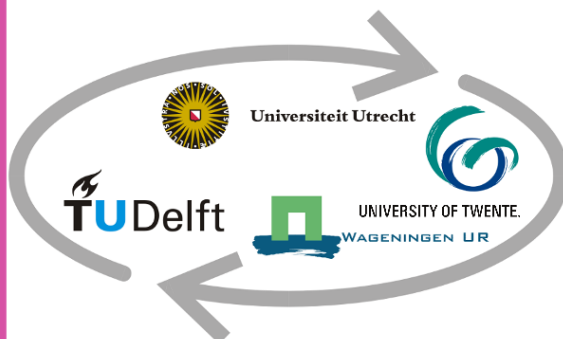


Investigating the influence of safety on cyclists route choice: a revealed preference study

C.J.M. de Vos



5th June 2018

Collaboration between

Utrecht University (Faculty of Geosciences) and ARCADIS



Universiteit Utrecht



Investigating the influence of safety on cyclists route choice: a revealed preference study

C.J.M. de Vos

Supervisor

Dr. Simon Scheider

Faculty of Geosciences
Utrecht University

Professor

Prof. dr. Stan Geertman

Faculty of Geosciences
Utrecht University

5th June 2018

Acknowledgements

For as long as I can remember I have been fascinated by geography and the movement of people, cycling in particular. During this master thesis, I had the opportunity to combine and explore my interests even further at Arcadis. Now, almost nine months later, the hard work of this thesis is completed.

As part of the master Geographical Information Management and Applications, I present to you my final thesis titled: *Investigating the influence of safety on cyclists route choice: a revealed preference study*.

Completing this research would not have been possible without the help from many people to whom I would like to express my gratitude.

First of all, I would like to thank my university supervisor Simon Scheider for always giving me valuable tips and feedback throughout the process. Secondly, special thanks go out to my professional supervisor at Arcadis, Marc Schenk, for supporting me throughout this rocky process, sharing my curiosity of the topic and providing me with valuable background information about cycling safety. However, this whole process would not have been possible without Joost de Kruijff who introduced me to this topic in the first place.

Additionally, I would like to thank all data providers that made this research possible. First of all, I would like to thank the Fietzersbond for providing the crucial cycling network data enriched with many different safety attributes. Secondly to the partner of the Fietstelweek, the NHTV, who were able to map match the data to the cycling network of my choice. Thirdly, many thanks go out to several people at the Municipality of Tilburg; Bram van Berkel for providing me with the data (and explanations) from their traffic model and Marit Beijers for sharing data from a local research project regarding social safety in Tilburg. And finally, I would like to thank Sanne van Zundert from Keypoint Consultancy, as one of the official partners of the Fietstelweek, who helped me investigate the representativity of the GPS-data for my specific study area.

Last but not least, I would like to thank my partner, fellow students and friends for supporting and motivating me, and for providing me with valuable advice, help and feedback during this process.

Summary

Compared to motorized vehicles, the behavior of cyclists is still, to this day, relatively unknown. Main traffic research assumes that people choose the shortest route to minimize their travel costs, but this does not seem to be the case for cyclists. For them, safety can be a factor that is as important (or even more important) than travel time or distance. Therefore safety can be the reason why a person travels a certain route. However, people tend to act differently than expected and that might especially be the case in a Dutch context, where cycling is very rooted in the culture and in general already very safe.

In this thesis, GPS data from the largest cycling project in the Netherlands (the Fietstelweek) is used to investigate on a large scale whether safety factors of the environment in fact explain route choice behavior. Route choice behavior is investigated by comparing the chosen routes with the shortest possible route alternatives in terms of travel distance (i.e. amount of deviation) and safety factors (Δ safety factor).

Extensive literature research has shown that the following objective and subjective safety factors are relevant for explaining route choice behavior: road type, road surface, road surface quality, intersections and control (traffic lights), obstacles, traffic intensity, cycling accidents and social safety. To measure these safety factors a wide array of data sources have been used. Furthermore, several GIS methods have been used to enrich the GPS tracks with the safety data.

In general, people did in fact choose significantly longer routes than the shortest possible alternatives. However, the amount of deviation is limited. The relationship between the amount of deviation and the difference in safety factors on both routes is captured in a simple linear model using a single linear regression method for all individual safety factors. It turns out that several safety factors do, in fact, influence cyclists' route choice behavior in the way that we would expect based on preferences found in literature. The difference between the presence of illumination, road surfaces and its quality, some road types and intersections on the chosen route compared to the shortest route alternative explains the size of the detour. However, also a number of safety factors show to influence route choice in a way that does not match preferences that were found in previous research. Many effects are difficult to explain because the measured safety factor might be closely linked to other characteristics

(more important than safety) of the route that are not taken into account in the simple linear models.

Despite the fact that many relationships exist that significantly explain the amount of deviation from the shortest route, the effect sizes and model fits are in almost all models are (extremely) small. In general, the simple linear models do not perform that well, and it proves that these relationships are not well captured in simple linear models with only one explanatory variable. Even though not all relationships are significant throughout the day, interestingly enough, the models performed better while investigating the relationships limited to a specific part of the day compared to a general model without this distinction.

