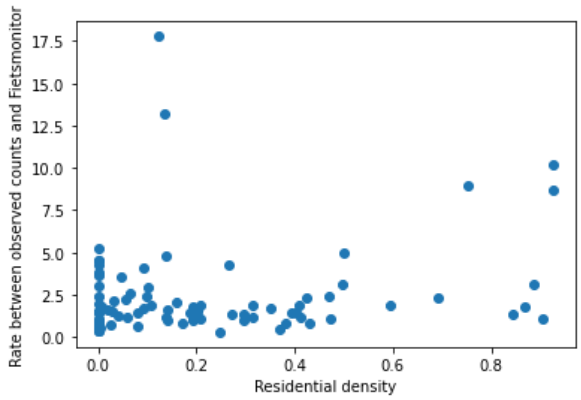

Figure 38 - Scatterplot of residential density and rate

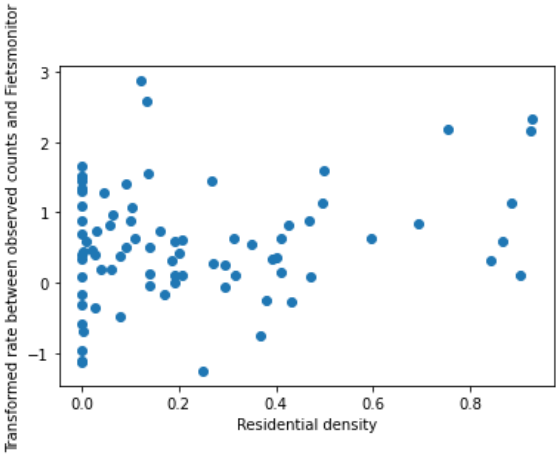

Figure 39 - Scatterplot of residential density and transformed rate

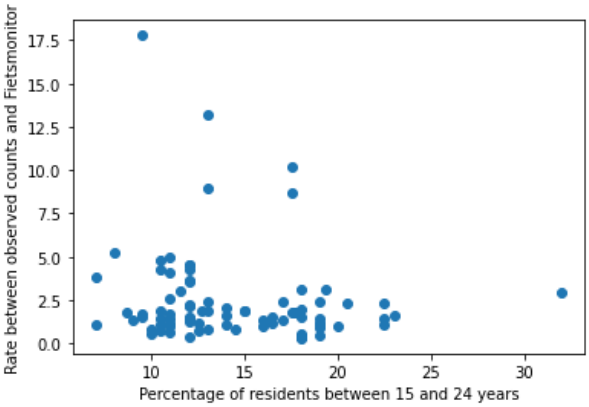

Figure 40 - Scatterplot of percentage of residents between 15 and 24 years old and rate

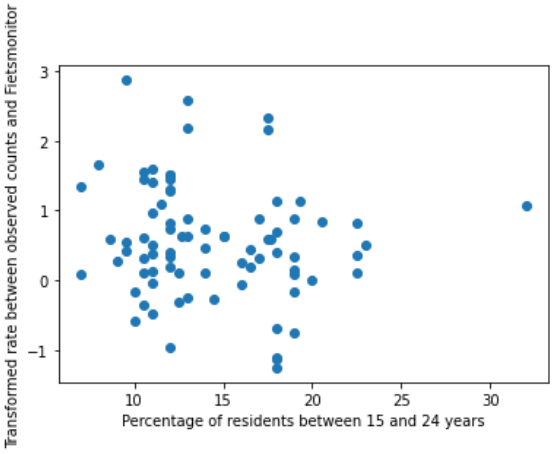

Figure 41 - Scatterplot of percentage of residents between 15 and 24 years old and transformed rate

