Predicting the volume of cyclists at road segments based on environmental characteristics

F.L.D. Witjes (Floris-Luc)
s2833697 - University of Twente

dr.ir. M.B. Ulak (Baran)
Internal supervisor - University of Twente

ir. S.A. Veenstra (Sander)
External supervisor - Witteveen+Bos

Final Thesis
24 June 2025

UNIVERSITY OF TWENTE.

Witteveen + Bos

Preface

Before you lies the final report of my Bachelor Thesis “Predicting the volume of cyclists at road segments based on environmental characteristics”. I carried out this research at the Witteveen+Bos office in Deventer, within the Mobiliteit en Verkeersveiligheid team. The goal of this thesis was to predict the volume of cyclists from a spatial planning perspective, considering a different approach than is usually done. With this approach I determined the effects of different environmental characteristics on cyclists not via a route choice methodology, but by purely analysing the characteristics of segments. I hope this research can contribute to an improved understanding of factors influencing cyclists and finally lead to an improved cycling experience.

During this thesis period I was able to learn many new things about QGIS, Python and regression modelling, which I very much enjoyed. I also found it very interesting to work on such a big project on my own, as that gave me the opportunity to really dive deep in to certain topics myself and truly become some sort of an expert in those areas. I want to thank Witteveen+Bos for giving me the opportunity to work on this research for the past ten weeks.

Next to that, I want to thank the colleagues in Deventer for the nice and welcoming experience and the fun chats at the coffee table, which made the internship very pleasant. I also want to thank my supervisor Sander Veenstra for all of the helpful feedback and fruitful discussions we had, which helped shape my research into what it now is. Besides that I also want to thank Baran Ulak for his comments and advice to improve my research even further.

In the past ten weeks I have had a great time carrying out this research and I hope you enjoy reading my thesis.

Floris-Luc Witjes

Deventer, 24 June 2025
f.l.d.witjes@student.utwente.nl

Abstract

In the Netherlands the number is cyclists is significant and expected to grow even further in the coming years. This makes it important for municipalities and other policy makers to have an insight cyclist intensities, as they can use this information when new developments affecting cyclists are made. Currently, models that estimate cyclist intensities often use the four step model as their core approach, which estimates the shortest route between the origin and destination of a trip. In research it was found that many trips do not follow the shortest route, but are instead influenced by environmental factors to take a different route (De Jong et al., 2023). Models that follow a spatial planning approach, which are often used to estimate pedestrian intensities use only environmental characteristics to predict and could be a solution to this problem. An example of a model that functions as such is the Loopmonitor of Witteveen+Bos, which predicts pedestrian intensities.

In this research it was evaluated if a model that only took into account environmental characteristics, like the Loopmonitor, can predict cyclist intensities on a case study of Apeldoorn, the Netherlands. Based on a literature study relevant characteristics that affect cyclists were determined, whilst also characteristics that are used in spatial planning analyses were found. These characteristics were used in a regression model, where Random Forest regression was found to be the most suitable regression type.

The findings from the regression model reveal that network characteristics were the most influential, whilst infrastructural characteristics had very little effect on the model results. In the validation of the results a Mean Absolute Error (MAE) of 436 cyclists on a daily basis was found, with the Root Mean Square Error (RMSE) being 681. These errors were more significant for locations with many cyclists. In a comparison with a four step model, the Fietsmonitor of Witteveen+Bos it was found that the constructed model has a greater predictive capability on Apeldoorn.

From this research it is furthermore recommended to incorporate network analysis more into research related to predicting cyclist intensities, as these were determined to be very impactful in this study. Lastly, it is recommended to compare the model and the Fietsmonitor again on a different city, so that an independent comparison can also be made about the accuracies of both models.

Table of Contents

1. Introduction

1.1 Problem Context

In the Netherlands a lot of people cycle for all kinds of different purposes. Some cycle to commute to their work or studies, whilst other cycle as leisurely activity. The national government of the Netherlands also hopes to achieve a 20% increase in in the distance travelled by cyclists in 2027, compared with 2017 (Ministerie van Algemene Zaken, 2024). In order to realise this increase, investments to improve cycling infrastructure are needed to making cycling more attractive and to ensure that the infrastructure has enough capacity for the increased demand. These developments are preferably implemented at locations where they can have a high impact.

To gain insight into how cyclists travel, where they go and which roads they use, traffic models are used. With these traffic models policy makers can gain insight into the behaviour of cyclists, which they can use to determine how and where it is possible to improve cycling infrastructure. to ensure that it can cope with the increased demand in the future.

Traditionally traffic models are based on a four step model and made using an engineering approach. A four step model consists of the four steps trip generation, trip distribution, mode choice and route assignment. In the trip generation step the total amount of trips that have their origin in a zone or their destination in a zone are analysed and determined. In the trip distribution step the origins and destinations of trips are matched to create a trip matrix. In the mode choice step the trips are divided per mode, so car, bicycle or public transportation for example. In the route assignment step finally the routes of the trips are decided (McNally, 2007). Based on this traffic engineering principle most traffic models are made, one of which is the ‘Fietsmonitor’, or bicycle monitor in English, which is a model of engineering firm Witteveen+Bos.

In four step models for cyclists and pedestrians the shortest routes are chosen, as these modes are in general not affected by differing speed limit. Nevertheless, it has been seen in research that cyclists cover 59% more distance on average than the shortest route (de Jong et al., 2023). They showed that this difference in route choice was due to several environmental factors. It was found that cyclists take detours in order to cycle on route sections with some form of cycling infrastructure or on flatter and water facing routes. In addition to the study of de Jong et al. (2023), more studies have been conducted, also finding additional factors influencing the route choice of cyclists and causing them to travel on other routes than the shortest possible route.

The effects of environmental characteristics on route preferences of cyclists are incorporated into the Fietsmonitor and other four step models, but it still remains a question if this effect can be correctly considered in a four step model, as the majority of trips follow a different route than the shortest route.

To analyse the volume of pedestrians at different segments, the 'Loopmonitor', or walking monitor in English, has been developed by Witteveen+Bos. This model is not based on the core principle of the four step model, but is instead based on a spatial engineering approach, where spatial, infrastructural and land-use characteristics, proximity of different facilities and network connectivity of the segments are used to predict the volumes (De Wit et al., 2021). This is a different approach than the traditional traffic engineering approach that is used for four step models such as the Fietsmonitor.

The goal of commissioning party Witteveen+Bos is to determine if a spatial planning approach, like the approach of the Loopmonitor, of purely looking into spatial, infrastructural, land-use and network characteristics to predict pedestrian intensities can also be applied to a model for cyclists.

1.2 Research Gap

In the existing research, frequently the assumption on the behaviour of cyclists comes from the four step model. Using this, the route choice of cyclists is studied. In several studies it was found that cyclists do not only take the route length into consideration, which is the core assumption from the four step model, but that also environmental characteristics play an important role. From this it is shown that this traditional traffic engineering approach does not fully accurately predict the volume of cyclists and thus new methods should be analysed.

In the field of spatial planning, the volumes of pedestrians at roads can be predicted with model that only take into account environmental characteristics and follows a significantly different approach from the traditional four step model. This approach has however not been fully implemented in an analysis of cyclists yet, as route choice still is deemed an important factor. This shows a gap in the research on knowledge about predicting cyclist intensities from a spatial planning approach. The aim of this thesis is to contribute to bridging this research gap.

1.3 Research Aim

Based on the found knowledge gap in the literature, the aim of research is set-up.
The aim of this research is to predict the volume of cyclists at road segments based on the environmental characteristics of the segments, by using a spatial planning approach to create a cycling model, on a case study on the city of Apeldoorn and comparing these results with a traditional traffic engineering four-step model, to determine what the differences and similarities between the different types of models are.

1.4 Research Scope

The scope of the research illustrates the boundaries of this research and is in clarification and addition to the research aim.

1.4.1 Environmental characteristics

Environmental characteristics are an important aspect of this research and therefore it is necessary to properly define this concept in the context of this research. Under the definition of environmental characteristic all elements of the (built) environment fall. This therefore ranges from infrastructural factors like cycling infrastructure and traffic control installations, to types of land-use.

Characteristics such as pavement types, slopes and speed limits can all fall under infrastructural factors, just as cycling infrastructure or traffic control installations, but it is not limited to these. Also characteristics about the volumes of other traffic users are included, so for example motor vehicles or cyclists fall under this. Types of land use that are included consist of, but are not limited to, residential, commercial, industrial, greenery land use and land use mix.

Lastly, network characteristics about the connectivity or accessibility of segments are also included as these influence the volume of cyclists at segment. Under network characteristics also falls the proximity of facilities towards segments. Which environmental characteristics will actually be included in the research will be discussed in Section 2, Theoretical Framework.

Based on the aim of the study, some characteristics can also already be excluded. As the aim of the study is to predict the volume of cyclists at a segment and not the volume of cyclists taking a route, characteristics about routes can be excluded. For this reason, the route length and the amount of turns on a route are excluded from the study.

Next to this, to keep the study feasible temporal and weather related factors are not taken into account, as this would potentially make the study too broad to carry out in the time set for it.

1.4.2 Road segment

A road segment is a bounded section of the road network between two intersections. This division is used, as each segment can have unique characteristics.

1.4.3 Spatial planning approach

With a spatial planning approach, a modelling approach is meant that uses environmental characteristics with a focus on a network analysis to predict cyclist volumes and does not incorporate the shortest route between origins and destinations.

1.4.4 Traffic engineering approach

With a traffic engineering approach, a modelling approach is meant where the four step model is used to predict cyclist volumes, based on the shortest route between an origin and destination.

1.4.5 Study Area

The research will be a case study on the city of Apeldoorn in the Netherlands. The city of Apeldoorn and its surrounding villages can be seen in Figure 1 below. This study will only consider the city of Apeldoorn, including the segments just outside it, but not the entire municipality or the surrounding villages.

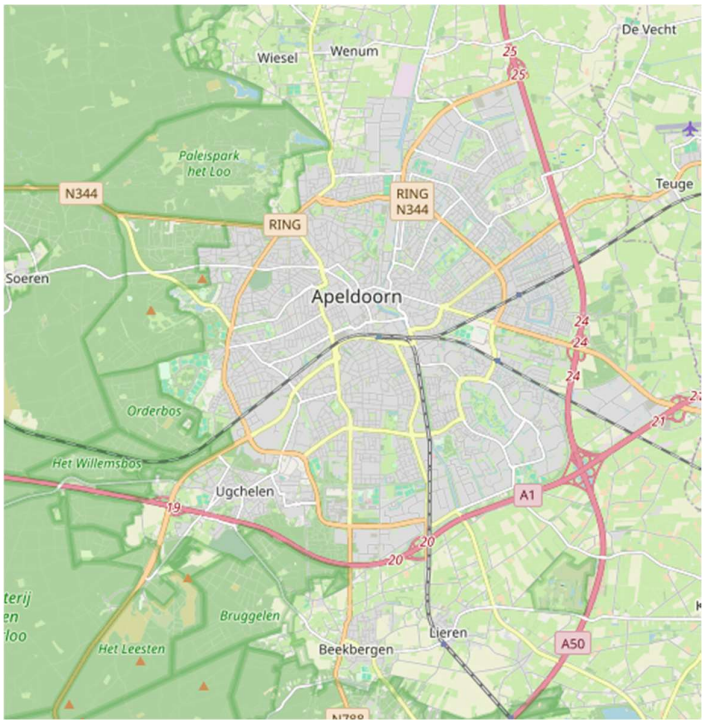

Figure 1: The city of Apeldoorn, the Netherlands (OpenStreetMap, n.d.)

1.5 Research Questions

Following from the scope and the aim of the research, to predict the volume of cyclists at road segments based on the environmental characteristics of the segments, using a spatial planning approach to create a cycling model, on a case study on the city of Apeldoorn and comparing these results with a traditional traffic engineering four-step model to determine what the differences and similarities between the different types of models are, the research questions are set up.

There are four different research questions. The questions are all shortly explained below and the full methodology can be found in Chapter 3.

Research Question 1:
Which environmental characteristics affect cyclists route choice according to literature?

The first research question is about analysing which environmental characteristics affect cyclists and should therefore be taken into account in the model. This will be done through a literature review.

Research Question 2:
What modelling approaches are appropriate to predict the volume of cyclists at segments with a spatial planning approach?

After it has been determined from the first research question which factors affect cyclist route choice, it is useful to analyse which modelling approaches could be used. For this it is important that the chosen approach takes into account the factors that were found relevant in the first research question, gives a clear output which predicts the volume of cyclists at the segments and uses a spatial planning approach.

Research Question 3:
What is the accuracy of the model, when comparing its results with cyclist counts?

After the model has been constructed, the accuracy of the model needs to be determined. This will be done by comparing the predictions of the model of cyclists volumes to cyclists counts of Apeldoorn that were not used to calibrate the model.

Research Question 4:
What are the differences and similarities in results between a model with a spatial planning approach and a model with a traffic engineering approach?

Finally, with the results of the developed model known and validated, a comparison between the results of the developed model with a spatial planning approach and another model that has a traffic engineering approach and the four step model as its core can be made. This will be done to see what the differences and similarities between them are and how they can potentially complement each other.

1.6 Report outline

This report is structured in different chapters. In Chapter 2 Theoretical Framework, previously carried out studies about factors affecting cyclists route choice and relevant characteristics for a spatial planning approach are discussed, just as regression modelling approaches that could be used in the model. Furthermore in Chapter 3 Methodology, the procedure for the construction of the model is described, just as the validation and comparison methods. In Chapter 4 Results, the outcomes of the model, validation and comparison are presented. Next to this is Chapter 5 Discussion, where the shortcomings and limitations of the research are discussed, followed by the Conclusion in Chapter 6 and finally the Recommendations for future research being presented in Chapter 7.

2. Theoretical Framework

In this chapter the current state of the literature on environmental characteristics affecting cyclists is discussed. This is first done in the context of route choice analysis, as this type of research is commonly carried out. Next to this, also characteristics that affect cyclists from a spatial planning perspective are discussed. Finally, different regression modelling approaches are discussed to determine which approach is the most suitable for this research.

2.1 Environmental characteristics affecting cyclists route choice

In literature a lot of research has already been conducted with regards to the preferences of cyclists in their route choice based on environmental characteristics. For this type of research, both stated and revealed preference studies are used. In stated preference studies, participants are asked to rank their preferences for different environmental characteristics in a route in a simpler version, or to choose between a set of routes in a more complicated version of a stated preference study (Yang & Mesbah, 2013). In a stated preference study always a hypothetical route or scenario is studied. In revealed preference studies, GPS data used to study the actual choices of cyclists (Prato et al., 2018). There are some downsides to stated preference studies, as it can be difficult for participants to correctly recall their actual choices, making the results from these studies potentially inaccurate (Yang & Mesbah, 2013). With the recent increasing availability of GPS data such as 'Fietstelweek' data, most research nowadays that is conducted is uses data that was gathered through a revealed preference approach and these are also mainly studied in this theoretical framework.

In studies regarding cyclists route choice, discrete choice models such as multinomial logit, mixed logit or path size logit models are often used to estimate the effects of different characteristics on cyclists (Khatri et al., 2016; Koch & Dugundji, 2021; Meister et al., 2023 & Patro et al., 2018). Using this approach, researchers determine which route characteristics have a positive effect on cyclists choosing a route and which characteristics have a negative effect.

Several studies concluded that segregated cycling infrastructure, cycling paths and also suggestive, painted cycling lanes on roads that are shared with cars, are preferred by cyclists on their routes (Chen et al., 2016; De Jong et al., 2023; Khatri et al., 2016; Koch & Dugundji, 2021; Lukawska et al., 2023; Meister et al., 2023; Prato et al., 2018 & Van Nijen et al., 2024).

Next to this, in several studies researchers found that cyclists prefer flatter roads over roads that go uphill (Prato et al., 2018; de Jong et al., 2021; Meister et al., 2023; Lukawska et al., 2023 & Chen et al., 2017).