Overall, this project should be considered as a first attempt at data-driven research to investigate route choice behavior and its relation to safety using GIS methods. Even though using the open- and freely available cycling data from the Fietstelweek has had many benefits, also some limitations have been encountered. Many of them related to privacy. Even though some insight is acquired regarding the effect every (Δ) safety factor individually has on the amount of deviation from the shortest route, these models are not suitable for other purposes such as predicting route choice behavior.

Understanding the underlying reasons for route choice is the key to improving the cycling network and infrastructure so that it fits the needs of the users. To gain more and better insight into the complex relationship between route choice and safety, other new (non-linear) methods and techniques should be explored and multiple (safety) factors should be taken into account. Further research should also include travel time and cycling motives, social-economic characteristics and changing infrastructure.

Contents

1	Introduction	2
1.1	Context	2
1.2	Problem statement and relevance	3
1.3	Research questions	4
1.4	Research scope	5
1.5	Reader's guide	5
2	Theoretical background	6
2.1	Route choice behavior	6
2.2	Accidents	7
2.3	Road type	9
2.4	Road surface	13
2.5	Obstacles	14
2.6	Traffic intensity	15
2.7	Social safety	16
2.8	Illumination	17
2.9	Conclusion	18
3	Methodology	20
3.1	Measuring route choice behavior	20
3.1.1	Stated preference studies	20
3.1.2	Revealed preference studies	20
3.1.3	Chosen method and study area	21
3.2	Research methods	22
3.2.1	Prepare & enrich network	23
3.2.2	Create shortest & chosen routes	23
3.2.3	Track enrichment	24
3.2.4	Statistical analysis	29
4	Dataset descriptions	32
4.1	Cycling GPS data	32
4.1.1	Fietstelweek project	32
4.1.2	Fietstelweek (2016) data	34
4.2	Cycling network	36

4.3	Safety data sets	36
5	Data preparation and data quality	40
5.1	Data preparation network	40
5.1.1	Topology of the cycling network	40
5.1.2	Network enrichment	41
5.2	Data preparation routes	47
5.2.1	Creating chosen routes	47
5.2.2	Creating shortest routes	47
5.3	Separation routes in timeslots	48
6	Data analysis	50
6.1	General route characteristics	50
6.1.1	Differences in trip length	50
6.1.2	Difference in safety factors	52
6.2	Regression analysis	54
6.2.1	General outcomes	54
6.2.2	Outcomes per safety factor	56
7	Conclusion	62
7.1	Answering research questions	62
7.2	Discussion & limitations	67
7.3	Recommendations for future research	69
	Bibliography	71
	Appendices	79
A	Machine learning	80
B	Cycling network	81
C	Representativity Fietstelweek 2016	82
C.1	Age and cycling motives on national scale	82
C.2	Cycling purposes sample Tilburg	83
D	Overview of safety datasets	84
E	Background information timeslots	86
F	Analysis	87
F.1	Descriptive statistics absolute and relative deviation	87
F.2	Overview significant positive and negative relationships	87

1.1 Context

A Dutchman without his bicycle is like a Cuban without his cigar.

For many years, the Dutch government has been stimulating its inhabitants to take up cycling as an alternative for public transport and the car. As a result, today, cycling has a central place in the Dutch DNA and people's daily life. Several studies relate the popularity of cycling to the high level of cycling safety in the Netherlands (Jacobsen, 2003; Pucher and Buehler, 2008a,b; Schepers et al., 2014).

Therefore it is hardly surprising that the Netherlands is one of the world leaders in bicycle use and safety.

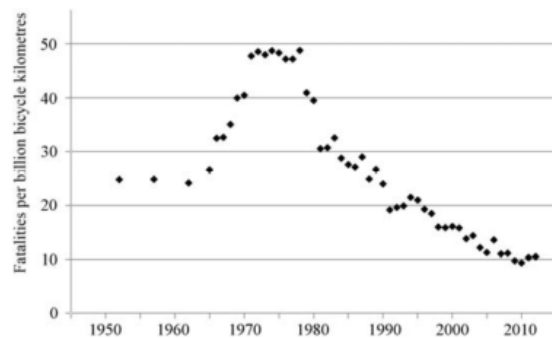


Fig. 1.1.: Numbers of recorded cyclist fatalities per billion bicycle kilometers. Source:Schepers et al. (2014)

In the 1970s the Dutch decided to invest generously in bicycle infrastructure and as figure 1.1 shows, at the same time the number of fatalities per billion bicycle kilometers has shown a substantial reduction. This reduction is impressive considering that the number of motor vehicle kilometers has doubled since 1950 (Schepers et al., 2014). However, now for the first time in years, the number of fatalities, injuries, accidents and collisions in Dutch traffic are increasing again, especially among vulnerable road users like cyclists and pedestrians (Rijksoverheid, 2017). Moreover, the Dutch Central Bureau of Statistics found that for the first time in 2017 more fatalities occurred among cyclists than motorized vehicles (CBS, 2018).

Cyclists are considered a vulnerable group of road users and according to the Rijksoverheid (2008), measures need to be taken to reduce the risk for this type of road users. The current government in the Netherlands has acknowledged the importance of the motive of this initiative. In the recent coalition agreement -in Dutch referred to as a *regeerakkoord*- stimulating cycling and improving cycling safety are considered essential goals and mentions that in the next four years extra budget will be made available to invest in cycling safety (Rijksoverheid, 2017). Not only in the Netherlands, vulnerable road users like cyclists are

getting more and more attention. Globally, vulnerable road users make up the majority of road fatalities. This number is substantially higher in developing countries due to the lack of proper cycling and walking infrastructure. According to Fishman et al. (2012), both actual and perceived levels of safety are deterrent to cycling. To be able to improve cycling safety, first it is important to investigate and understand what factors play a role in both actual and perceived cycling safety.