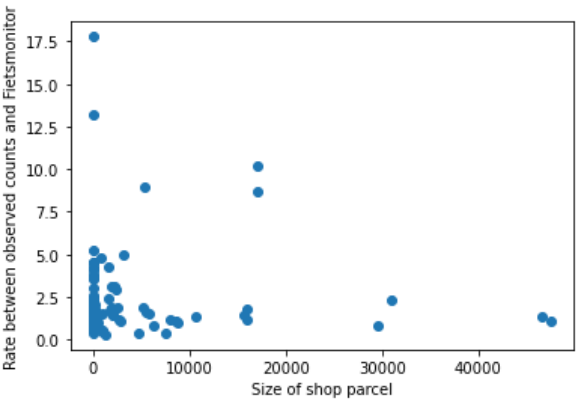

Figure 42 - Scatterplot of size of shop parcel and rate

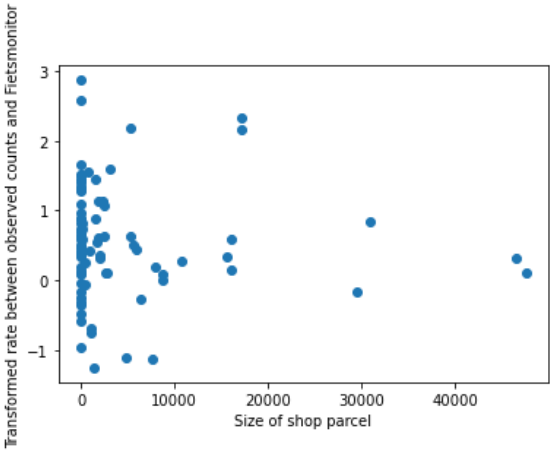

Figure 43 - Scatterplot of size of shop parcel and transformed rate

Appendix B: Correlation between dependent and independent variables Arnhem trained model