It was also concluded that cyclists prefer asphalt paved roads to cycle on (van Nijen et al., 2024; Prato et al., 2018 & Lukawska et al., 2023), whilst roads that are paved with stones paved are less preferred (Lukawska et al., 2023, Van Nijen et al., 2024).

About the influence of signalized traffic control installations on cyclists, there was no real consensus in the literature. Khatri et al. (2016) and van Nijen et al. (2024) found a positive effect, whilst other studies found a negative effect and that cyclists tended to avoid signalized intersections on their routes (Koch & Dugundji, 2021; Meister et al., 2021 & Prato et al., 2018).

Khatri et al. (2016) & van Nijen et al. (2024) had not expected this preference of cyclists for traffic control installations, due to delays associated with them, but it was explained by Khatri et al. (2016) that this could be due to infrastructural nature of their case study city Phoenix, United States, as signalized intersections could provide safe crossing over large roads, just as that the fact that the downtown area of Phoenix is highly signalized, leaving cyclists no choice. It was mentioned by van Nijen et al. (2024) that the found preference for traffic control installations in their case study of Enschede, the Netherlands could be (partially) correlated with the found preference for the ring road of Enschede which was presumably preferred for easy navigation, as the ring road also has a high number of traffic control installations.

The study of Meister et al. (2021) found that cyclists prefer roads that have a speed limit of 30km/h over roads with speeds limits of 50km/h in their study about Zurich, Switzerland. The study of De Jong et al. (2023) found that cycling demand is greater along roads with greater speed limits, but they mentioned that this could be due to a correlation of roads with a high connectivity and intensities to also often have greater speed limits. Next to this, the study of Prato et al. (2018) mentions that cyclists prefer routes along roads that have no more than 2 lanes for motorized traffic. Whilst the study of Prato et al. (2018) did not directly address speed limits, this finding can be seen as a proxy, as roads with fewer lanes often have low speed limits and vice versa. Lastly, the study of Jestico et al. (2016) found that cycling volumes are the greatest on roads with a 30km/h speed limit, showing a clear preference of cyclists.

Besides these infrastructural factors influencing bicycle route choices, also volumes of both motorized traffic and other cyclists on roads can have an impact on cyclists. A study from Uijtdewilligen et al. (2024) found that cyclists are affected by the crowdedness of other cyclists on their routes. Cyclists prefer routes that are not crowded, as crowded routes are perceived as less safe and therefore less preferred by cyclists. The type of cycling infrastructure that is used also effects the degree to which crowdedness is perceived. When crowdedness is high, separated cycling infrastructure becomes less preferred over mixed traffic conditions. Uijtdewilligen et al. (2024) mention that this could be the case due the fact that separated cycling paths are sometimes too narrow to handle high crowdedness, whereas mixed traffic conditions could provide more space for deviations.

Next to this, a majority of studies found that higher motorized traffic intensities also have a negative effect on cyclists (Khatri et al., 2016; Meister et al. 2023; Prato et al., 2018 & Uijtdewilligen et al., 2024). Next to this the study of Koch & Dugundji (2021) found that cyclists take detours to avoid roads with (high) levels of noise pollution coming from motorized vehicles. This also shows through an indirect relation that cyclists prefer roads that are not crowded with motorized traffic. An exception to this was the study of Van Nijen et al. (2024), as they found a positive effect of both heavy and medium motorized traffic intensities on cyclists. This was also assumed to be caused by the preference of cyclists in Enschede for the ring road, as there are heavy car intensities on the ring road.

Also several land-use zones were found to be an influence to the route choice of cyclists. It was found that a residential land use zone has a negative influence on cyclists (Koch & Dugundji, 2021; Prato et al., 2018 & Van Nijen et al., 2024). Next to this commercial land uses are preferred by cyclists, as was found by Koch & Dugundji (2021) Van Nijen et al. (2024), whilst industrial land uses are not preferred (Prato et al., 2018 & Van Nijen et al., 2024). The studied literature is not conclusive on the effect of greenery land uses. The studies of De

Jong et al. (2023) & Van Nijen et al. (2024) found that greenery lands uses are not preferred by cyclists on their routes. It was mentioned by De Jong et al. (2023) that this could be caused by difficulties with navigation through parks and forests. On the other hand, several studies have concluded that greenery land uses do attract more cyclists (Koch & Dugundji, 2021; Lukawska et al. 2023 & Prato et al., 2018). Studies were also not in agreement with about the effect of a mixed land use, as Van Nijen et al. (2024) found a negative effect, whilst Chen et al. (2018) found a positive relation.

Lastly, some studies also analysed whether segments that were water facing had any effect on cyclists route choices, where a clear positive effect was found (Chen et al., 2017; De Jong et al., 2023; Koch & Dugundji, 2021 & Lukawska et al., 2023).

Besides different types of land use, also the density around a segment can effect cyclists. It was found by Van Nijen et al. (2023) & Prato et al. (2016) that a greater degree of urbanisation attracts more cyclists or is more preferred by them, whilst the studies of De Jong et al. (2023) & Lukawska et al. (2023) found that cyclists prefer routes that do not pass through highly urbanised areas. This is elaborated upon by De Jong et al. (2023), as they mention that high urban areas, or areas with a high population and employment density do attract a lot of traffic as the amount of origins and destinations of trips are high there, but they are not seen as attractive areas to cycle through for those who pass by them. This influence of residential and employment densities on cyclists is supported by Knijnenburg (2021), as he found that they both influence cyclist route choices.

2.2 Environmental characteristics from a spatial planning approach

Traditionally, spatial planning approaches are mainly used to predict pedestrian intensities throughout cities. These approaches usually consist of analysing environmental characteristics, with a major focus on the network characteristics. The network is often analysed through a spatial-configurational approach, where topological features of the street network are analysed (Lerman et al., 2014). In the studies that were analysed this was often done via Space Syntax, which analyses the urban configuration (Hillier, 2007). The study of Raford and Ragland (2006) also argued that Space Syntax network analysis is the most suitable method on an urban scale, so this method will be explored further.

In a Space Syntax analysis the features of street network are analysed to determine the accessibility of all separate segments. For this an axial map of the network is created to show the lines of sight in the network and based on that the topological relationships of the segments can be determined (Drolsbach, 2022 & Lerman et al., 2014). It analyses how every segment relates to all other segments (Van Nes & Yamu, 2021). An axial map of an example city can be seen in Figure 2b, alongside the map of the actual town in Figure 2a. From the axial map the connectivity of a street, which is the number of connections to other streets it has, can easily be determined (Drolsbach, 2022). Using the connectivity, the integration of a street can be determined and seen in Figure 2c, as the integration showcases the amount of directional changes needed to reach all other streets in the network (Drolsbach, 2022, Lerman et al., 2014 & Van Nes & Yamu, 2021). The integration of the main street A is very high, as from there all other streets can be reached easily without the need for much turns.

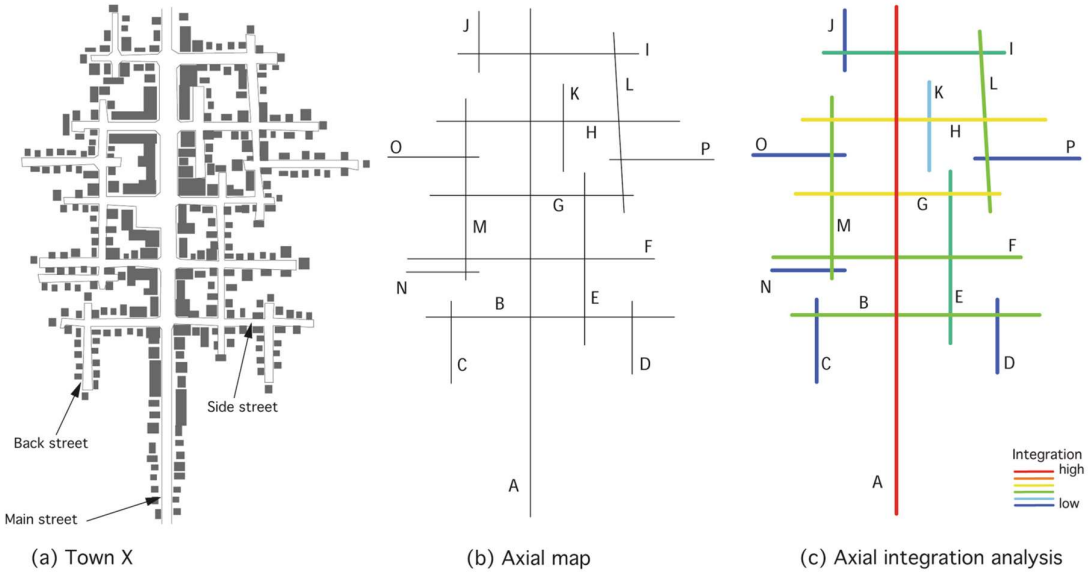

Figure 2: Representation of the transformation of the map of a town (a), towards an axial map (b) and an axial integration analysis (c), (Van Nes & Yamu, 2021)

Besides the integration measure, also angular choice is an often used index to show the network connectivity. The angular choice of a street depicts how often it is present on the shortest angular routes between other segments, to show how relevant its position is in the network (Lerman et al., 2014). An adaption of this index, which accounts weights for the different angles of turns also exists. In this way, if there is only a very slight turn, it would be seen as a different effect compared to a turn of 90 degrees (McCahill & Garrick, 2008).

More measures are possible and sometimes used in studies as mentioned by Lerman et al. (2014), but integration and angular choice are the most commonly used so the analysis will be kept to those. These Space Syntax measures can be analysed on multiple radii, Van Nes & Yamu (2021) mention that most commonly a radius of 400-800m is used when analysing smaller areas, whilst 4000-8000m radii are used for analysing entire urban areas or cities. Sometimes also a combination of two different radii is used, to incorporate effects of some streets being important on a local scale, whilst not on a city wide scale. In the remainder of this section different studies who used these measures to analyse pedestrian and cyclists intensities will be discussed.

In the research of Jiang (2009), where pedestrian movements in central Londen were analysed through integration, it was found that 60% of the human movement in their study could be explained from a topological, network perspective whereas the remaining 40% of the human movement is caused by other environmental characteristics such as land uses or road widths. This finding is supported by Raford and Ragland (2004), as they found that around 55% of the predicted pedestrian intensities were correlated with the integration measure, with the remaining 45% being correlated with the population and employment densities that were taken into account in their research.

In the study of Dhanani et al. (2017) both integration and angular choice were used to analyse pedestrian volumes, along with land use mix, public transport accessibility and residential density. Through a regression model they found that integration on a scale of 2000m, or 25 minutes walking influenced pedestrian volumes the most. In addition to this, the study of De Wit et al. (2021) about the creation in the Loopmonitor in Rotterdam found that the angular choice on scales of both 400m and 2400m were highly influential for pedestrian volumes, just as the amount of facilities near a segments, the proximity to a train station and park and the width of the footpath at the segment. The factor of proximity to public transportation can be a highly relevant factor, as it was found by the Dutch Railways or NS (2023) that 39% of the travellers at the main train station of Apeldoorn arrive or leave cycling.

In the study of De Wit et al. (2021) the exact definition of what is meant with facilities is not elaborated on, except that residential land uses are not taken into account. From this it can probably be expected that buildings with a commercial land use are meant by this, but also other facilities that are not directly classified with a commercial land use could be influential. Public transport was already explicitly mentioned by De Wit et al. (2021), but educational and potentially also healthcare facilities could also have a significant influence on cyclists, as nearly 80% of high school students travel by bike (Fietsersbond, 2022). Next to this, supermarkets could potentially be extra relevant, besides their commercial land use, as Veenstra (2008) found that the majority of all trips to supermarkets with a distance of around 1 km are made by bicycle in the Netherlands, showing the vast amount of cycling traffic from and towards supermarkets.

Next to the studies in a pedestrian context, some literature also exists about Space Syntax measures in a spatial planning context on cyclists. The research of Drolsbach (2022) found that the angular choice metric on a radius of 2500m was the best suited for explaining the predicted cyclists routes in his research. A set of radii were tested, ranging up until 2500m, with each greater radius having a higher predictive value. Due to data limitations in his study the maximum radius was limited to 2500m, as otherwise a greater radius would potentially have been preferred. Besides that, Drolsbach (2022) found that land use mix, the degree of urbanisation, greenery or water facing routes, and cycling infrastructure were influential on cyclists.

In a study of Orellana et al (2019a) about cyclist volumes, it was found that a radius of 4500m for the angular choice was the most suitable. In another study Orellana et al (2019b) found that a global radius on the angular choice also worked well, as it explained over 40% of the cyclists routes in their study. In these studies several other environmental factors were also analysed and found to be relevant, with the angular choice metric being the most important. The other factors that were analysed were land use mix, density, greenery, open water bodies, cycling infrastructure and the type of pavement, just as slope, the amount of facilities and the amount of intersection at routes (Orellana et al 2019a & Orellana et al. 2019b).

Concluding, many different environmental characteristics are important from a spatial planning perspective. A network analysis with Space Syntax is the most influential factor, but there is no consensus in the studied literature whether integration or angular choice works best. The same goes for the radius on which a Space Syntax analysis should be carried out, as both smaller and greater radii were used, whilst the study of De Wit et al. (2021) used both. Besides the network analysis, land use factors were found to be relevant in addition to the proximity to different facilities such as train station and education.

2.3 Visualisation of relevant characteristics

After researching relevant characteristics from both a route choice and spatial planning context, the final determination of all relevant characteristics that will be included in the model can be made.

From the characteristics that were found to be relevant, the slope is excluded. This is chosen for as the research is about segments, where it is impossible to include the slope in as it is unknown which direction one is travelling on two-way segments. The characteristic is mainly relevant in a route choice context and is therefore not applicable to this research. Next to this also the characteristics about population and employment density are excluded from the research, as no proper data course could be found for the employment density and the degree of urbanisation already takes into account the expected effect at least partially. Lastly, also the volume of cyclists is removed from the research. This is chosen as including the characteristic could give a too clear pattern of where cyclist intensities are great and where not to the model, which could have a negative effect on the research aim to evaluate if a spatial planning approach could predict cyclist intensities.

The characteristics that will be included in the remainder of this research can be seen in Table 1, along with their expected effect on cyclists from the literature review.

Table 1: Relevant characteristics

Characteristic	Influence	Sources
Infrastructural		
Segregated cycling path	+	Chen et al. (2018), De Jong et al. (2023), Khatri et al. (2016), Koch & Dugundji (2021), Lukawska et al., 2023), Meister et al. (2023), Orellana (2019b), Prato et al., (2018), Van Nijen et al. (2024)
Cycling lane	+	De Jong et al. (2023), Koch & Dugundji (2021), Lukawska et al., 2023), Meister et al. (2023), Prato et al., (2018), Van Nijen et al. (2024)
Pavement type asphalt	+	Lukawska et al., 2023), Prato et al., (2018), Van Nijen et al. (2024)
Pavement type paving stones	-	Lukawska et al., (2023), Van Nijen et al. (2024)
Traffic control installations	+ / -	Khatri et al. (2016), Koch & Dugundji (2021), Meister et al. (2023), Van Nijen et al. (2024)
High speed limits	-	De Jong et al. (2023), Jestico et al. (2016), Meister et al. (2023), Prato et al. (2018)
Traffic		
Motorized vehicle intensities	+ / -	Koch & Dugundji (2021), Khatri et al. (2016), Meister et al. (2023), Prato et al., (2018), Uijtdewilligen et al., (2024), Van Nijen et al. (2024)
Crowdedness of cyclists	-	Uijtdewilligen et al. (2024)
Land-use		
Residential land-use	-	Koch & Dugundji (2021), Orellana (2019a), Prato et al., (2018), Van Nijen et al. (2024),
Commercial land-use	+	Koch & Dugundji (2021), Van Nijen et al. (2024)
Greenery land-use	+ / -	De Jong et al (2023), Drolsbach (2022), Koch & Dugundji (2021), Lukawska et al. (2023), Prato et al., (2018), Van Nijen et al. (2024)

Industrial land-use	-	Prato et al., (2018), Van Nijen et al. (2024)
Water facing	+	Chen et al. (2018), De Jong et al (2023), Drolsbach (2022), Koch & Dugundji (2021), Lukawska et al. (2023)
Land-use mix	+ / -	Chen et al. (2018), Dhanani (2017), Orellana (2019b), Van Nijen et al. (2024)
Degree of urbanisation	+ / -	De Jong et al. (2023), Drolsbach (2022), Lukawska et al. (2023), Prato et al. (2018) , Van Nijen et al. (2024)
Network		
Network accessibility or connectivity	+	De Wit et al. (2021), Dhanani (2017), Drolsbach (2022), Jiang (2009), Lerman et al. (2014), McCahill & Garrick (2008), Orellana (2019a), Orellana (2019b), Raford & Ragland (2006), Van Nes & Yamu (2021)
Proximity to train station	+	De Wit et al. (2021), Dhanani (2017), NS (2023)
Proximity to educational facilities	+	De Wit et al. (2021), (Fietsersbond, 2022)
Proximity to hospital	+	De Wit et al. (2021)
Proximity to supermarkets	+	De Wit et al. (2021), Veenstra (2008)

Using all the gathered information from literature, a visualisation of the state of the available knowledge can be set up. Figure 3*Figure* below illustrates the state of the available knowledge, divided in to which characteristics have a positive or negative influence on the perception of cyclists, with also some characteristics having an unclear influence.