1.2 Problem statement and relevance

Even though cycling is a massive part of Dutch culture, most cycling research (especially regarding safety) is not from the Netherlands. Due to very different geographical contexts, these results are not always valid for the Netherlands. Besides, most of the cycling research that is available today focuses on *modal choice* and the promotion of cycling when travelling to one's destination from point A to point B. Yet, the route between point A and B itself is a matter of choice which often remains unexplored (Spinney, 2009; Van Duppen and Spierings, 2013).

Safety can be the reason why a person decides not to cycle at all, but it can also be the reason why a cyclist travels a certain route. To advance our (scientific) understanding of human mobility and the transportation system, it is essential to investigate the factors that play a role in cyclists' route choice. From a policy perspective, governments often want to improve traffic safety and therefore they require a framework of both cycling safety and cycling behavior. Additionally, planners need to be able to predict cycling behavior when improving and developing bike infrastructure. For this purpose, large amounts of data are required that are representative of the population in their region (Zhu and Levinson, 2015). Yet, at this point, cycling behavior is not sufficiently researched on that scale to accurately do so.

Most bicycle research up until this point are small-scale studies and deal with limitations for generalizing and interpreting cycling behavior. Consequently, the behavior of cyclists is still poorly understood. Large-scale quantitative studies can therefore help to gain more insight into cycling behavior.

Large volumes of GPS data can help to provide insight into cycling behavior and other social processes that were previously undersampled or poorly understood (Romanillos et al., 2015). Since we are currently living in the era of large computing power, Big Data and Apps, we are now able to collect large volumes of data with relatively low costs and process these large volumes of data (Garrard et al., 2008).

Currently, cycling networks and traffic models are derived from car networks and models which are not at all suited to the behavior of cyclists (Hendriks and Bussche, 2016).

The reason for this is that compared to motorized vehicles, the behavior of cyclists is relatively unknown. Due to the amount of freedom in movements and choices cyclists have compared to motorized vehicles (who are constrained by far more rules and regulations), tackling the behavior of cyclists is notoriously difficult (AMS, 2016; Blokker, 2013).

Mainstream traffic research since the 1950s often uses Wardrop's first principle which assumes people take the shortest route to minimize their 'travel costs' (Wardrop, 1952). Yet, studies investigating route choice found that in fact cyclists often cycle longer routes than the shortest possible route or they are willing to cycle longer routes for better safety conditions (Krizek et al., 2007; Tilahun et al., 2007; Menghini et al., 2010; Broach et al., 2012; Vedel et al., 2017). Ehr Gott et al. (2012) and Heinen et al. (2010) argue that the generalized costs for cyclists are made up differently from motorized vehicles. They argue that for cyclists, factors, such as safety, cannot be ignored because safety can be a factor that is as important (or even more important) than travel time or distance.

However, despite the fact that previous research suggests that safety plays a role in cycling and route choice behavior, people often tend to act differently than expected. In a Dutch context specifically, where cycling is very popular and in general already very safe, would safety even show up to be relevant in choosing to travel a certain route? Using large volumes of GPS data could provide insight into the route choice behavior of Dutch cyclists by comparing their chosen routes to the shortest possible route alternatives.

1.3 Research questions

The general research objective of this thesis is to gain insight into route choice behavior by exploring the relationship between cycling safety and route choice. To realize this objective the following research question will be answered:

Main research question

To what extent do safety-factors of the environment influence cyclists' route choice behavior and how can this be empirically measured using GPS tracks and GIS methods?

To answer this question, several sub-questions need to be answered first. First of all, it is necessary to gain insight into safety factors of the environment that influence route choice. Therefore the following sub-questions will aim to identify and measure these safety factors.

1A. What safety factors of the environment are relevant in terms of cyclists' route choice?

1B. How can the safety factors of the environment be measured?

The second sub-question will elaborate more on the used methods to measure route choice behavior in this project by the answering the following question:

2. How can route choice behavior be measured using GIS methods?

Finally, the final sub-question investigates the influence of safety on cyclists' route choice using statistical analysis.

3. To what extent do these safety-factors statistically influence cyclists' route choice?

1.4 Research scope

The scope of the research, i.e., what is included in this thesis, is reflected in the research objectives and questions stated above. The route choice behavior is investigated by studying why and how far people detour by comparing the chosen routes to shortest possible alternatives.

It is also important to elaborate on what will **not** be covered in this research.

The goal of the literature study is to identify safety-related factors that influence route choice (limited to cyclists). Therefore, safety-related factors that do not directly affect route choice are only briefly mentioned or left out altogether.

Due to time constraints of this research route choice behavior is investigated, but is limited to investigating travel *distance* and **not** travel *time*. Also travel motives are not included in this research, because a person's travel motive is often considered personal information. The data is therefore not available on the scale of individual routes. Finally, transportation forecasting and route-choice modeling are also not in the scope of this research.