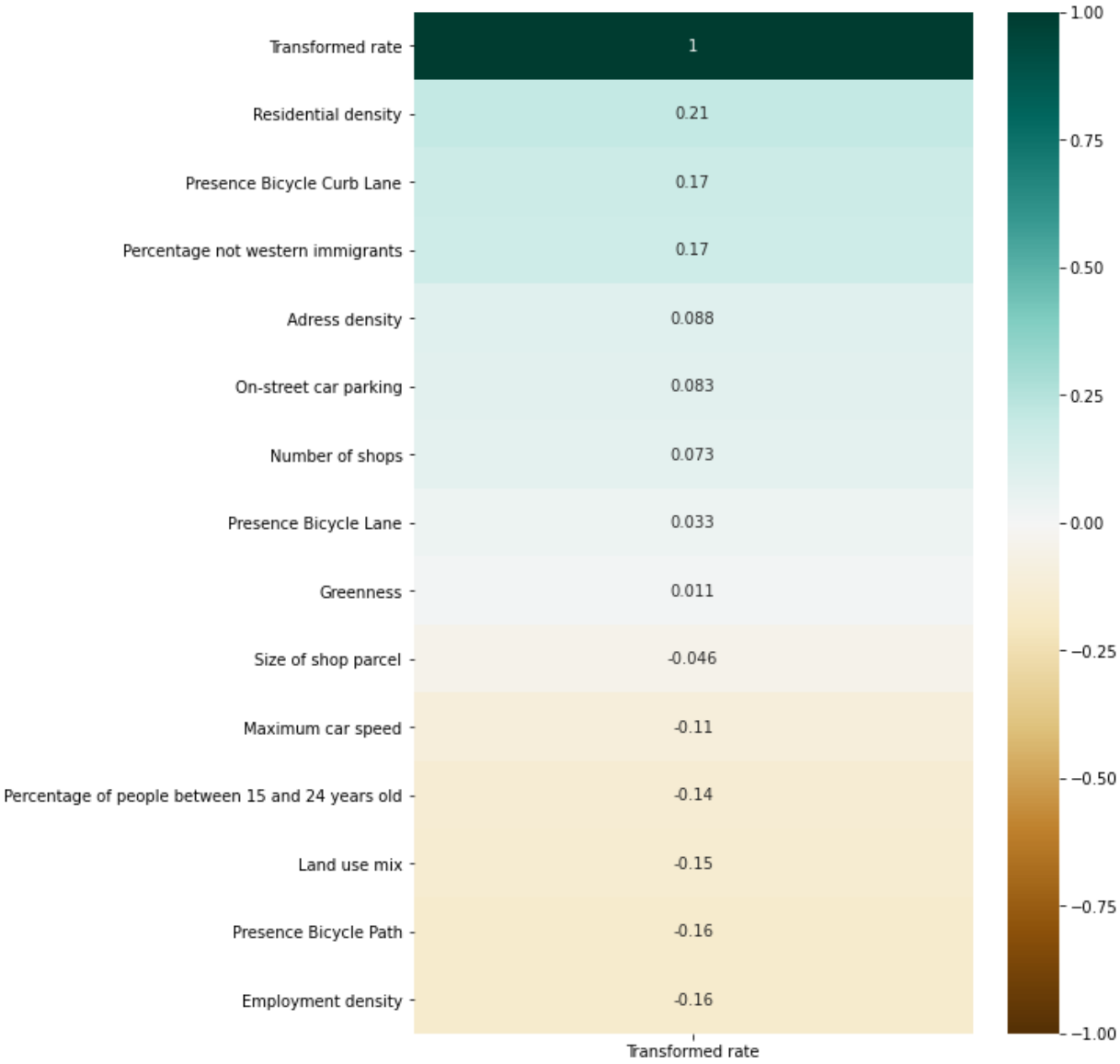

Figure 44 - Pearson correlation between dependent and independent variables of Arnhem trained model

Appendix C: Correlation results k-fold splits