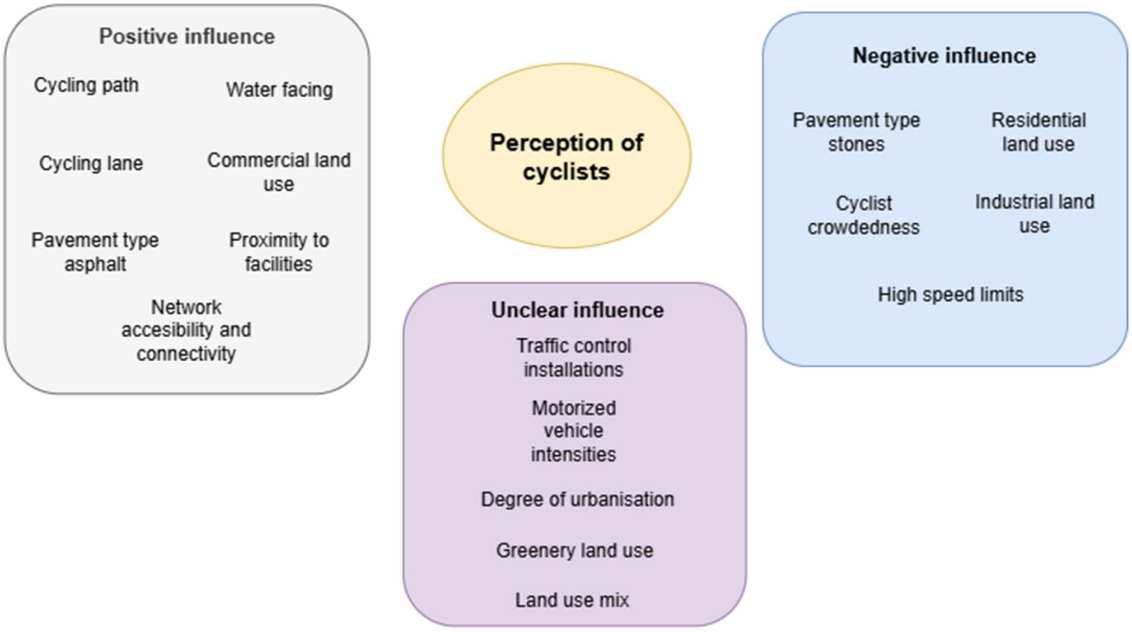

Figure 3: Visualisation of the relevant characteristics and their expected influence

2.4 Regression modelling approaches

In the model, a regression approach will be used to predict the cyclist intensities on all segments present in the study area of the research. Several different regression could be used for this and these will be explored in this section to determine which regression approach is the most suitable for this study.

First of all, the most simple and often used regression approach is (multiple) linear regression, due to it being easily understandable. Multiple linear regression is used to predict an independent variable (y), from several dependent variables (x). This is illustrated in equation 1 below, where β describes the coefficients of the dependent variables and ε describes the error. The β of the dependent variables can be determined through Ordinary Least Squares (OLS) analysis, which minimizes the sum of the squared differences between the observed values and the predicted values of the model (GeeksforGeeks, 2025a).

An important assumption of linear regression is that the relationship between the dependent and independent variables is linear, which can cause problems when many different dependent variables are used. Next to this, linear regression is also sensitive to outliers in the data and can be prone to overfitting.

Equation 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

Linear regression models have as a big advantage that their working is simple and very understandable, but they function less good with large and more complex datasets. In these datasets linear regression models can function less optimal in capturing the nonlinear relationships between the different features (Yang et al., 2025). Additions to linear regression to improve its performance such as Ridge, Lasso or Elastic-Net regression exist, but these still have the same underlying assumption that the relationships are linear. Other regression models, such as Random Forest regression models can handle these large datasets better and can increase the accuracy of a model. (Yang et al., 2025). Therefore, Random Forest regression will be explored in more detail.

In Random Forest regression decision trees with a certain depth are used to make predictions. All trees are given different subsets with features of the entire dataset to make its own predictions. All these predictions are later on combined through averaging to create the final predictions of the model. Using this random data selection for the different trees, overfitting on the train data is in general prevented which improves the accuracy of the overall model (GeeksforGeeks, 2025b). Random Forest regression was also used by De Wit et al. (2021) in their spatial planning model of pedestrians in Rotterdam, so this approach could also align very well with the objective of this research.

As both linear regression and Random Forest could be suitable for the purpose of this research and have their advantages, both models are evaluated to determine which has the best fit and can predict cyclist intensities the best.

3. Methodology

In this section the entire methodology of research will be discussed. This is first done with regards to the operationalisation of the characteristics and construction of the model. After the model construction has been discussed, the process of the validation and comparison is presented.

3.1 Methodological framework

For the research a methodological framework is set up, to illustrate clearly what the research is about and what steps will be taken. The conceptual framework can been seen in Figure 4 and there the 5 different components of the methodology are illustrated; literature study, data gathering, model construction, validation and application.

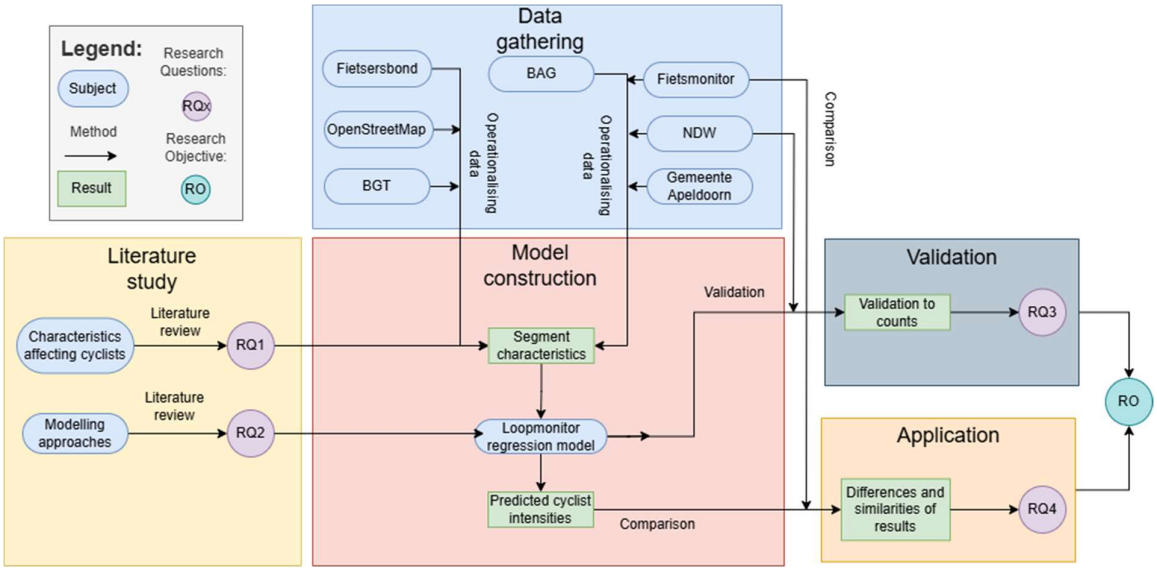

Figure 4: The methodological framework of this research

Using the literature study, which can be read in Section 2, it has been determined which characteristics are relevant to include and which modelling approaches can be used for the model. All of the relevant characteristics are operationalised throughout different data sources, which will be discussed in the next section.

3.2 Data

To operationalise the characteristics, data is needed. For all characteristics the method of operationalisation within the model and the corresponding source can be seen in Table 2*Table* . These data sources will be discussed the in the remainder of this section.

Table 2: The characteristics and their data sources

Characteristic	Method of operationalisation	Data source
Infrastructural		
Segregated cycling path	Binary variable present or not	*(Fietsersbond, 2021)*
Cycling lane	Binary variable present or not	*(Fietsersbond, 2021)*
Pavement type asphalt	Binary variable present or not	*(Fietsersbond, 2021)*
Pavement type paving stones	Binary variable present or not	*(Fietsersbond, 2021)*
Traffic control installations	Binary variable present or not within buffer around traffic control installation	*(Fietsersbond, 2021)*
Speed limits	Categorical, 5 categories	*(Fietsersbond, 2021)*
Traffic		
Motorized vehicle intensities	Categorical, 3 categories	*(Gemeente Apeldoorn, 2023)*
Crowdedness of cyclists	Ratio of intensity and capacity, between 0 and 1	*(CROW, 2006 & Witteveen+Bos, 2023)*
Land-use		
Residential land-use	Continuous ratio of buffer area, between 0 and 1	*(PDOK, 2025)*
Commercial land-use	Continuous ratio of buffer area, between 0 and 1	*(PDOK, 2025)*
Greenery land-use	Continuous ratio of buffer area, between 0 and 1	*(PDOK, n.d.)*
Industrial land-use	Continuous ratio of buffer area, between 0 and 1	*(PDOK, 2025)*
Water facing	Continuous ratio of buffer area, between 0 and 1	*(PDOK, 2025)*
Land-use mix	Continuous ratio of buffer area, between 0 and 1	*(PDOK, 2025 & PDOK, n.d.)*
Degree of urbanisation	Categorical, 5 categories	*(CBS, 2025)*
Network		
Network accessibility and connectivity	Continuous value	*(depthmapx, 2020)*
Proximity to train station	Continuous value	(*OpenStreetMap*, n.d.)
Proximity to educational facilities	Continuous value	(*OpenStreetMap*, n.d.)
Proximity to hospital	Continuous value	(*OpenStreetMap*, n.d.)
Proximity to supermarkets	Continuous value	(*OpenStreetMap*, n.d.)
Cyclist counts		
Count data	Continuous value	(NDW, 2025)

For the infrastructural characteristics the Fietsersbond (2021) data is used. The Fietsersbond provided a very detailed dataset with the cycling network of Apeldoorn and its surrounding area. The high level of detail makes the network very convenient to implement, but as the data is from 2021, the potential outdatedness should be taken into account when making conclusions.

Next to this, traffic models from the municipality of Apeldoorn and Witteveen+Bos are used to estimate the motorized vehicle and cyclists intensities. Using guidelines on capacity from the CROW, the crowdedness of segregated cycling paths can also be determined.

Data on land use types comes from the BAG (Basisregistratie Adres Gegevens) and BGT (Basisregistratie Grootschalige Topografie), both distributed by PDOK. Next to this, data on the degree of urbanisation comes from the ‘Wijken en Buurten data’, provided by the Central Bureau of Statistics (CBS). All these data sources are seen as trustworthy, as they are provided by the government of Netherlands and are highly detailed.

For the network analysis with Space Syntax, the depthmapx software is used. The exact working and implementation of this program will be described in Section 4 Methodology. Next to this, OpenStreetMap information will be used to determine the proximity of segments towards different facilities. In OpenStreetMap a vast amount of data is available about facilities, making it very useful for this. This data is all available because OpenStreetMap is an open source platform, allowing anyone to contribute. This however has the side effect that the data is not validated, adding an uncertainty to the results of the model.

Lastly, whilst not a data source related to the characteristics of the segments, the count data also needs to be discussed. For the cyclist counts in Apeldoorn, data from the Nationaal Dataportaal Wegverkeer (NDW, 2025) is available. At 159 locations counts were gathered for several periods since 2020 and the count locations can be seen in Figure 5. The periods that the counts were gathered differ per location, from just 2 weeks to over multiple months. For all of these locations the cyclist intensities on a daily basis are used. The count data is overall seen as trustworthy, as the data is validated by the NDW.

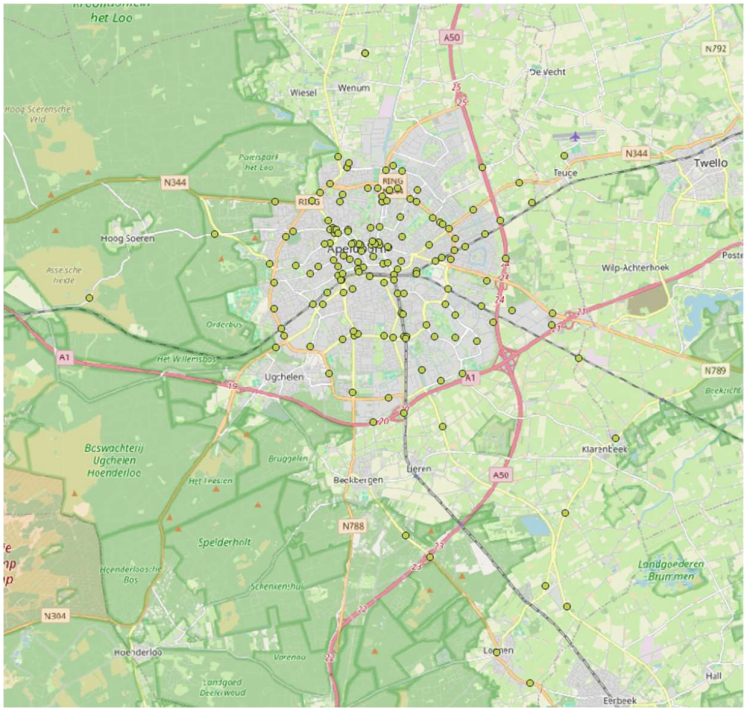

Figure 5: The 159 count locations in and around Apeldoorn (Background taken from OpenStreetMap, (n.d.))

3.3 Operationalising characteristics

In this subsection the methods of operationalisation used for all the characteristics of the segments will be discussed.

3.3.1 Infrastructural characteristics

For the characteristics about cycling infrastructure or paving types very few alterations were needed, as the Fietsersbond data had this information readily available. The only step that was needed was to determine which subcategories of the Fietsersbond data corresponded to cycling path and cycling lane or paved and unpaved, as the data was very detailed.

Traffic control installations

The traffic control installations were linked to the segments by creating a buffer of 50 meter around the installation and checking which segments lay within that buffer. The segments were then divided via a binary approach.

This method only takes into account segments directly adjacent to a traffic control installation and not segment further away from the segment, but directed towards an installation. This is a downside of the method, as cyclists who might potentially take a detour to avoid a traffic control installation, would probably also avoid earlier segments in the direction of the traffic control installation. This is not taken into account in the model and should be considered when analysing the results.

Speed limits

The speed limits are included in categorical manner and also only for roads that are shared by cyclists and motorists, as the impact of the speed of motorized traffic is assumed to be minimal for cyclists on a segregated cycling path. The speeds are categorised into 15, 30, 50, 60 and 80km/h limits, and linked accordingly to the relevant segments.