1.5 Reader's guide

This section provides a brief explanation of the structure of this thesis. The following chapter, chapter 2, will provide an extensive literature review to identify the safety-related factors of the environment that influence cyclists' route choice behavior. Chapter 3 will further mention how the research is carried out and describe the methods and techniques that are used how to measure route choice behavior. Then, chapter 4 will provide an overview of the used datasets to represent the safety factors. Furthermore, chapter 5 will mention how the data is prepared and modeled using GIS methods. In chapter 6 the results of the data analysis are provided. Finally, in the concluding chapter, chapter 7, the research questions are answered.

This chapter will serve as a theoretical framework that will identify the safety-related factors that affect cyclists' route choice.

2.1 Route choice behavior

Studies regarding bicycle route choice behavior have shown that people do not always travel the shortest route to minimize their travel costs, because the generalized costs for cyclists are made up of more factors than only travel distance. Ehrgott et al. (2012) and Heinen et al. (2010) argue that for cyclists, factors, such as safety, cannot be ignored because safety can be a factor that is as important (or even more important) than travel time or distance.

Travel distance or time however, is one of the most important factors regarding cyclist's route choice. Therefore, the current Dutch government is investing in aligning the safest and shortest routes to guarantee safe and convenient cycling on the network (Rijksoverheid, 2017). Route decisions, in fact, are a trade-off between the cyclist's purpose or motives, the time one has available and the need for safety. This means that for some purposes one has more time available to detour to a route that is safer, but also sometimes the degree of safety does not matter that much at all and the main goal is minimize travel distance or time. An example of this trade-off process is the research of Zimmermann et al. (2017). In their models distance enhances the effect of the road characteristics. They suggest that the way individuals perceive the travel distance of a road is influenced by other characteristics of the road, e.g. slope or safety factors. In reality, people do not choose for either distance or road characteristics; they always make a trade-off. For instance, if a road has a high slope, it will cost more per unit of length, making the road less attractive to a cyclist. According to Zimmermann et al. (2017, p. 189), this behavior is plausible for cyclists, because:

"... travelers might be willing to cope with negative attributes, but more so for relatively short distances."

Also Joolink (2016) found in her research that safety (*choosing a safe route*) is a high-impact factor on route choice. A general conclusion in bicycle research is that cyclists significantly deviate from the shortest path, but that there is a limit to this deviation. The amount of deviation is often relative to the total travel time (or distance) and one's purpose (Standen et al., 2017; Zimmermann et al., 2017; Broach et al., 2012; Winters et al., 2010; Menghini et al., 2010). Also, significant differences in deviations exist between different cycling climates over the world. Safety might play a more prominent role in determining one's route in car-

dominated countries with fragmented and sparse cycling facilities than in the Netherlands, where cycling is strongly rooted in the culture, and the cycling network is already relatively safe and well-connected throughout the whole country (Pucher and Dijkstra, 2003).

In general, safety plays a role in route choice when people prefer or avoid certain areas and infrastructure, and it often explains why people are willing to detour from the shortest path. In their research Heinen et al. (2010), distinguish between objective and subjective safety. Heinen et al. (2010, p. 63) defines objective safety as: "...*'real' safety for cyclists, measured in terms of the number of bicycle-related incidents per million inhabitants.*" and subjective safety as: "*safety that refers to how individuals perceive safety, and is mostly measured in terms of the stated safety experience of users or other respondents.*"

In their research they found that cyclists' preferences in choosing a route are often based on subjective notions of safety (Heinen et al., 2010). Even in the Netherlands, where cycling safety is high on the priority list, Rietveld and Daniel (2004) found that the variation in the amount of bicycle use in Dutch municipalities can be explained by perceived and actual safety. Similarly, the same has been found in a Belgian study of Vandenbulcke et al. (2011) when comparing cycling levels from Flanders and Wallonia.

The following sections will further identify and explain several objective and subjective safety-related factors that affect cyclists' route choice.

2.2 Accidents

Road or traffic safety is currently often measured by the number of accidents, crashes, deaths etcetera. The definition of a traffic accident in the Netherlands is an occurrence that takes place on a public road that causes damage or an injury and includes at least one driver (including cyclists). To this day there are dangers present in traffic and because people travel they are exposed to risk in traffic. Crashes and accidents or the risk thereof result from the interaction between the three traffic safety pillars: road users, vehicles and infrastructure (Schepers et al., 2014). Wegman et al. (2012) found that cycling is associated with a considerably higher risk of injury accidents than traveling by car. This effect is noticeable in figure 2.1 that shows, respectively, the relative amount of road users involved in accidents and the relative amount of road users that were injured.