Split 1

Figure 45 - Pearson correlation heatmap of split 1

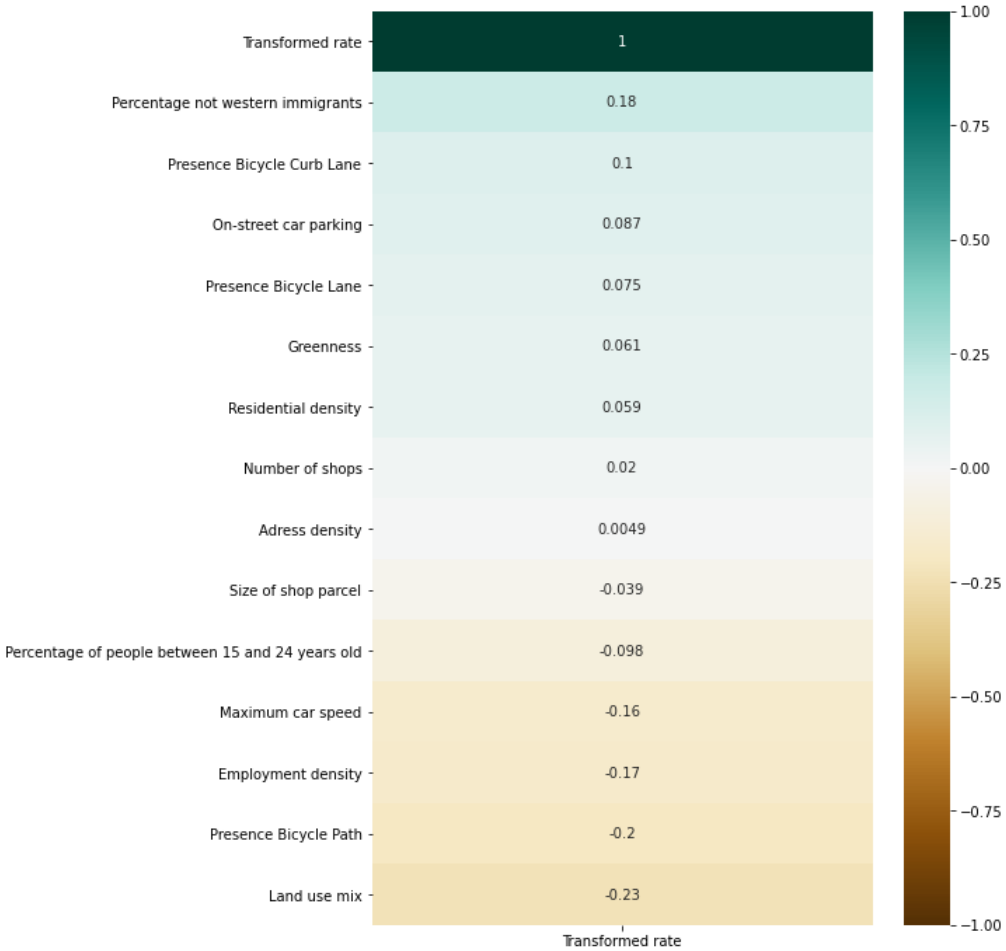

Figure 46 - Pearson correlation between independent and dependent variable of split 1