3.3.2 Traffic characteristics

Motorized vehicle intensities

Of the motorised vehicle network the roads are linked to the nearest segment of the cycling network available, with a maximum distance of 10 meters. Using this method, all roads that are occupied by motorised vehicles are linked to a nearby cycling network segment if present, ensuring that any effect of motorised vehicle intensities are taken into account. These volumes are then categorised into three categories. This division can be seen in Table 3.

Table 3: Categorical division of motorized vehicle intensities

Motorized vehicle intensities	
Intensity (Number per day)	Classification
Less than 500	1
Between 500 and 3,000	2
More than 3,000	3

Crowdedness of cyclists

For segregated cycling paths, the capacity of the path can become important to ensure a great enough overtaking possibility exists for cyclists. The capacity of a cycling path is decided by its width and guidelines from the CROW (2022) for widths at certain intensities per peak hour can be seen in Table 4. Using the width of a cycling path, its capacity can be calculated which can then be put in a ratio with its intensity. Here a ratio of 1.0 would describe a path with sufficient capacity, e.g. an intensity of 400 with a capacity of 750, whilst 0.5 would describe a capacity of 150 with an intensity of 300. These intensities are all for the intensity in a peak hour, but as in the rest of this study the cyclist intensities are analysed on a daily basis this must be converted. For this a rule of thumb from the Fietsberaad (2023) is used, which states that the peak hour intensities are 15% of the daily intensities. With this conversion the crowdedness can be operationalised in the model.

Table 4: Intensity per peak hour and width recommendations for segregated cycling paths (CROW, 2022)

Intensity 1 way path	Width	Intensity 2 way path	Width
0 - 150	2.30m	0 - 75	2.60m
150 - 250	2.50m	75 - 150	2.70m
250 - 500	2.70m	150 - 500	3.60m
500 - 700	3.30m	500 - 700	4.40m
700 - 900	3.50m	700 - 900	4.80m
> 900	3.60m	> 900	5.20m

3.3.3 Land use characteristics

Land uses

For all the different land use types, residential, commercial, industrial, greenery and water a 250 meter buffer was made around each segment of the network. The amount of area covered by each land use within this buffer was calculated and put in a ratio over the total area to determine the respective influence of each land use.

Land use mix

To calculate the land use mix around the segments, the Shannon Index was used and the formula for this index can been below in Equation 2 (Bobbitt, 2022).

Equation 2:

$$S = \frac{-\sum_k p_i * \ln(p_i)}{\ln k}$$

In this equation k is the total number of land use classes taken into account and p_i is the total area of each land use type in the buffer. This equation results in a ratio between 0 and 1 of the land use mix, where 1 describes a very good mix between the all present land uses. (Bobbitt, 2022).

Degree of urbanisation

The degree of urbanisation describes the amount of addresses per km^2 in a categorical manner. The data is already divided into 5 different categories by the data provided CBS on a neighbourhood basis. This division can be seen in Table 5.

Table 5: Degree of urbanisation categories (CBS, 2025)

Address density per km^2	Degree of urbanisation	Category
More than 2,500	Very high urbanisation	1
Between 1,500 and 2,500	High urbanisation	2
Between 1,000 and 1,500	Moderate urbanisation	3
Between 500 and 1,000	Low urbanisation	4
Less than 500	Not urbanised	5

3.3.4 Network characteristics

Network accessibility and connectivity

The network analysis will be done using the depthmapXnet QGIS plugin (2020), where the integration and angular choice will be calculated. The integration (NAIN) stands for the direct connectivity of a street, as it is calculated by how much turns are needed to reach other segments from there. The angular choice (NACH) on the other hand in calculated by analysing how often the segment is passed through when connecting other segments. In Figure 6 an example of the results of the calculation of NACH and NAIN is shown. Here it can be seen that streets which are often travelled through to reach other streets have a high NACH and streets that have a high direct connectivity have a high NAIN.

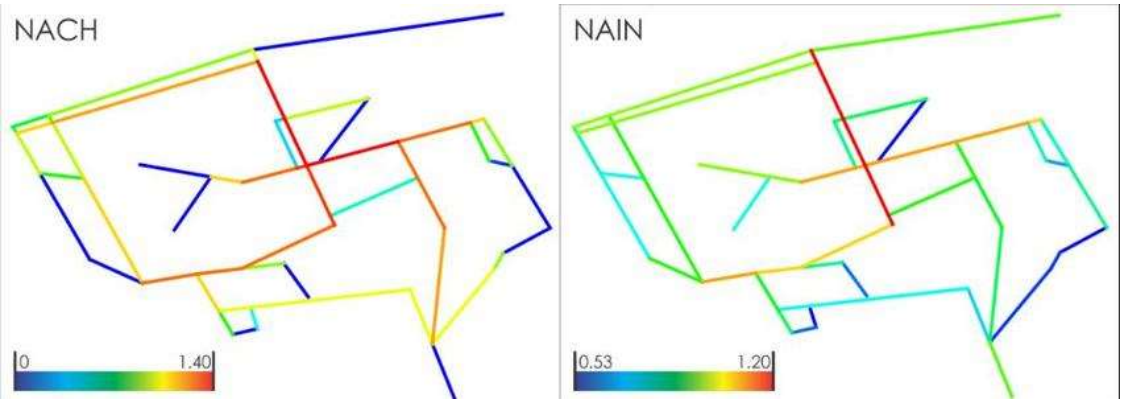

Figure 6: Example of NACH and NAIN (Ourique et al., 2017)

These measures can be calculated on several scales, depending on which scale is the most relevant. For this reason multiple scales will be calculated, after which the most relevant ones will be determined. The scales that will be calculated are 400m, 800m, 1200m, 2000m, 3000m, 3600m, 4800m and 7200m and the entire network. These measures will all be calculated on a normalised scale, which means that the angle of the turn used is taken into account.

In addition to these values, also the angular connectivity of each segments is calculated, which analyses just how many other segments can be reached directly from each segment, so not on a radius. The angular connectivity is then also corrected to take into account the angle of each turn, where greater angels are penalised.

All these different scales will be compared and analysed for correlation and importance, to determine which should actually be included in the model. These values will be included in the model on a continuous scale.

Proximity to facilities

The proximity to educational facilities, train stations, hospitals and supermarkets is determined by taking information about all relevant facilities in Apeldoorn from OpenStreetMap (n.d.). From all segments the distance to these facilities will be calculated as the crow fly flies, and not via a network route to ease calculations. For train stations also a division is made between the major train station of Apeldoorn and its smaller stations, Apeldoorn Osseveld and Apeldoorn De Maten as these two stations get significantly less travellers on a daily basis than the main station (NS, 2023).

3.3.5 Cyclist counts

Besides all the environmental characteristics that need to be linked to the segments, also the cyclist counts need to be made ready to include in the model. During this process some problems with the count data became clear.

For example, as can be seen in Figure 7 in red, some count points were found to be in between two roads, making it unclear to which segment the count should be linked. This was handled by removing the count point, to ensure that it would not be linked to the wrong segment and creating incorrect results.

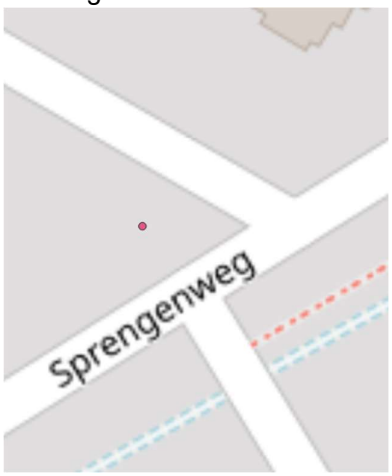

Figure 7: Count point in red exactly between two roads (Background taken from OpenStreetMap (n.d.))

Next to this, count points were also sometimes found to be in the middle of the motorized traffic road, exactly between the cycling paths on both sides as in Figure 8. These were handled by creating two points on the cycling paths and giving both of them half of the count value and deleting the original point. In the end these processes resulted in 189 count points being used, compared to the original 159, where some count points also had to be deleted as they were outside of the scope of the city of Apeldoorn.

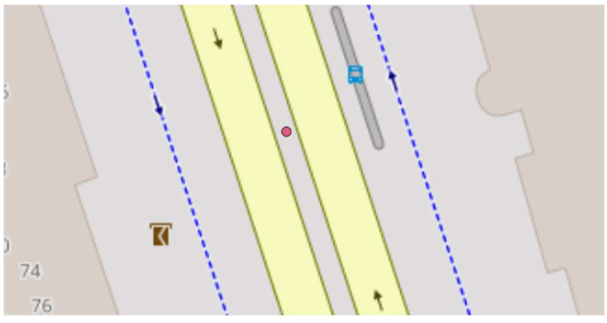

Figure 8: Count point in red in the middle of a road, in between two cycling paths (Background taken from OpenStreetMap (n.d.))

After all counts were properly linked, the values were analysed to see how the counts were spread out. This can be seen in Figure 9. Here it can be seen that there are some count values that have a significantly larger values than the rest of the counts, with the majority of the counts being spread around the 1000 to 2000 count value. This shows that that are few counts at locations with very few cyclists and also very few counts at locations with many cyclists.

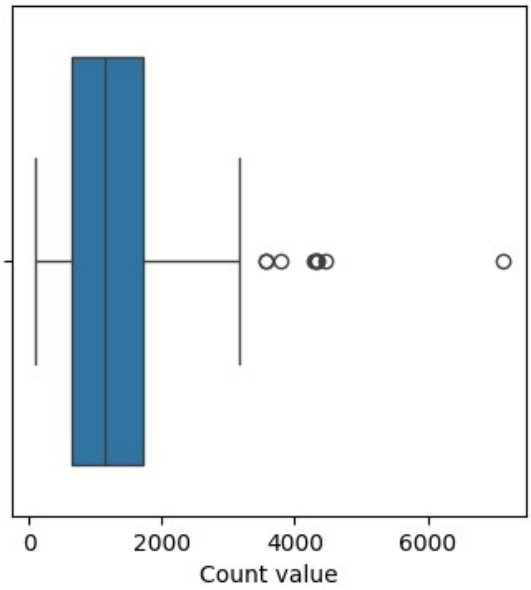

Figure 9: Spread of the count values

3.4 Model construction

In this section the construction of the regression model that will be used to predict the cyclist intensities is explained. The already existing Loopmonitor of Witteveen+Bos is used for this, as this model already has a nice base for predicting pedestrian intensities based on a spatial planning approach. In the Loopmonitor model, first the correlation scores of the different features are calculated, to determine whether some features have to be excluded from the analysis. After this, several different types of regression models are iterated through and evaluated to determine which model is the most accurate and should be used for the final prediction. Lastly, the cyclist intensities are predicted and exported to be used for further analysis.

In this study the Loopmonitor will be used, but with some adaptations and additional features. As was explained in Section 2.4 a linear regression and Random Forest regression models are used in this study, so these two models are evaluated. Based on the criteria of the validation, as is described in Section 0, the type of regression model to use is determined.

Random Forest regression models have different parameters that can be changed to optimise the performance, so these parameters will be evaluated to determine the most optimal configuration with the highest accuracy. This is specifically for Random Forest regression models, as their greater complexity also allows for more configurations. Besides this a feature analysis is conducted on the network accessibility and connectivity features, to determine which features are relevant to include in the model and which should be excluded from the model to reduce the amount of network accessibility and connectivity features in the model, to just the important features.

All of the different steps that are taken in the model are explained in detail in the remainder of this section.

3.4.1 Correlation analysis

Before the model can make predictions, it needs to be analysed if correlation between two different features exists. If a high correlation exists between two variables, the same effect could potentially be described twice in the model, skewing the results.
To prevent this, a correlation analysis is carried out. Using this analysis it can be determined if there are highly correlated features, and if so one of them can be taken out. The correlation will be analysed using the Pearson correlation coefficient, which can be seen in Equation 3.

Equation 3:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Using this formula, the correlation between the two characteristics can be determined. If the *r* coefficient has a value of -1 or 1, it describes a perfectly negative or positive correlation, whist a coefficient of 0.5 describes a moderate positive correlation (Akoglu, 2018).
The threshold for when a correlation is too great that a variable should be excluded differs per field of study as is mentioned by Akoglu (2018). For this research a coefficient of -0.8 or 0.8 is chosen, which is supported by Knijnenburg (2021) who also studied the effect of environmental characteristics on cyclists.

3.4.2 Model parameters

In a Random Forest regression model there are several parameters that determine the inner working of the model. As was already explained in Section 2.4, a Random Forest regression model uses trees with a certain depth to train itself. The standard values for the amount of trees is 100, with no maximum depth existing for the trees.

To get the most accurate and optimal predictions from the model the amount of trees, also known as the amount estimators of the model and the maximum depth of the trees, should be optimized. This is done by iterating over different values to find the optimal value with the smallest Root Mean Square Error. For the amount of trees it will be iterated over intervals of 10 until 250 trees, with the max depth being evaluated for each depth up until 22. These parameters are used for the optimal configuration.

3.4.3 SHAP feature importance analysis

To analyse which features are the most important and have the most influence on the final predictions of the model, SHAP values will be used. SHAP stands for SHapley Additive exPlanations, and is a concept from game theory that can be used to interpret predictions from regression models (Yamaguchi, 2020). SHAP values show how much each feature contributes to the final predictions of the model, offering interesting insight into how the model operates.

3.5 Validation

With the model constructed, the accuracy of the model can be determined. For this, 20% of all counts are used to test the model, with the remaining 80% being used as train data.

The model will be evaluated on multiple metrics, which will be explained below. The model is evaluated on both the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) metrics to determine the accuracy of the model compared to the actual cyclist counts. The equations for the MAE and RMSE can be seen in below in equations 4 and 5. Here y_i are the predicted intensities, and x_i are the actual counts in the test set (Wang & Lu, 2018b).

Equation 4 & 5:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - x_i|}{N} \qquad RMSE = \sqrt{\frac{\sum_{n=1}^{N} (y_i - x_i)^2}{N}}$$

Next to these metrics, the R^2 score of the model will also be calculated, to analyse the predictive power of the model. The R^2 ranges from 0 to 1, where a score of 1 means that the features in the model can perfectly explain the variance of the cyclist counts. A higher R^2 score means that model can predict well based on its input.

The validation analysis is done for 500 model runs, where for each run different train and test data sets are used. The average values of these 500 runs are used as the results. This will be done for both the linear regression and Random Forest regression models, to determine which model fits the best.

3.6 Comparison

To determine the differences and similarities between this spatial planning based model and a four step model, a comparison of the model with the Fietsmonitor is made. The predictions of both models are compared to analyse how both operate. This comparison is made on the entire city of Apeldoorn and after that also on a specific subset of roads with different characteristics, to compare the predictions for specific locations throughout Apeldoorn.

For the comparison on Apeldoorn the differences between the predictions of the spatial planning model and the Fietsmonitor is analysed, to determine if they predict similarly or very differently. This is analysed by performing a paired t-test, to determine if the means of both predictions are statistically significant different or not. From this it can be concluded if the spatial planning approach predicts in an other way than the traditional four step model. For this a two-tailed t-test is conducted, as it just about the difference between the means and not about if one of the models specifically predicts greater intensities.

The equation that is used to calculate the t-statistic in Equation 6 below. Here $\bar{d}$ is the difference in the prediction between the model and the Fietsmonitor, $s_{\bar{d}}^2$ the variance of the differences and N the amount of segments that are compared (Hedberg & Ayers, 2015). For the t test a confidence interval of 95% is set up to determine if the differences are significant or not, with a null hypothesis that the mean difference is 0.

Equation 6:

$$t = \frac{\bar{d}}{\sqrt{\frac{s_{\bar{d}}^2}{N}}}$$

For the second comparison method four different categories of cycling roads in Apeldoorn are chosen to perform the comparison on. For each category two different roads are analysed, at different locations to ensure variability in the analysis. The first type is a regular segregated cycling path, where no motorized traffic can travel on. For this the Kanaal Zuid and Baron Sloetkade cycling paths are chosen. Next to this a cycling path or lane alongside a major motorized traffic road is chosen. Here the Koning Stadhouderlaan and Deventerstraat are chosen. Also access roads for residential areas will be analysed. These are not major roads, but very relevant for carrying traffic in, out and around residential areas. Here the Beethovenlaan and Sluisoordlaan will be used for the comparison. Lastly, roads within residential areas will be analysed. This comparison will be carried out on the Spreeuwenweg and the Schopenhauerstraat. Of all these locations pictures and more details on their location and present infrastructure can be found in Appendix A.