Aside from the severity of the injuries, two types of accidents can be categorized for road users. First of all, there are *one-sided accidents* in which only the cyclist is involved. This type of accident can be caused by either an interaction between the cyclist and the infrastructure or a person's cycling skills. Secondly, there are *two-sided accidents* in which both a cyclist and

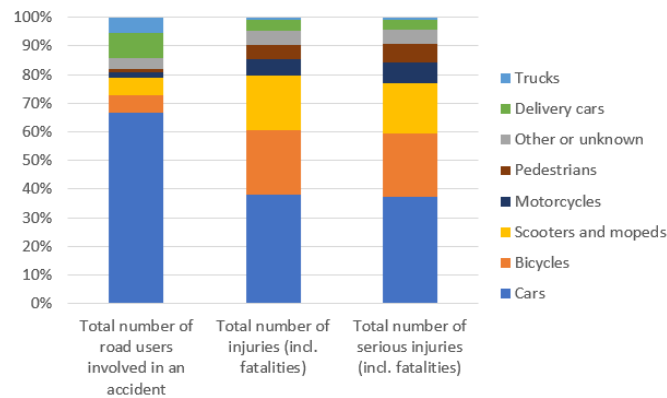


Fig. 2.1.: Relative amount of road users in traffic accidents 2001-2015 (BRON)

another party interact. This other party can be a motorized vehicle, but also a pedestrian or another cyclist. Pinder (2001) states that:

"The network of streets both produce and contain memories. [...] Each new angle, each new experience on the streets could produce another memory."

Therefore people that witness, are familiar with or experience accidents or near-misses themselves, might associate these events with a feeling of unsafety on that specific location. Shankwiler (2006) even found that people remember what they perceive to be dangerous route segments better than 'normal' route segments. This can result to people avoiding specific road segments that they associate with this *feeling* of unsafety but also road segments that have a higher risk of accidents due to a number of other reasons. Van Duppen and Spierings (2013) found that instead of avoiding these parts, people might also adopt a different riding style.

One of the biggest problems with bicycle accidents, however, is that they are often not registered; meaning that a lot more accidents occur than the official numbers suggest. Therefore, the relative amount of cyclists represented in figure 2.1 is even higher in reality than the figure suggests. Especially non-fatal accidents are highly under-reported because in that case the police are often not involved. This is problematic because then the information about accidents is missing; meaning that it is not possible to know what, where and why a traffic accident has occurred. Especially between 2009 and 2013 (see figure 2.2) the quantity and quality of reporting has been deficient which makes it difficult to address the exact cause of the accident (SWOV, 2017). In a pilot in the province of Friesland, this was tested when analyzing the traffic victims that visited the emergency room. The results showed that 80% more people showed up to the emergency room as a result of a traffic accident than the police had registered (Verkaik, 2017). Many studies have acknowledged the problem of under-registration (Tirry and Steenberghen, 2014; Wijnhuizen and Aarts, 2014; Reurings et al., 2012; Langley et al., 2003).

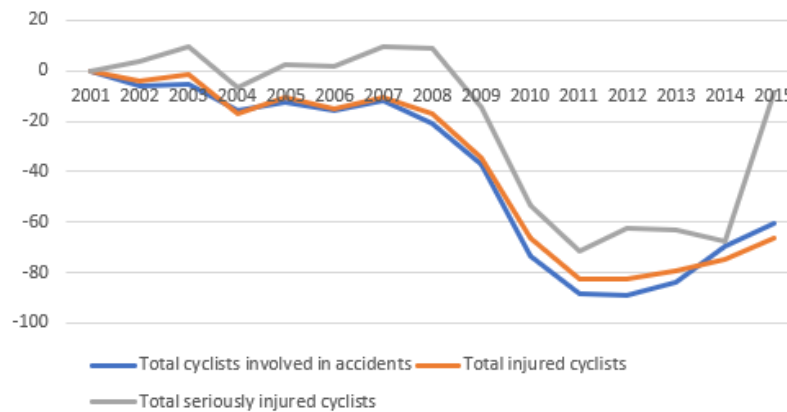


Fig. 2.2.: Relative development of bicycle accidents in % over period 2001-2015. Baseline = 2001

Even though accident data is often used to estimate road safety, accidents are not the sole indicator of safety. Safety should therefore not only be measured by the results (accidents) but also by the causes. Accidents often reflect the tip of the iceberg in terms of perceived safety (Van der Schaaf et al., 2013). Sanders (2015) found that cyclists experienced near misses more often than actual crashes and argues that this is problematic since it represents traffic risk for cyclists that is generally not reflected in official crash statistics and arguably may be even of more influence on perceived risk and safety than actual collisions. After all, locations that are accident-free are not per definition safe roads.

2.3 Road type

Bicycle studies all over the world, in general, have shown that people have a strong preference for bicycle-dedicated infrastructure, because cyclists tend to feel safer on this type of infrastructure (most likely due to the allocated place for cyclists on the road) compared to cycling in mixed heterogeneous traffic (Buehler and Dill, 2016; Beura et al., 2017). In the study of Broach et al. (2012) in Portland (USA) this preference was also found. The cyclists appear to place relatively high value on the presence of off-street bike paths, neighborhood bikeways with traffic-calming features (such as bicycle boulevards) and bicycle bridges or tunnels. The study showed that around 50% of kilometers cycled occurred on bicycle infrastructure even though these facilities only account for 8% of the bikeable road network (Broach et al., 2012). This is consistent to the results of Winters et al. (2010) and an extensive literature study of Heinen et al. (2010) and suggests that people go out of their way to use these facilities. Fishman et al. (2012) found that one of the cyclists coping strategies included avoiding to cycle in mixed traffic, even if it means cycling on a longer route.