Split 2

Figure 47 - Pearson correlation heatmap of split 2

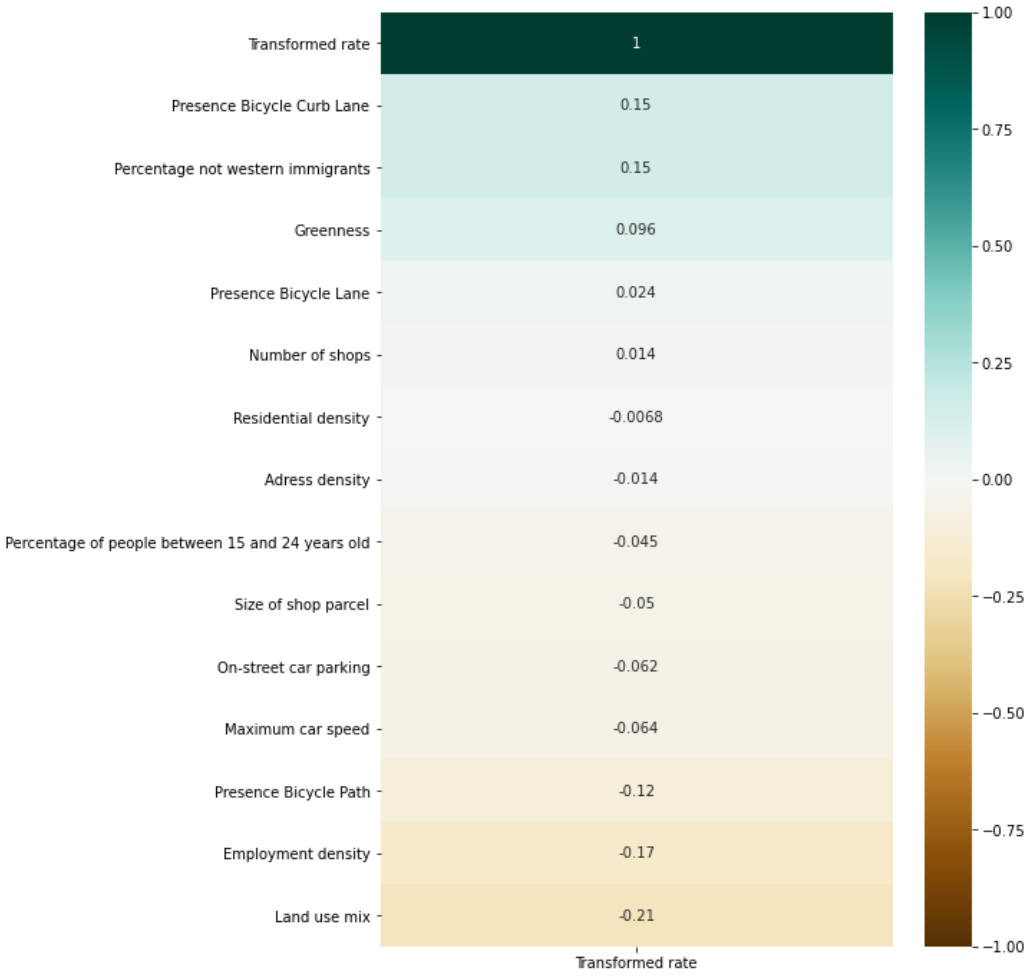

Figure 48 - Pearson correlation between independent and dependent variable of split 2