4. Results

In this section the results of the model and the analysis of those results can be read.

4.1 Model

It was found that the Random Forest regression performed significantly better in predicting the cyclist counts than the linear regression model, so the Random Forest regression model is used for the remainder of this research. The results of the evaluation of both models can be seen in Appendix A. This means that all results that are in presented in the remainder of this section are from the Random Forest Regression model.

4.1.1 Feature analysis

For the feature set with all features, also a feature analysis is conducted, but only on the network accessibility and connectivity features. This analysis is conducted through the SHAP values, to see which features have the greatest impact on the model. In Figure 10 this can be seen and from that it can be concluded that the five features with the greatest impact on the model are the NACHr2000m, NAINr2000m, NAIN on the network level, NAINr400m and the angular connectivity. From this it must be noted that these are the features with the greatest model impact of this subset of features, not per se the most important of all features. The meaning of all network features was explained in Section 3.3.4. These five features, together with all infrastructural, traffic, land use and other network features are used in the model to predict the cyclist intensities. From these features it is interesting to note that a great difference of radii is seen as important, as 400m, 2000m and the entire network are seen as important, not showing a clear indication for a radius or cycling time that has a greater impact.

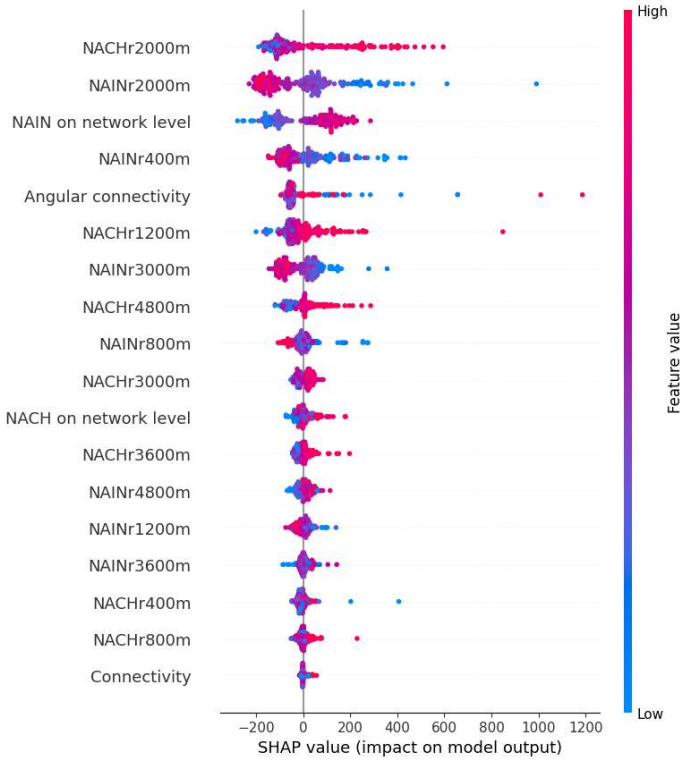

Figure 10: SHAP values of the network accessibility and connectivity features

4.1.2 Correlation analysis

The Pearson correlation analysis on all features in the model is presented in Figure 11 and also enlarged in Appendix B. From this it can be concluded that two combinations of features are too greatly correlated, as they exceed the thresholds that were set at 0.8 and -0.8. The pavement type asphalt and pavement type paving stones have a correlation score of -0.86, whilst the distance to a small station and the distance to the hospital have a correlation score of -0.81. These negative correlations could have been expected, as a segment cannot have asphalt and paving stones as its pavement type at the same time. Next to this, in Apeldoorn only few small stations and hospital locations exist, which are also located at relative opposite ends of the city. This makes that if the distance to the small stations is low, the distance to the hospital areas is large and vice versa, causing a negative correlation.

To determine which features have to be removed, the correlation with the count value is used. The features with the greatest correlation with the count value are kept, which are the distance to small station and pavement type paving stones features. This means that the distance to the hospital and pavement type asphalt are dropped as features, as the values of 0.27 for the small station and 0.08 for the paving stones show a greater effect than the values of 0.02 for the distance to hospital and -0.01 for the pavement type asphalt respectively.

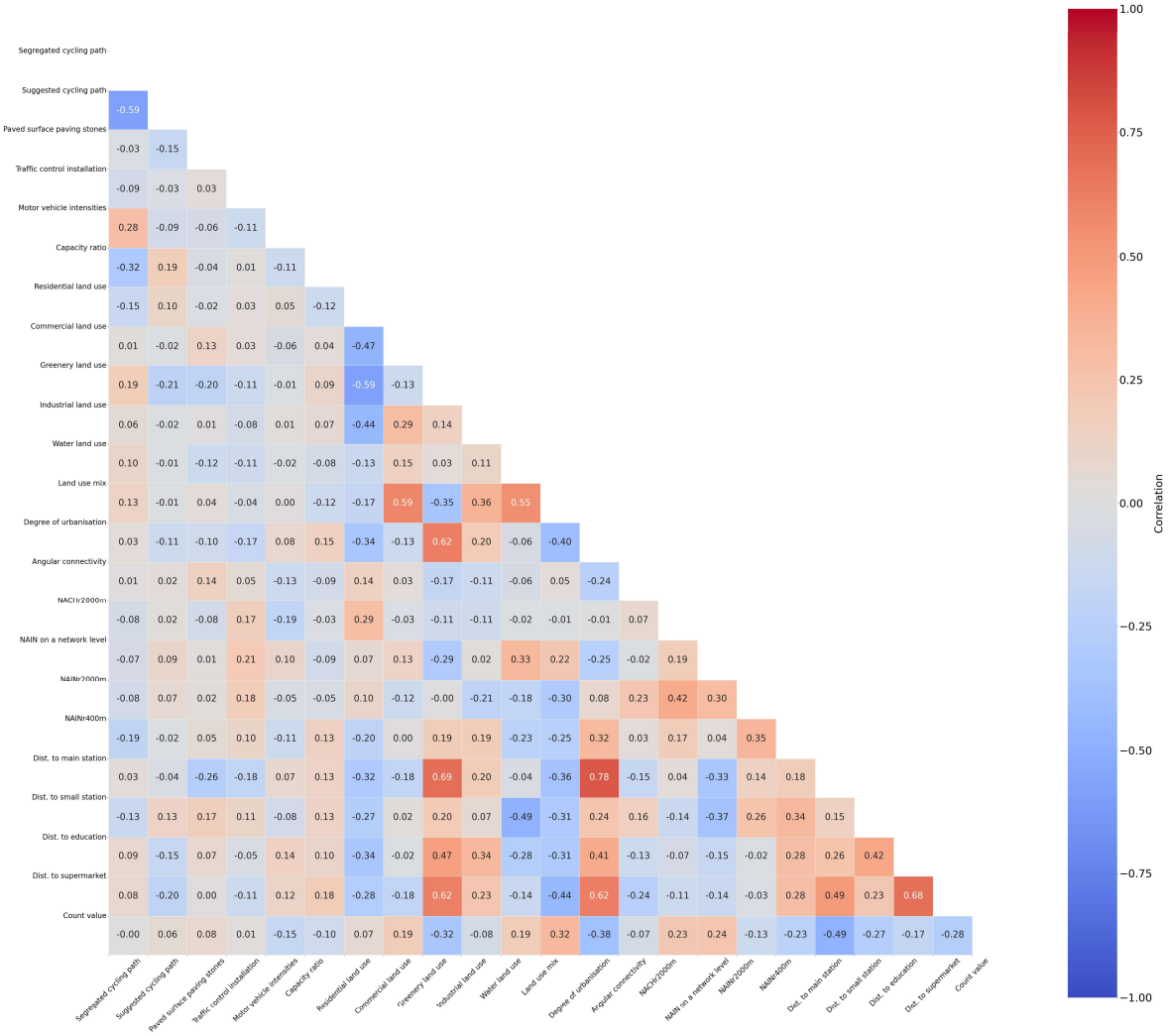

Figure 11: Pearson correlation matrix of the features and count value

4.1.3 Model parameters

In Figure 12 the results of the iteration of the amount of trees, plotted against the RMSE can be seen. From this it can be concluded that most optimal configuration for this is 90 trees in the Random Forest regression model. Other configurations such as 20 or 130 also perform well, but these configuration also have some outliers with very high RMSE values. Therefore the 90 trees configuration is preferred to be used in the model.

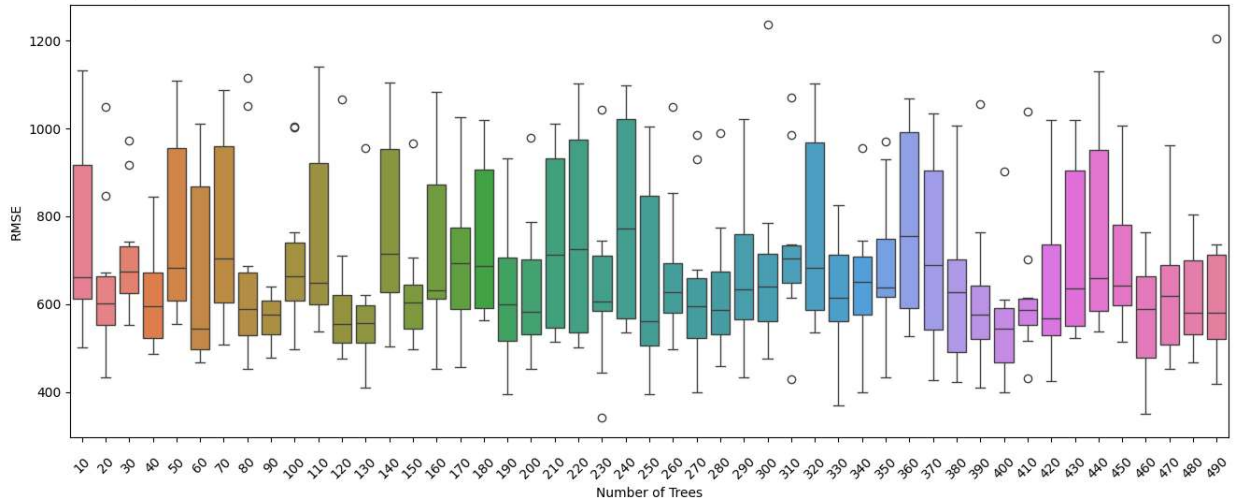

Figure 12: RMSE plotted against the number of trees

In Figure 13 the maximum depth of the trees plotted against the RMSE can be seen. From this it can be concluded that a maximum depth of 14 overall results in the most optimal results, as no significant outliers with a very high RMSE exist and overall the RMSE is also low. For other depths with an overall low RMSE such as a depth of 7, 25 or 34 outliers with a significantly larger RMSE exist, making them less optimal. A depth of 16 also results in a relatively low RMSE, but a depth of 14 still performs slightly better.

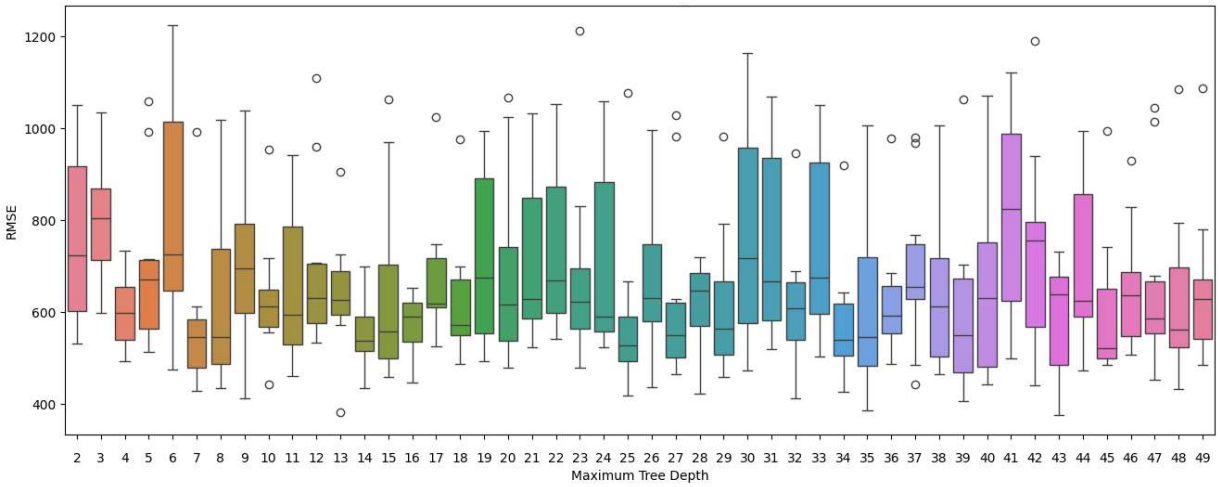

Figure 13: RMSE plotted against the maximum tree depth

4.1.4 SHAP feature importance analysis

The SHAP feature importance analysis is also conducted for all features in the model and the results are presented in Figure 14 below. The features are all ordered on their impact on the model output, with the most impactful features at the top. It is shown that the distance to the main station to the main station is the most important feature, with low distances resulting in high predictions. It can be seen that especially all network features are important, as of the five most impactful features only the land use mix is not a network characteristic. Next to this it is also interesting to note that all infrastructural features that are in the model are seen as having barely any impact on the results.

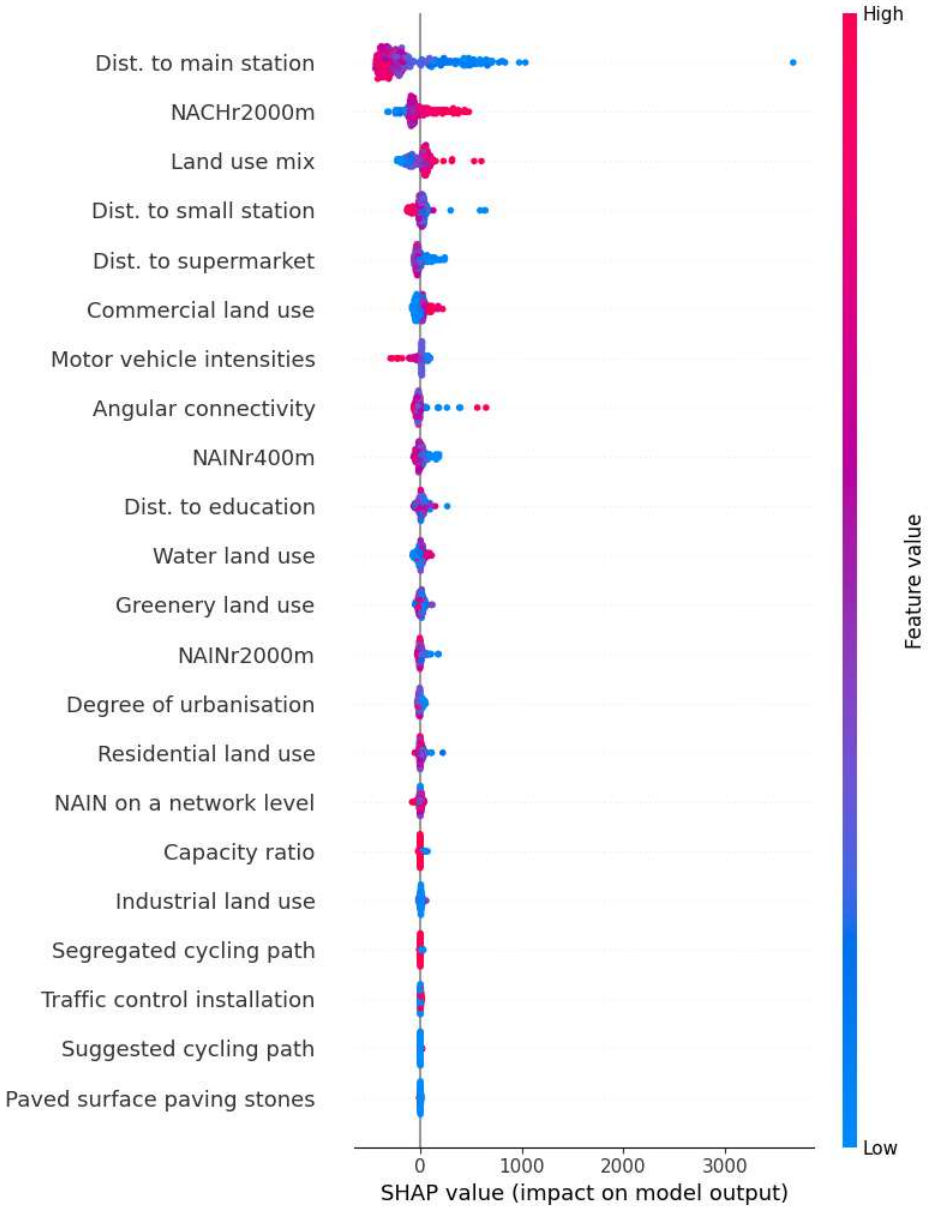

Figure 14: SHAP model importance values of all features

4.1.5 Predicted intensities

The predicted daily cyclist intensities for the entire city of Apeldoorn can be seen in Figure 15 below. The intensities are categorised in five categories, ranging from very low to very high. It can be seen that in general mostly low or very low intensities are predicted, which is logical. Nearly all high or very high intensity segments are cluttered around the main station area, which makes sense as proximity to the main station was the most impactful feature of the model.