It seems, however, that there is a limit to detouring for bicycle dedicated infrastructure. Sener et al. (2009) argue that especially in the USA a lack of (connected) bicycle facilities contributes to the bicycle use in general. Because of the limited and fragmented supply of bicycle dedicated infrastructure in many countries cyclists would have to travel quite some distance to reach or stay on the bicycle-dedicated infrastructure (Pucher and Buehler, 2008a). Moreover, Caulfield et al. (2012) found that cyclists actually prefer connected bicycle facilities over bicycle facilities in general. Connected cycling routes and an accessible, continuous cycling network will allow direct traveling will minimize interactions with other road users by a sudden need to cross the road, reducing the risk of accidents.

Even though Broach et al. (2012) found that cyclists go out of their way to use bicycle facilities, they also note that it is unlikely that a cyclist would choose a route that is seven times longer to avoid traveling on a busy road without a bicycle lane. In that case, the person would either accept a route that is safe enough (even though one could be safer when making a bigger detour) while cycling or would not be traveling by bicycle at all.

In contrast, in the Netherlands, the creation of a complete and integrated system of cycling routes has lead to such a wide range of facilities for cyclists to cover almost any trip on either completely separate paths, lanes or traffic-calmed residential streets Harms et al. (2016). Pucher and Buehler (2008b) mention that these measures (for a large part) explain the high quality cycling network in the Netherlands. They state that these measures enhance the overall bicycle network because they offer much safer, less stressful cycling compared to cycling on streets that are filled with fast-moving motorized vehicles.

Bicycle-friendly countries such as Denmark or the Netherlands have a long history of building bicycle paths separated from roadways, but it is getting increasingly popular in other cities around the world as well (Schepers et al., 2017; Buehler and Dill, 2016). Standen et al. (2017) found that in Sydney (Australia) people will take a longer route to use the separate bicycle infrastructure but to a lesser extent for commuting purposes. Routes for commuting purposes should be as direct as possible because travel time has a more significant impact in that case than other factors contributing to a person's route choice. The preference for separate bike paths or infrastructure applies particularly to women, non-regular cyclists and people over the age of 30 (Standen et al., 2017). Also, a Dutch study found that people using an e-bike value the presence of separated bicycle path more than regular cyclists (Van Genugten and Van Overdijk, 2016).

The Dutch street hierarchy, already since the 1960s, follows the *homogeneity principle* (Schepers et al., 2017) and therefore makes sure that road users (that share the same road) with differences in speed, direction and mass are not too large. It regulates that mixing cyclists and motorized traffic happens safely regarding speed differences. Due to substantial speed differences between different types of road users, separate lanes or paths can help separate and protect cyclists from motorized vehicles. Tingvall and Haworth (2000) consider 30km/h a safe speed where vulnerable road users can be mixed with motorized vehicles. This threshold

should prevent severe injuries in case of crashes between road users. This type of roads are called *access roads* (Schepers et al., 2017).

According to Jacobsen (2003) and Schepers et al. (2014), a high incidence of cycling increases safety for cyclists in general because of the "safety in numbers" phenomenon. Basically, it means that due to a high share of bicycles on the road, drivers of motorized vehicles are more aware of cyclists and therefore adapt their behavior. When the share of cyclists is low, the awareness of cyclists among drivers is also low. In the Netherlands, some access roads are different from others and are designed to promote cycling safety on these roads specifically. On streets with relatively a high number of cyclists compared to the number of motorized vehicles, bicycle boulevards (in Dutch *fietsstraten*) are created. Motorized vehicles are allowed to ride on a bicycle boulevard, however, cyclists are considered the main road users. Therefore motorized vehicles are seen as 'guests' on these types of roads, and the infrastructure helps enforce this. In their study, Broach et al. (2012) state that the safety in number phenomenon is also part of the reason why bicycle boulevards are valued so highly. Even though many types of bicycle boulevards exist over the world, an example of a bicycle boulevard is given in figure 2.3f.