Split 3

Figure 49 - Pearson correlation heatmap of split 3

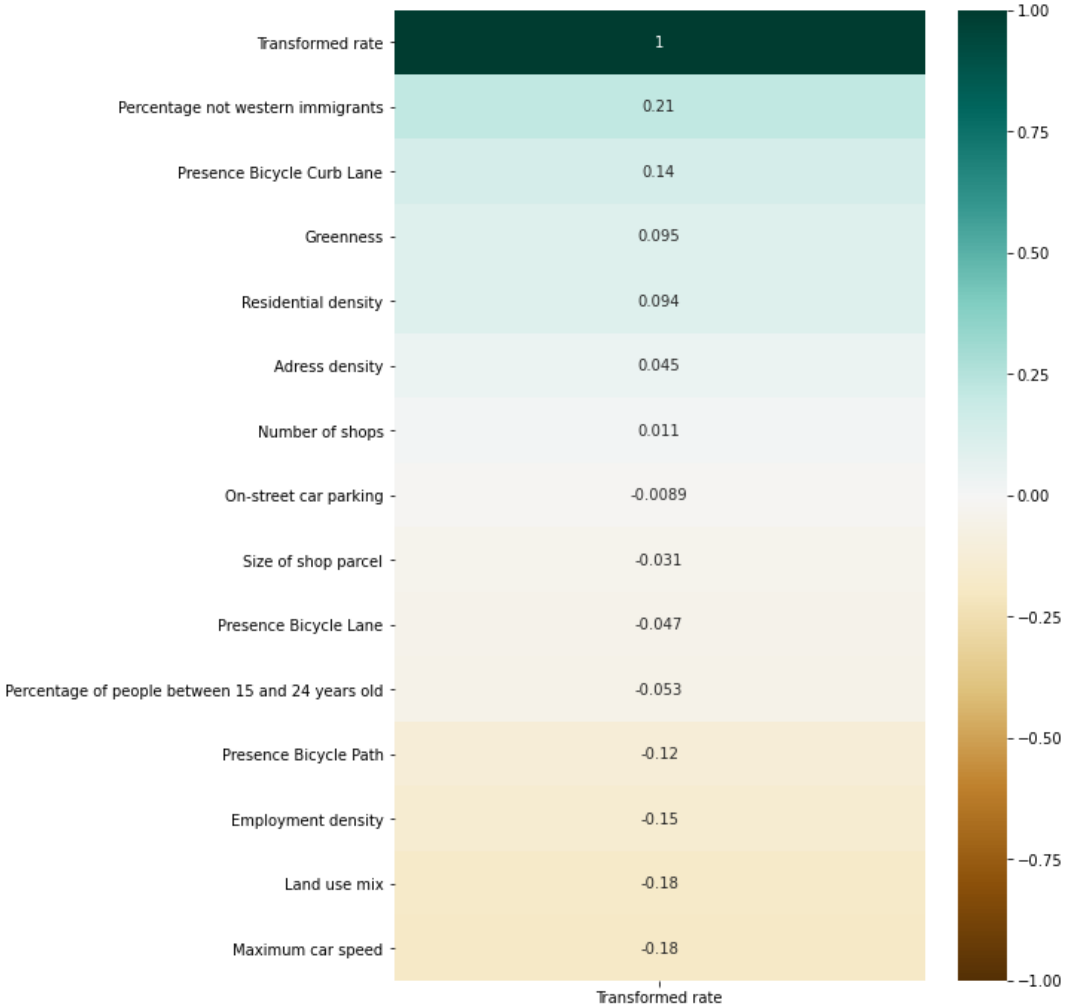

Figure 50 - Pearson correlation between independent and dependent variable of split 3