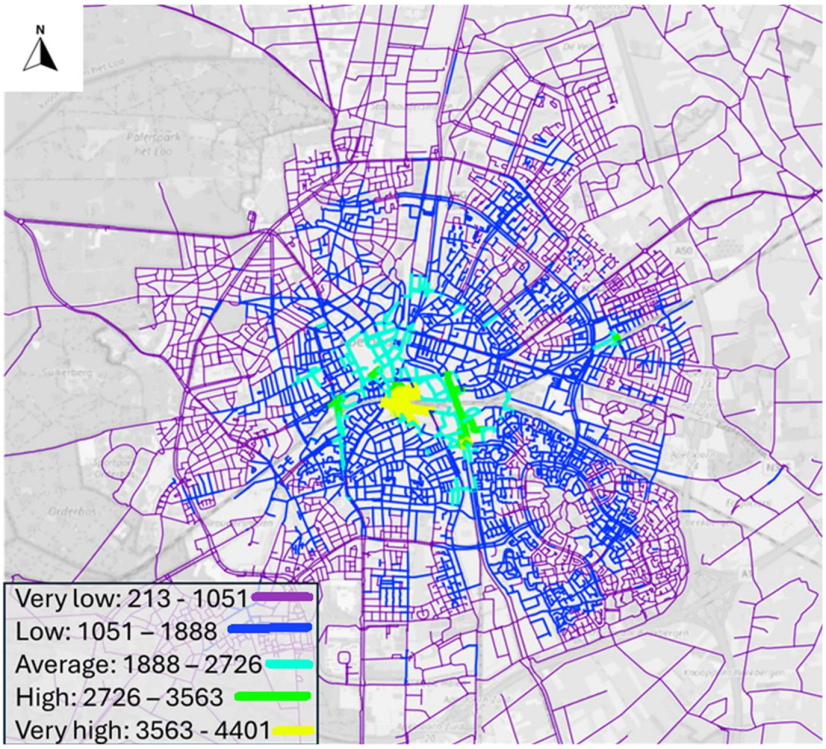

Figure 15: Predicted cyclist intensities for Apeldoorn (Background adapted from OpenStreetMap, n.d.)

4.2 Validation

The results of the model are compared to 20 percent of the actual counts, and the results of this for the different criteria can be seen in Table 6 below. These are the average values of 500 model runs, with each time different train and test sets. As the model predicts cyclist intensities on a daily basis, the MAE and RMSE also have an error of cyclist per day.

Table 6: Results of the validation

Criterion:	Value:
Mean Average Error (MAE)	436.40
Root Mean Squared Error (RMSE)	681.69
Train set R^2	0.92
Test set R^2	0.45

From these results it can be concluded that there are some larger errors in the predicted intensities, as the RMSE is higher than the MAE.

Next to this, the R^2 score of the training set is good, but becomes much worse for the test set. This shows that the model can only predict accurately on its training data and is potentially overfitted on this data.

As it is important to know the error of each prediction relative to the magnitude of the count, the residuals are calculated for the test set a randomly selected model run. Residual values denote the difference between counts and predictions, with positive residual values showing a too low prediction. These residuals can be seen in Figure 16. From this it can be concluded that the predictions that for up until 1800 cyclists the models functions good, but that for greater intensities the predictions of the model are mainly too low. This means that higher cyclist intensities are often predicted too low by the model. In Figure 17 this is also described, as here the predictions are plotted against the actual counts, showing that the counts are greater than the predictions, especially for greater values.

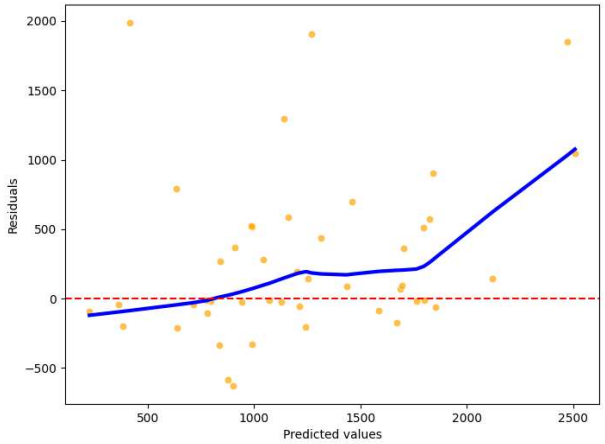

Figure 16: Residual values of the test set

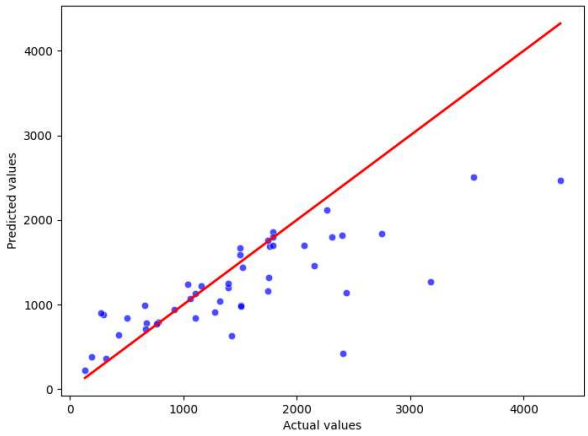

Figure 17: Predicted intensities plotted vs count values of the test set

4.3 Comparison

The results from the t-test to determine if the spatial planning model and the Fietsmonitor are different are presented in Table 7 below. With a confidence interval of 95% this means that the difference between is statistically significant and that it can be concluded that both models produce different results.

Table 7: Results of the t-test

T-statistic	P-value
122.40	< 0.01

Besides this, also multiple plots are made to get a clearer picture of how these differences are spread over the different predictions. In Figure 18 a scatterplot illustrating these differences is shown, illustrating a clear pattern of the model predicting higher intensities than the Fietsmonitor, especially for predictions below 1000 cyclists. Further analysis between the model, Fietsmonitor and the count values can be seen in Appendix D. From this it can be concluded that the model can predict the cyclist intensities more accurately in Apeldoorn than the Fietsmonitor can. This was to be expected, as the model is of course trained on this data whilst the Fietsmonitor is not.

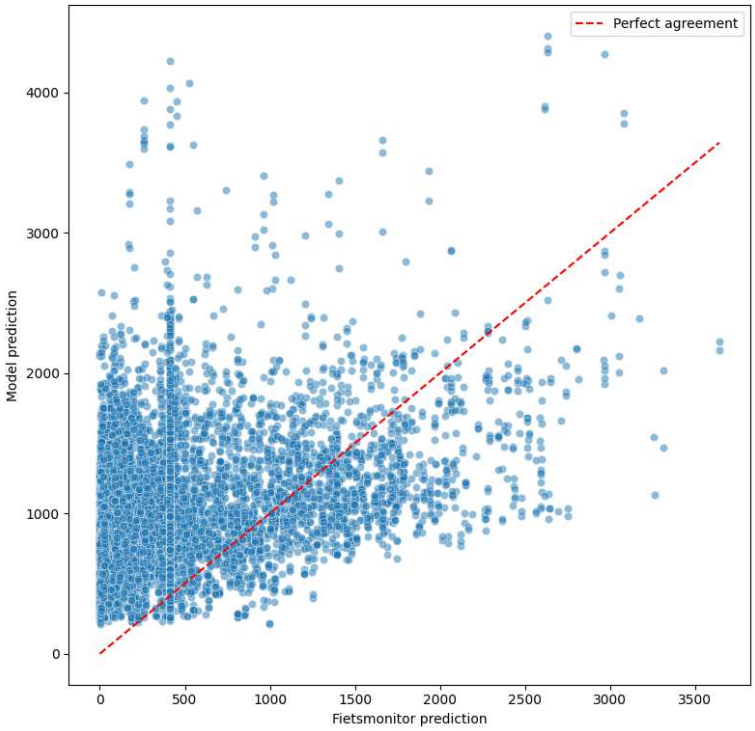

Figure 18: Scatterplot of the model and Fietsmonitor predictions

The results of the comparisons on different streets can be seen in Tables 8 up until 11 below. In these results it can be seen that the spatial planning model predicts the intensities higher for most types of streets. This is especially the case for the segregated cycling paths without motorized traffic and for streets in residential neighbourhoods. This is an interesting taking into account the Validation which is discussed in Section 5.2, as there it was found that the predictions are often lower than the counts.

The fact that the model predicts high intensities for residential neighbourhoods and their access roads is not unexpected, as the model was not trained on low intensities and therefore predicts high intensities mainly. The fact that the Fietsmonitor predicts lower intensities than the model, whilst the model already predicts too low intensities shows that the Fietsmonitor can not predict these count values accurately.

Table 8: Results of the comparison: Segregated cycling path with no motorized traffic

Location:	Predicted intensity spatial planning approach:	Predicted intensity Fietsmonitor:
Kanaal Zuid	1060	268
Baron Sloetkade cycling path	1510	442

Table 9: Results of the comparison: Segregated cycling path or lane with motorized traffic

Location:	Predicted intensity spatial planning approach:	Predicted intensity Fietsmonitor:
Koning Stadhouderlaan	1834 west side, 1202 east side	2226 west side, 2318 east side
Deventerstraat	1120 north side, 1083 south side	1155 north side, 1303 south side

Table 10: Results of the comparison: Access roads to residential neighbourhoods

Location:	Predicted intensity spatial planning approach:	Predicted intensity Fietsmonitor:
Beethovenlaan	632	291
Kruizemuntstraat	1126	501

Table 11: Results of the comparison: Streets in residential neighbourhoods

Location:	Predicted intensity spatial planning approach:	Predicted intensity Fietsmonitor:
Spreeuwenweg	909	119
Schopenhauerstraat	937	248

5. Discussion

During this research several assumptions were made and shortcomings were found, which all influence the accuracy of the final results. These findings will be discussed in this section.

First of all, the main subject of this research are the cyclist counts that were used as input for the model. It was already noted that not many high counts (3000+) were present and that the counts that did have such great values were therefore some sort of outlier. In the model however also very few low counts were used as input, as the lowest count value was 101 cyclists, with just 13 of the 189 count values being below 250 cyclists. As the model can only predict based on its input, it is not well suited to predict low intensity segments, as these are not often present in the train set. Because of this the model overall probably predicts relatively high intensities, also for neighbourhoods and segments were this would not be expected.

Next to this, the methods of operationalisation for all features was not fully accurate, which could also influence the results. The impact of signalised intersections was operationalised by taking a buffer around the intersection, but in dense urban areas such as Apeldoorn this method would also include segments that were nearby the intersection, but not actually leading towards the intersection and therefore not influenced by it. The proximity of segments towards facilities was implemented by taking a distance as the crow flies, instead of via a network route. This could cause for segments to appear much closer to facilities than how they can actually be reached, especially if there are infrastructural or natural barriers present such as train tracks or water ways. As Apeldoorn has both of these, this could have an impact on the results.

Also, as a segment approach was used, the network was also implemented in the Space Syntax network analysis in that way. In the network analysis however, each segment passed is seen as a turn, whilst in reality this would not be the case for cyclists on the segments, as a small street on the side would not affect them. This misalignment between reality and the network analysis could create inaccuracies in the results of the network analysis.

In the feature importance analysis, it was noted that infrastructural characteristics were not significant predictors for the model's outcomes. This finding may be caused by the nature of the data used for cyclist counts, which were mainly collected at locations where a high volume of cyclists was already present. Consequently, these locations often featured cycling infrastructure. When the majority of counts are obtained from areas that already have suitable cycling infrastructure, the model could encounter challenges in recognizing infrastructural characteristics as important features. Because of this it can not strongly be concluded that infrastructural characteristics are not important in a spatial planning context.

6. Conclusion

The aim of this research was to predict the volume of cyclists at road segments based on the environmental characteristics of the segments, via a spatial planning approach on a case study on the city of Apeldoorn and to compare these results with a traditional traffic engineering four-step model.

In order to realise this, first existing literature was studied to determine which characteristics would be relevant to include in the model, both from a traffic engineering and spatial planning perspective. This resulted in 38 relevant features that were eventually operationalised to be used as input for the regression model. Next to this regression model types were evaluated, where both linear and Random Forest regression models were seen as potentially suitable.

With the relevant characteristics a model was made, where both regression types were evaluated. From this it was concluded that Random Forest regression is the most suitable for this research and used to predict cyclist intensities. The results from the Random Forest regression model were optimised through a feature selection of the network characteristics, a correlation analysis and an optimization of the parameters of the Random Forest model. The results of the model were also compared with a four step model in the form of the Fietsmonitor, to analyse the differences and similarities between them.

The most important features of the model where the network characteristics, with especially the proximity to trains stations being important. From the results it was further concluded that the model can predict cyclist intensities, although with an error. The MAE and RMSE became more significant for greater predictions and it was also found that the model was potentially overfitted on its train data set. Next to this the model could predict cyclist intensities more accurately than the Fietsmonitor, but as the model was trained on this data no definitve conclusions can be made.

7. Recommendations

Based on the limitations and inaccuracies in this research, several recommendations for future research are set up and these can be read in this chapter.

In this research only the Random Forest regression model was used to handle the complex dataset, but other high end machine learning models were not evaluated. For example, XGradient boosting or other models could potentially predict more accurate results and this could be analysed in the future. The XGradient boosting is a more complex regression model, which could be better suited to prevent overfitting on the train data and be more capable of finding complex relationships between features.

For future research on spatial planning approaches to predict cyclists intensities, more diversity in the cyclist counts is advised so that locations with also very low and very high cyclist intensities can be predicted accurately.

As it was concluded that the network analysis via Space Syntax resulted in the second most impactful feature to predicted the cyclist intensities, it is recommended to incorporate these features more often in research related to cyclists as it worked well in this research. The features about proximity to facilities were also seen as very impactful, as the distance to train stations and supermarkets scored as very impactful, so these network features could potentially also give insightful results in future studies. These findings can also be very useful for municipalities and other policy makers, as this shows that the location of a segment related to facilities has a significant impact on the amount of cyclists passing through.

It was concluded that the spatial planning model scores better than the Fietsmonitor in this study, but as the model was trained on this count data and the Fietsmonitor this is not a definitive conclusion. To fully and accurately determine this, both models should be evaluated on a dataset that both models were not trained on.