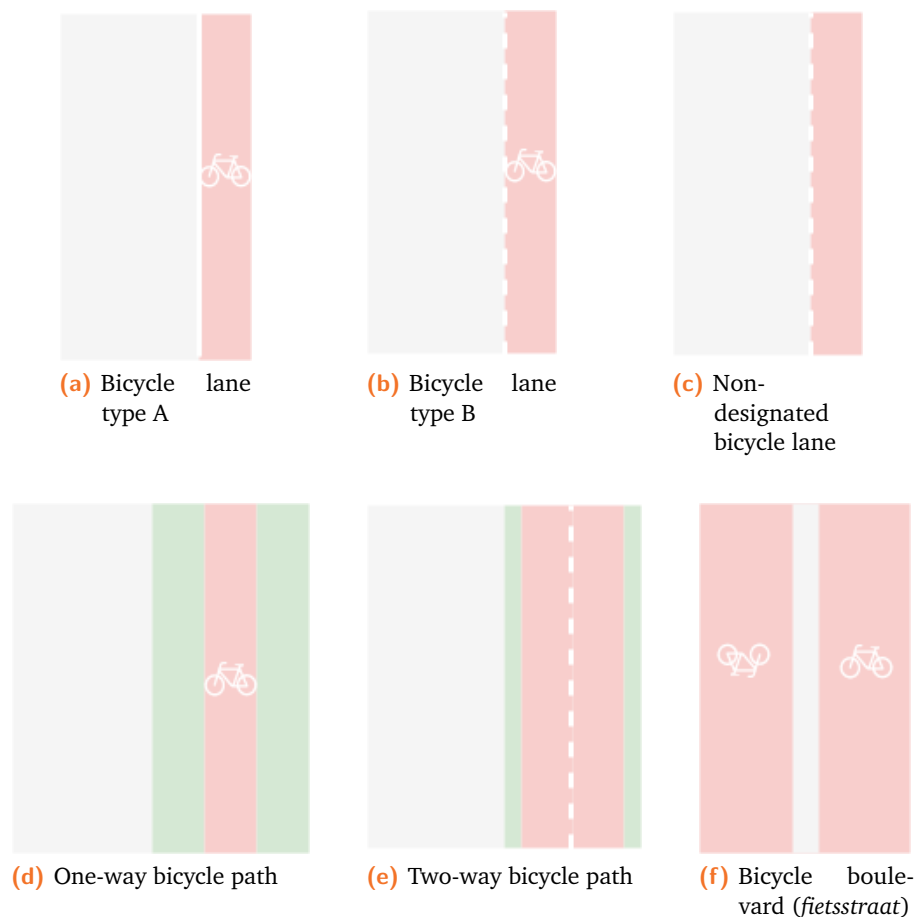


Fig. 2.3.: Types of bicycle infrastructure

Another measure that is taken to increase safety for cyclists is the creation of bicycle-specific grade-separated intersections (e.g., tunnels and bridges) in order to cross high-speed roads safely. Another example is the Dutch concept of 'shared space'; by organizing the public space in such a way, it allows different road users to mix without providing them instructions on how to do so. Because the perceived subjective safety decreases due to the absence of any guidelines, road users are more cautious and alert, making eye-contact to communicate with other road users to determine what behavior is acceptable (Methorst, 2007).

The second category of roads that exist in the Netherlands are called *distributor roads* and allow a speed limit ranging from 50-70km/h. On these roads cyclists are, for safety reasons, separated from motorized traffic by providing the cyclists separate bicycle paths or lanes (Schepers et al., 2017). By assigning scooters to the road on distributor roads since 2000, the speed differences on the bicycle lanes and paths were strongly reduced (Reurings et al., 2012), contributing to both the objective and subjective safety. Different bicycle facilities exist for this type of roads. Figure 2.3 shows the different types of bicycle infrastructure. Figure 2.3a, 2.3b and 2.3c are three different types of bicycle lanes. Bicycle lanes provide cyclists with their own place on the road, but there is no physical separation (e.g., shoulders) that prevent cars from driving on the bicycle lane. On almost half of all 50 km/h roads in the Netherlands, cyclists still share the road with motorized vehicles, among which heavy goods vehicle traffic (e.g., buses, lorries) (Van der Knaap, 2017). Figure 2.3a shows a barrier line; meaning that cars are not allowed to cross it. Figure 2.3b and 2.3c both show a broken line; however they are different. Figure 2.3b is often used to provide cyclists with their own place on the road but allowing cars to cross it for parking; resulting in potential interactions between cyclists and drivers (Heinen et al., 2010). Parking spots are often located directly next to the bicycle lane. Figure 2.3c, showing no cycling symbol, provides no real protection for cyclists. This lane (especially when they are non-colored) only *suggests* that this lane might be a location where cyclists can ride. In reality, cyclists may use this lane to ride their bicycle, but it is not compulsory. Often these lanes are very narrow and consequently very uncomfortable to use as a cyclist. Also other road users, e.g. cars, are allowed to use this lane (or part of it).

Finally, the third type of roads that Schepers et al. (2017) categorize in the Netherlands are called *through roads*; these roads have speed limits of more than 100 km/h and consequently cycling is prohibited on these type of roads. Near these types of roads, one would only find separate bicycle paths with a substantial physical barrier in between or service roads (which allows mixed traffic and in most cases maximum speeds around 60 km/h). Figure 2.3d and 2.3e shows two different bicycle paths; this type of bicycle infrastructure is known for its physical separation from the road and sometimes also allows two-way bicycle traffic.

An overview of all types of roads, maximum speeds and the location of cyclists is given in table 2.1

Type of roads	Speed limit	Location of cyclist
Access roads	30 km/h	Mixed with other traffic
Distributor roads	50 or 70 km/h	Separated from motorized traffic by bicycle lanes/paths
Through roads	≥ 100 km/h	Cycling not allowed

Tab. 2.1.: Location of cyclists on categorized roads. Source: (Schepers et al., 2017)

2.4 Road surface

Pucher and Buehler (2008b) mention that in the Netherlands aside from investing in the expansion of the separate cycling facilities, the design, quality and maintenance (e.g., road surface) of the cycling network has continuously improved the cycling network by providing safer, more convenient and more attractive cycling every year.

When investigating cycling accidents, the road surface is one of the critical factors that is to blame due to loss of control Reurings et al. (2012); Wijnhuizen and Aarts (2014). This is corroborated by Dozza and Werneke (2014) who found that poor maintenance of the road increases the risk of accidents tenfold. Aside from the maintenance of the road surface, the type of road surface itself can have this effect as well. For instance, cobblestones that are wide apart can lead to cyclists losing control of their bicycles. Also, because of rain, the cycling infrastructure can become slippery (some types more than others) causing cyclists to slip resulting in accidents. Hölzel et al. (2012) also link cyclists preference for specific type of road surface to energy input. Comfortable cycling infrastructure requires smooth rolling over the road surface by providing the lowest possible energy input. In their research regarding the resistance of the road surface, they found that asphalt and concrete slabs are advisable materials for comfortable cycling pathway surfaces and self-binding gravel and cobblestones are less suitable (Hölzel et al., 2012).