Split 4

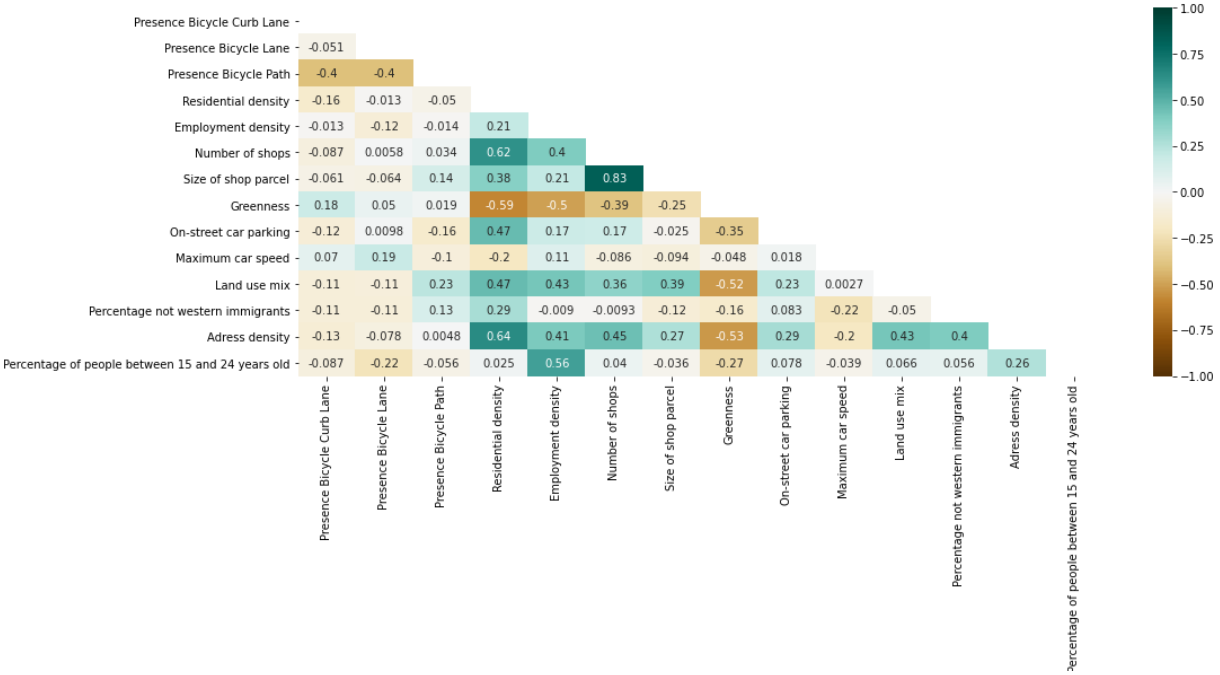

Figure 51 - Pearson correlation heatmap of split 4

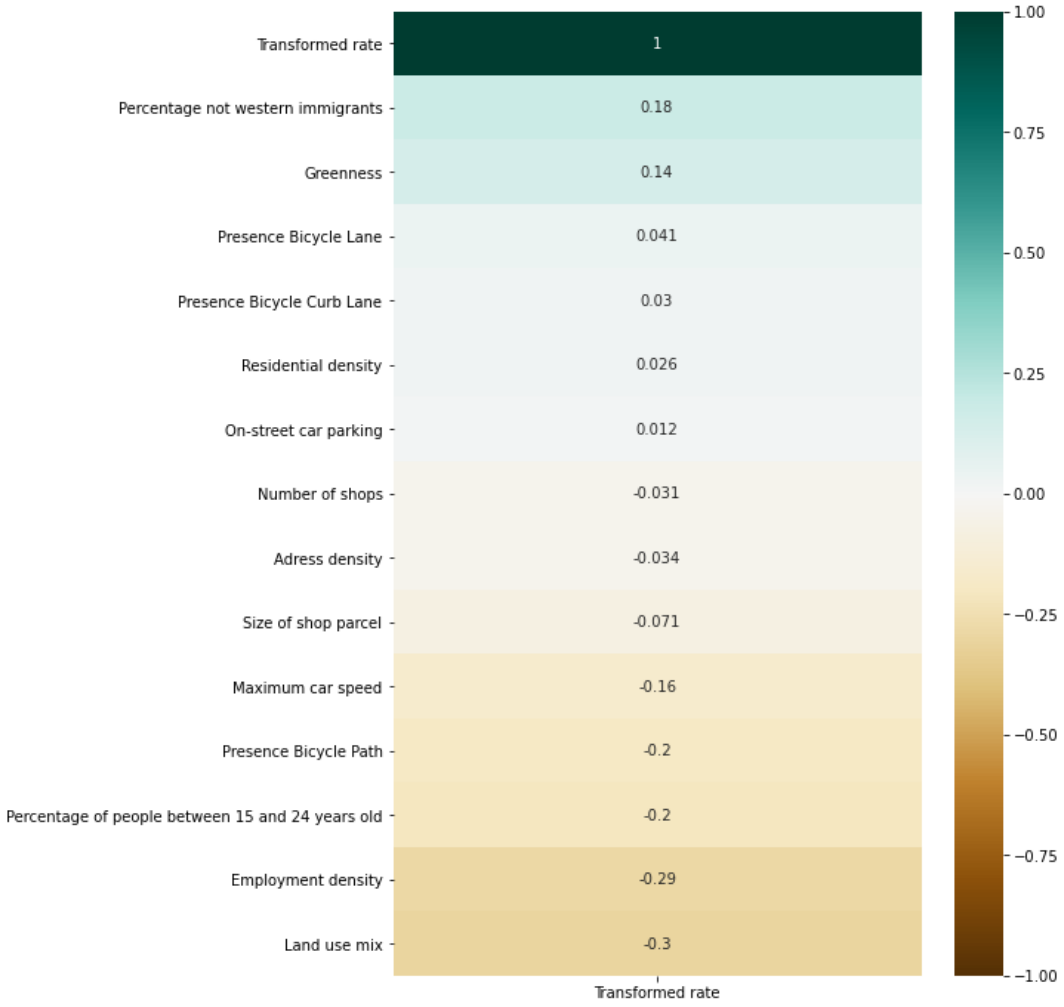

Figure 52 - Pearson correlation between independent and dependent variable of split 4