Lastly, Witteveen+Bos could use this spatial planning model in the future to predict cyclist intensities alongside the already existing Fietsmonitor. For this it must of course be noted that the model has an error, especially for the very low and very low intensity locations. It could also be very interesting for Witteveen+Bos to do a second comparison between the spatial planning model and the Fietsmonitor on another city, to properly determine how both models function. Next to this a combination of the spatial planning model into the Fietsmonitor could also be possible, where for example the network characteristics are especially incorporated.

8. References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Keywords: Correlation coefficient; Interpretation; Pearson's; Spearman's; Lin's; Cramer's
Bobbitt, Z. (2022, April 20). *Shannon Diversity Index: Definition & example*. Statology. https://www.statology.org/shannon-diversity-index/

Centraal Bureau voor de Statistiek, (CBS). (n.d.). *Wijk- en buurtkaart 2025*. Centraal Bureau Voor De Statistiek. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2025

Chen, P., Shen, Q., & Childress, S. (2018) A GPS data-based analysis of built environment influences on bicyclist route preferences, International Journal of Sustainable Transportation, 12:3, 218-231, DOI: 10.1080/15568318.2017.1349222

CROW. (2022). Geactualiseerde aanbevelingen voor de breedte van fietspaden 2022. *https://fietsberaad.nl/getmedia/e57e2986-5719-413f-bc6a-7d37bb7a36c5/Geactualiseerde-aanbevelingen-voor-de-breedte-van-fietspaden-2022_versie2.pdf.aspx?ext=.pdf#page=6.30*

de Jong, T., Böcker, L., & Weber, C. (2022). Road infrastructures, spatial surroundings, and the demand and route choices for cycling: Evidence from a GPS-based mode detection study from Oslo, Norway. Environment and Planning B, 50(8), 2133-2150. https://doi.org/10.1177/23998083221141431

depthmapX development team. (2020). depthmapX (Version 0.8.0) [Computer software]. Retrieved from https://github.com/SpaceGroupUCL/depthmapX/

De Wit, A., Versluis, L. & Leferink, T. (2021). Van eerste stappen tot sprint: de ontwikkeling van een nieuw soort voetgangersmodel, de Rotterdamse LoopMonitor

Dhanani, A., Tarkhanyan, L., & Vaughan, L. (2017). Estimating pedestrian demand for active transport evaluation and planning. *Transportation Research Part a Policy and Practice*, *103*, 54–69. https://doi.org/10.1016/j.tra.2017.05.020

Drolsbach, S. K. (2022). *The search for cycling routes: Analysing the influence of spatial characteristics on cycling routes in Amsterdam* [MSc Thesis]. AMS Institute, TU Delft & Wageningen University.

Fernando, J. (2024, July 29). *The correlation Coefficient: what it is and what it tells investors*. Investopedia. https://www.investopedia.com/terms/c/correlationcoefficient.asp#toc-what-is-the-correlation-coefficient

Fietsberaad. (2023). *Fietsstromen modelleren in bestaande simulatiemodellen.* Retrieved May 24, 2025, from https://www.fietsberaad.nl/Kennisbank/Fietsstromen-modelleren-in-bestaande-simulatiemode

Fietsersbond. (2021). *Cycling network of Apeldoorn for 2021.* https://www.fietsersbond.nl

Fietsersbond. (2023, August 28). *Fietsroutes naar middelbare scholen vaak nog te onveilig voor scholieren*. Fietsersbond. https://www.fietsersbond.nl/nieuws/fietsroutes-naar-middelbare-scholen-vaak-nog-te-onveilig-voor-scholieren/#:~:text=Onder%20deze%20cijfers%20vallen%20ook,school%20(DUO%2C%202022).

GeeksforGeeks. (2025a, May 20). *Multiple Linear Regression using Python ML*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/ml-multiple-linear-regression-using-python/

GeeksforGeeks. (2025b, May 30). *Random Forest algorithm in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

Gemeente Apeldoorn (2023). Motorized Vehicle Intensities Model

Google Maps. (n.d.). Google Maps. https://www.google.com/maps

Hedberg, E.C., & Stephanie Ayers, S. (2015) *The power of a paired t-test with a covariate.* Social Science Research, Volume 50, 2015, Pages 277-291, ISSN 0049-089X, https://doi.org/10.1016/j.ssresearch.2014.12.004.

Hillier, B. (2007) Space Syntax. *Space is the machine*. Space Syntax. https://www.spacesyntax.com

Interpretable Machine Learning. (n.d.). Chapter *18 SHAP* https://christophm.github.io/interpretable-ml-book/shap.html

Jestico, B., Nelson, T. & Winters, M. (2016) *Mapping ridership using crowdsourced cycling data*, Journal of Transport Geography, Volume 52, 90-97, ISSN 0966-6923, https://doi.org/10.1016/j.jtrangeo.2016.03.006.

Jiang, B. (2009). Ranking spaces for predicting human movement in an urban environment. *International Journal of Geographical Information Science*, *23*(7), 823–837. https://doi-org.ezproxy2.utwente.nl/10.1080/13658810802022822

Koch, T., & Dugundji, E. R. (2021). Taste variation in environmental features of bicycle routes. Proceedings of the 14th ACM SIGSPATIAL International Workshop on Computational Transportation Science, IWCTS 2021, article 2, (pp. 1–10). https://doi.org/10.1145/3486629.3490697

Khatri, R., Cherry, C. R., Nambisan, S. S., & Han, L. D. (2016). Modeling Route Choice of Utilitarian Bikeshare Users with GPS Data. Transportation Research Record, 2587(1), 141-149. https://doi.org/10.3141/2587-17

Knijnenburg, D. (2021). *Attractiveness of roads influencing bicycle traffic*. [BSc Thesis]. University of Twente

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Lerman, Y., Rofè, Y., & Omer, I. (2014). Using space syntax to model pedestrian movement in urban transportation planning. *Geographical Analysis*, *46*(4), 392–410. https://doi.org/10.1111/gean.12063

Łukawska, M., Paulsen, M., Rasmussen, T. K., Jensen, A. F., & Nielsen, O. A. (2023). A joint bicycle route choice model for various cycling frequencies and trip distances based on a large crowdsourced GPS dataset. Transportation Research Part A: Policy and Practice, 176, 103834. https://doi.org/10.1016/J.TRA.2023.103834

McCahill, C., & Garrick, N. W. (2008). The applicability of space syntax to bicycle facility planning. *Transportation Research Record Journal of the Transportation Research Board*, *2074*(1), 46–51. https://doi.org/10.3141/2074-06

McNally, M.G. (2007), "The Four-Step Model", Hensher, D.A. and Button, K.J. (Ed.) *Handbook of Transport Modelling* (*, Vol. 1*), Emerald Group Publishing Limited, Leeds, pp. 35-53.

Meister, A., Felder, M., Schmid, B., & Axhausen, K. W. (2023). Route choice modeling for cyclists on urban networks. Transportation Research Part A: Policy and Practice, 173, 103723. https://doi.org/10.1016/J.TRA.2023.103723

Ministerie van Algemene Zaken. (2024, August 22). *Kabinet: meer mensen op de fiets*. Fiets | Rijksoverheid.nl. https://www.rijksoverheid.nl/onderwerpen/fiets/fietsbeleid

Mwiti, D. (2023, September 1). *Random forest regression: When does it fail and why?* neptune.ai. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why

NDW *(2025). Cyclist count data. Retrieved from:* https://docs.ndw.nu/producten/fietsdata/

NS Dashboard. (2023). https://dashboards.nsjaarverslag.nl/reizigersgedrag/apeldoorn

OpenStreetMap. (n.d.). *OpenStreetMap.* https://www.openstreetmap.org/#map=12/52.2196/5.9790

Orellana, D. & Guerrero Balarezo, M. L. (2019a). *The influence of Space Syntax on Cycling Movement. Proceedings of the 12 Space Syntax Symposium.*

Orellana, D., & Guerrero Balarezo, M. L. (2019b). *Exploring the influence of road network structure on the spatial behaviour of cyclists using crowdsourced data.* Environment and Planning B, 46(7), 1314-1330. https://doi.org/10.1177/2399808319863810 (Original work published 2019)

Ourique, L., Eloy, S., Resende, J.R.P., & Dias, Miguel, S.D. (2017). Spatial perception of landmarks assessed by objective tracking of people and Space Syntax techniques.

PDOK. (2025). *Dataset: Basisregistratie Adres Gegevens. Retrieved from:* https://service.pdok.nl/lv/bag/atom/bag.xml

PDOK. (n.d.). *Dataset: Basisregistratie Grootschalige Topografie. Retrieved f*-topografie-bgt-*rom:* https://www.pdok.nl/introductie/-/article/basisregistratie-grootschalige

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, 12(10), 770 781. https://doi.org/10.1080/15568318.2018.1437236

Raford, N., & Ragland, D. (2004). Space Syntax: innovative pedestrian volume modeling tool for pedestrian safety. *Transportation Research Record Journal of the Transportation Research Board*, *1878*(1), 66–74. https://doi.org/10.3141/1878-09

Raford, N., & Ragland, D. (2006). Pedestrian Volume Modeling for Traffic Safety and Exposure Analysis: The Case of Boston, Massachusetts. *UC Berkeley: Safe Transportation Research & Education Center*. Retrieved from https://escholarship.org/uc/item/61n3s4zr

Rivera, Viridiana (n.d.). *Cyclists commuting through Amsterdam streets* [Photograph]. Pexels. https://www.pexels.com/photo/cyclists-commuting-through-amsterdam-streets-28683733/

Uijtdewilligen, T., Baran Ulak, M., Wijlhuizen, G. J., & Geurs, K. T. (2024). Effects of crowding on route preferences and perceived safety of urban cyclists in the Netherlands. Transport Research Part A: Policy and Practice, 183, 104030.

Van Nes, A., & Yamu, C. (2021). Introduction to space Syntax in urban Studies. In *Springer eBooks*. https://doi.org/10.1007/978-3-030-59140-3

van Nijen, N., Ulak, M. B., Veenstra , S., & Geurs , K. (2024). Exploring factors affecting route choice of cyclists: A novel varying-contiguity spatially lagged exogenous modeling approach. *Journal of Transport and Land Use*, *17*(1), 557–577. https://doi.org/10.5198/jtlu.2024.2452
Veenstra, S. A. (2008). *Verkeerspatronen rond supermarkten: onderzoek naar patronen met betrekking tot supermarktverkeer [MSc Thesis]. University of Twente.*

Wang, W., & Lu, Y. (2018). Analysis of the mean Absolute Error (MAE) and the Root Mean square Error (RMSE) in assessing rounding model. *IOP Conference Series Materials Science and Engineering*, *324*, 012049. https://doi.org/10.1088/1757-899x/324/1/012049

Witteveen+Bos. (2023). *Fietsmonitor to estimate cyclist intensities*

Yamaguchi, K.(2020) "Intrinsic Meaning of Shapley Values in Regression," 2020 11th International Conference on Awareness Science and Technology (iCAST), Qingdao, China, 2020, pp. 1-6, doi: 10.1109/iCAST51195.2020.9319492.

Yang, C., & Mesbah, M. (2013). *Route choice behaviour of cyclists by stated preference and revealed preference*. Australasian Transport Research Forum 2013 Proceedings, Australia. https://australasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2013_yang_mesbah.pdf

Yang X., Chu Y., Hu S., Jin L, Liu H. & Tao N., *Evaluating the influence of environmental factors and route characteristics on leisure-oriented active travel: A case study in Skåne Province*, Landscape and Urban Planning, Volume 259, 2025, 105343, ISSN 0169-2046, https://doi.org/10.1016/j.landurbplan.2025.105343.

9. Appendix

9.1 Appendix A: Locations used in the comparison

Comparison of the segregated cycling path - no motorized traffic

Kanaal Zuid: A 2-way segregated cycling path. This can be seen in Figure 19.

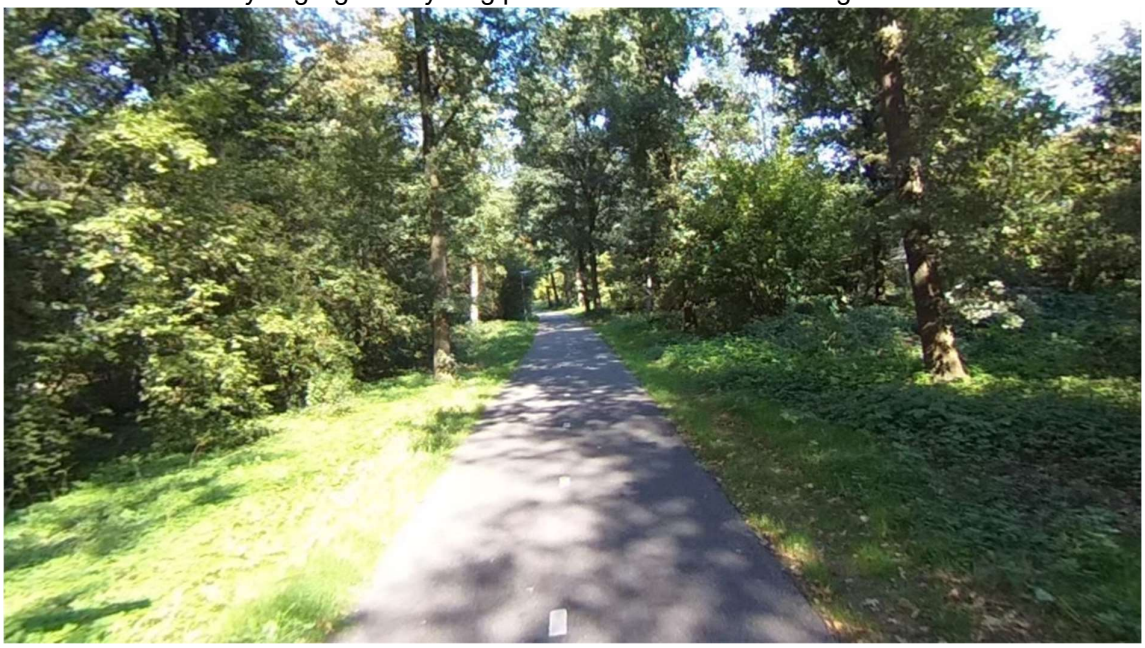

Figure 19 Kanaal Zuid (Google Maps, n.d.)

Baron Sloetkade cycling path: A 2 way segregated cycling path. Can be seen in Figure 20.

Figure 20: Baron Sloetkade cycling path (Google Maps, n.d.)

Comparison of the segregated cycling path - with motorized traffic

Koning Stadhouderlaan: On both sides a 2 way segregated cycling path. Can be seen in Figure 21.

Figure 21: Koning Stadhouderlaan (Google Maps, n.d.)

Deventerstraat: On both sides a 1 way segregated cycling path. Can be seen in Figure 22.

Figure 22: Deventerstraat (Google Maps, n.d.)

Comparison of residential access roads

Beethovenlaan: A shared road with motorized traffic. Can be seen in Figure 23.

Figure 123: Beethovenlaan (Google Maps, n.d.)

Kruizemuntstraat: A shared road with motorized traffic. Can be seen in Figure 23.

Figure 24: Kruizemuntstraat (Google Maps, n.d.)

Comparison of residential streets

Spreeuwenweg: A shared road with motorized traffic. Can be seen in Figure 25.

Figure 25: Spreeuwenweg (Google Maps, n.d.)

Schopenhauerstraat: A shared road with motorized traffic. Can be seen in Figure 26.

Figure 26: Schopenhauerstraat (Google Maps, n.d.)

Location in the network:

All these streets are also located at different locations throughout Apeldoorn, this can be seen in Figure 27.