Split 5

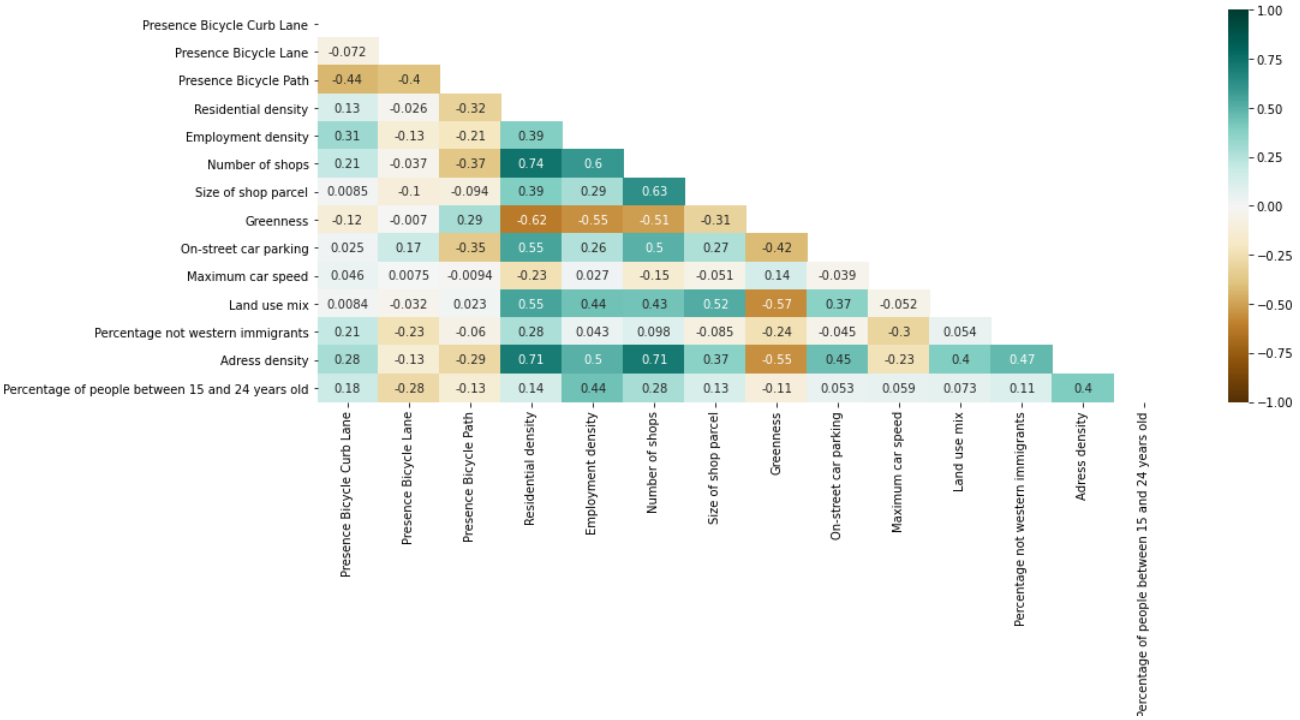

Figure 53 - Pearson correlation heatmap of split 5

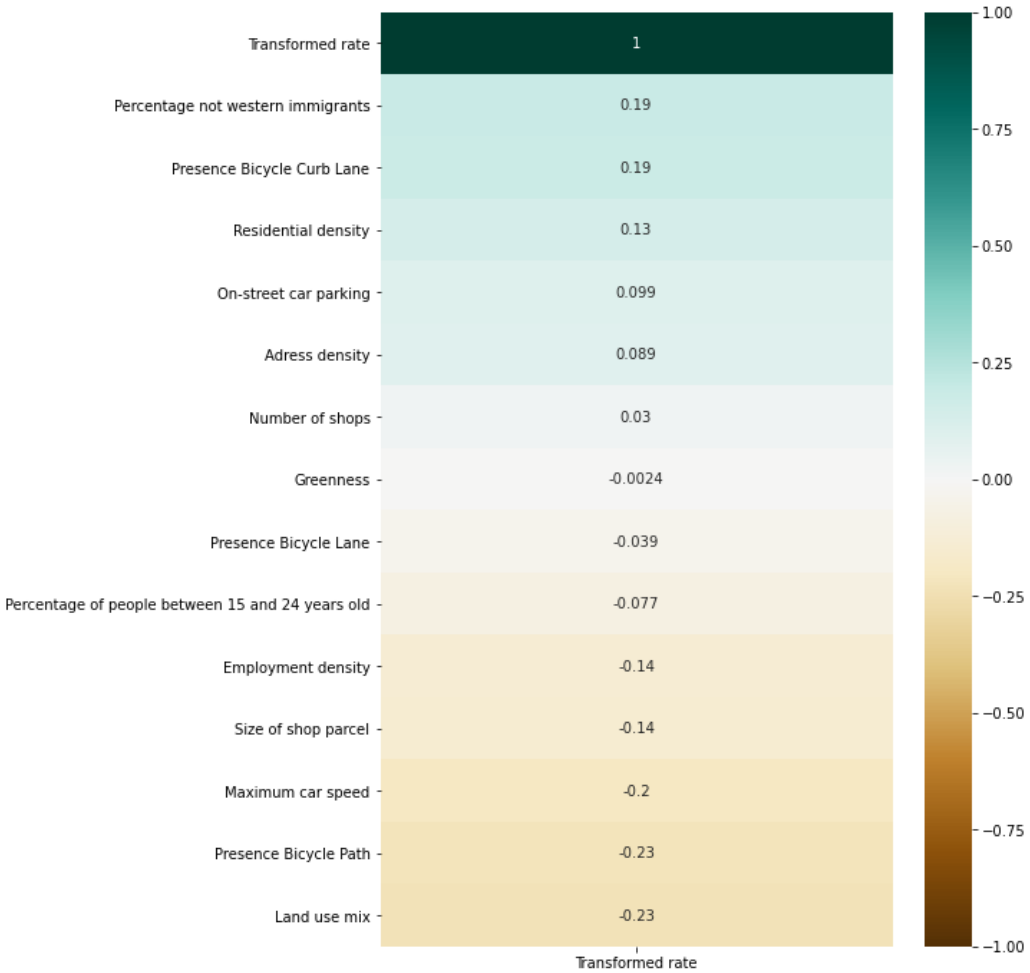

Figure 54 - Pearson correlation between independent and dependent variable of split 5