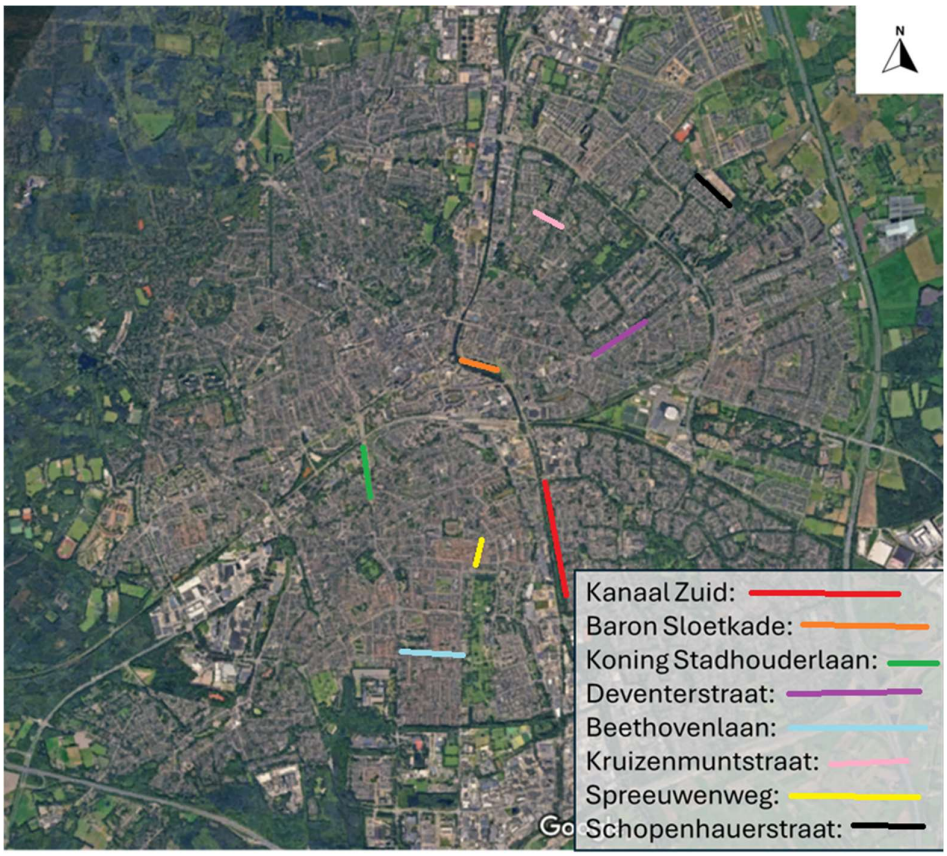

Figure 27: Locations of the compared streets (Google Maps, n.d.)

9.2 Appendix B - Evaluation of the regression models

Table 13: Evaluation of the Random Forest and linear regression models

Criterion:	Values for Random Forest regression:	Values for linear regression:
Mean Average Error (MAE)	436.40	572.87
Root Mean Squared Error (RMSE)	681.69	785.02
Train set R^2	0.92	0.46
Test set R^2	0.45	0.24

9.3 Appendix C - Pearson correlation matrix enlarged

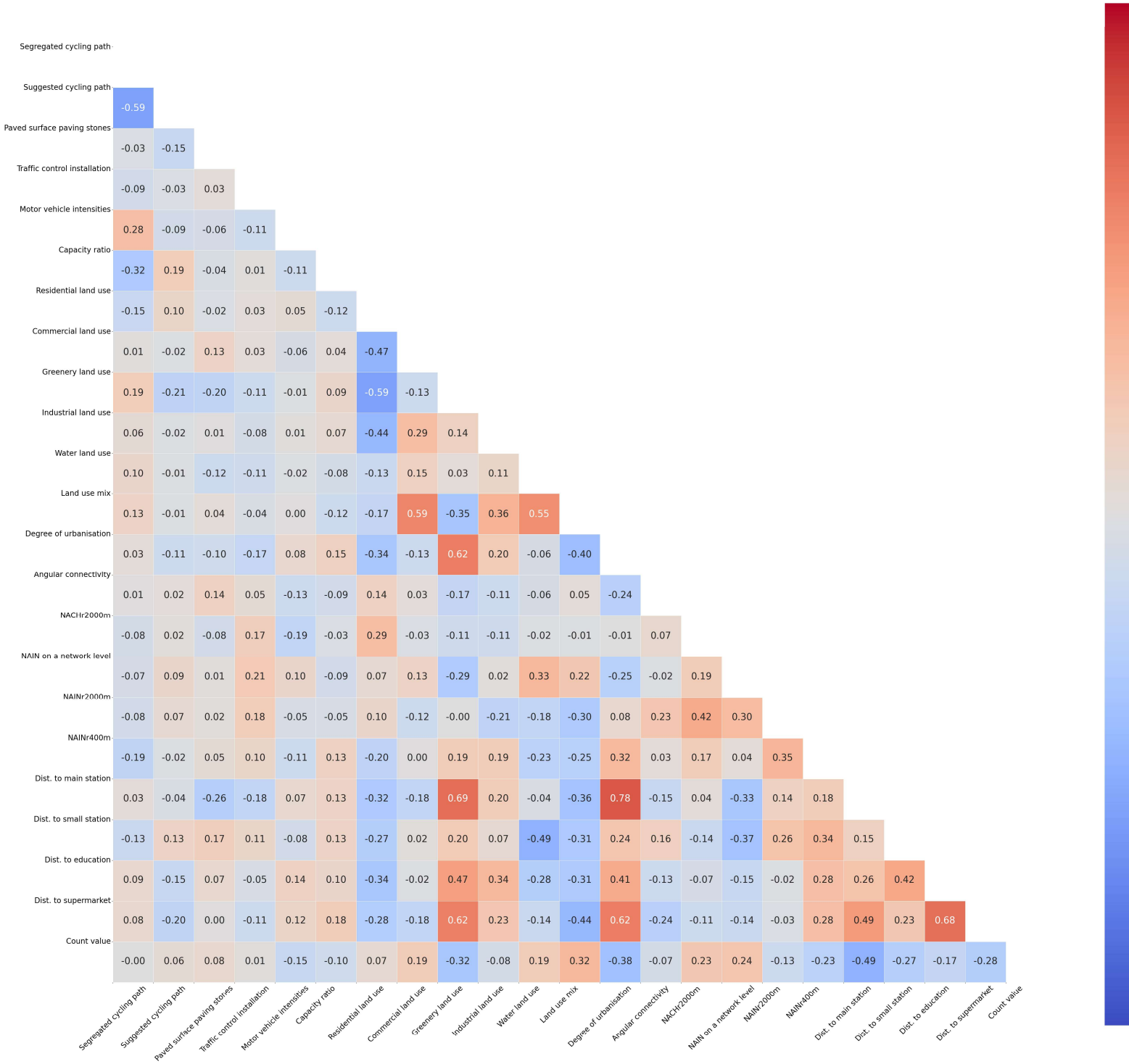

Figure 11: Pearson correlation matrix of the features and count value

9.4 Appendix D: Additional results of the comparison with the Fietsmonitor

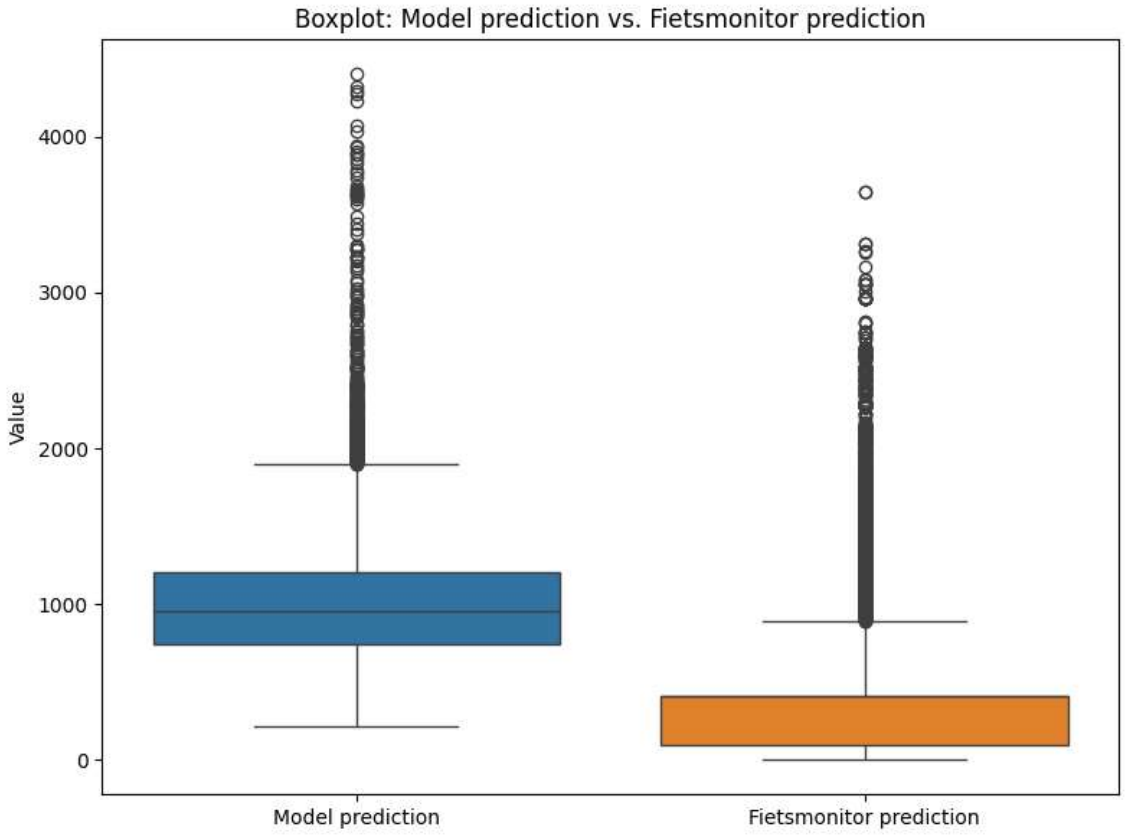

Figure 28: Boxplots showing the differences in predictions between the model and Fietsmonitor

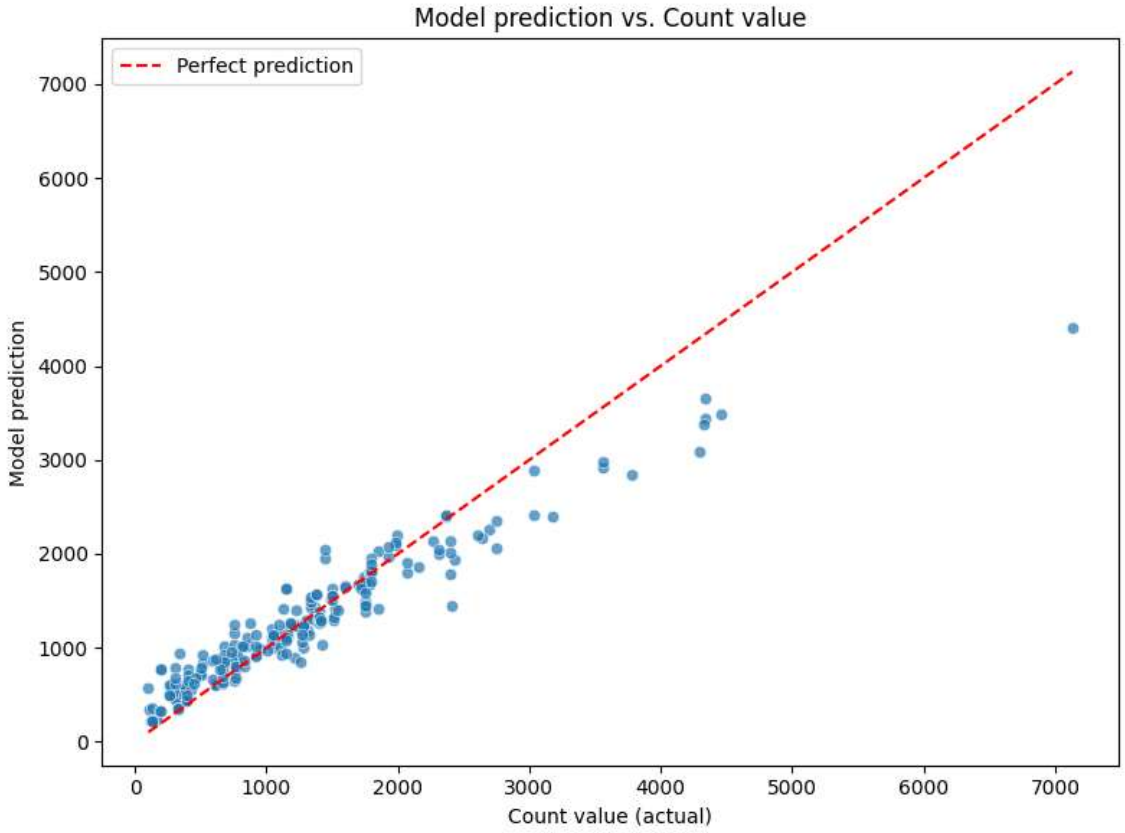

Figure 29: The model predictions plotted against the count values

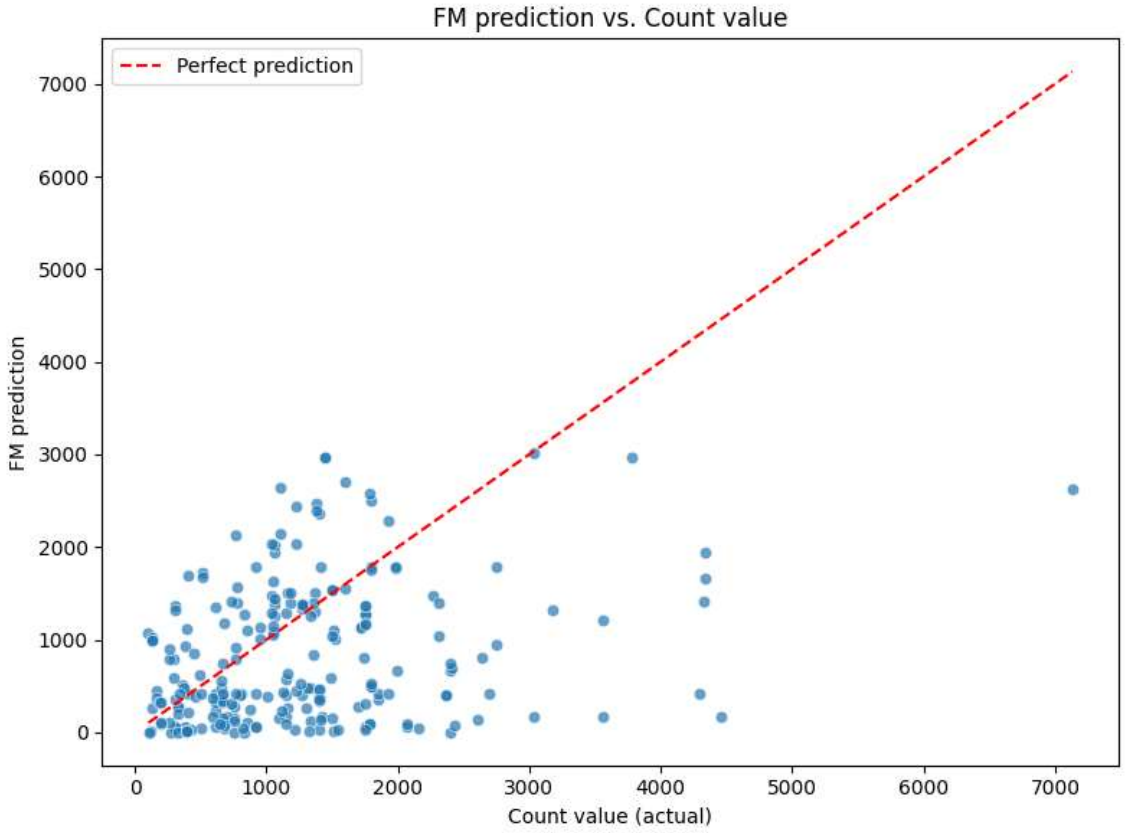

Figure 30: The Fietsmonitor predictions plotted against the count values