In a Dutch study about route choice, Van Genugten and Van Overdijk (2016) found that the quality of the road surface turned out to be the second most important factor that influences cyclists' route choice. Especially regarding trips with an overall longer distance, the impact of quality of the road surface on route choice increases. On short trips, however, a gain in time turned out to be more important than the road surface quality. Also in Dublin, people felt a strong aversion for roads with poor quality surfaces. Almost 32% of the respondents stated that they would alter their routes to avoid these roads (Lawson et al., 2013).

Stinson and Bhat (2003) found that the surface quality is especially important for women, older people (due to decreasing stability), and experienced cyclists (probably due to long cycling distances and cycling for recreational purposes).

2.5 Obstacles

In the research of Reurings et al. (2012) 669 cycling accidents were investigated that were registered in the emergency room. A large part of these accidents was explained by crashing into objects that are part of the infrastructure (e.g., curbs or bollards) or people. The presence of obstacles, either people or objects are also sometimes called 'hindrances' and can have a significant effect on route choice.

Rietveld and Daniel (2004) describe the effect of hindrances as follows:

"The generalized costs of a route depend on the quality of the infrastructure and elements connected with the hindrances experienced by the bicycle user with respect to the traffic or the other users of the road network." (Rietveld and Daniel, 2004, p. 540)

Parking

Around one-third of the cyclists in the research of CROW (2015) stated that they often experience inconvenience from parked cars. Again, the width of the bicycle lane has a key impact on the degree of safety cyclists experience while cycling next to parked cars. Parked vehicles can reduce sight distance for cyclists. Also, to park, motorized vehicles need to cross the bicycle lane to park their cars resulting in interaction between cyclists and motorized vehicles. Furthermore, drivers can open doors without looking; thereby possibly hitting a cyclist with the car-door. Cyclists' behavior regarding parked cars differ depending on their attitude. Van Duppen and Spierings (2013) found that in their study in the Netherlands people do not avoid these type of roads, but they adopt a different driving style. In Texas, USA, however, cyclists would be willing to cycle more than 6 minutes extra to avoid parallel parking on their bicycle route (Sener et al., 2009).

Whereas in the Netherlands bicycle lanes are rarely (either fully or partly) occupied by other vehicles, in India a bicycle lane is a popular place for street vendors or to use for parking. Due to the high roadside commercial and parking activities, cyclist state that the bicycle lanes offer (very) poor services to bicycles (Beura et al., 2017).

Intersections and control

One of the biggest irritations of cyclists is stopping because cycling requires physical effort to get going before getting into a certain flow once you reach some speed. Stopping, therefore, causes cyclists to lose their momentum (Blokker, 2013). Hence, due to delays, intersections, traffic lights and stop signs often irritate cyclists even though they are necessary for regulating traffic and creating safe traffic situations. Many studies have found that cyclists indeed

generally avoid intersection control (e.g., traffic lights, stop signs) when choosing a route (Van Genugten and Van Overdijk, 2016; Broach et al., 2012; Rietveld and Daniel, 2004; Stinson and Bhat, 2003). Winters et al. (2010) even found that the traffic lights and signs explain why people detoured more than 10% compared to the shortest route. However, there are also cases in which people prefer traffic lights or stop signs according to Stinson and Bhat (2003). It is often considered safer than no intersection control, and the traffic lights and stop signs might help cyclists to cross dangerous intersections safely. Dozza and Werneke (2014) found in their study that cycling near a regular intersection increased the risk of accidents by four times, due to the interaction with other road users. In that case, the positive effect of the traffic lights or signs outweighs the negative in such locations and make the intersection even attractive according to Broach et al. (2012). In their research, they found that cyclists generally avoided intersection control except in case of crossing high-traffic streets. Whether this can be explained by the signals actually reducing the delay or the increased perceived safety is unsure.

Be that as it may, especially experienced cyclists and commuters tend to avoid controlled intersections on their route because travel time can be more important (even though covering a greater distance in total) and they feel safe enough to cross a street without the help of traffic lights (Heinen et al., 2010).

2.6 Traffic intensity

Investigating cycling accidents has shown that many accidents also happen because of the interaction with other road users. Cyclists either lose control due to physical contact or last-minute avoidance of pedestrians, cyclists or motorized vehicles (Wijlhuizen and Aarts, 2014). When more road users are sharing the road and the infrastructure is becoming more crowded, less room is available. Therefore the chance of more physical interaction between bicycles and other road users can increase (Reurings et al., 2012). This is often the case on certain parts of the network in rush hour periods. Then, car drivers are forced to pay greater attention to other car drivers as well as cyclists and pedestrians. This can result in distracting their attention from cyclists, especially in countries where there is a low share of cycling.

Yet, in their research, Van Duppen and Spierings (2013) found that people deal differently with crowdedness; some people change their route to avoid these high traffic volumes and others take the crowded route but adopt a specific riding style to ensure safety. Most bicycle studies conclude that cyclists have a negative perception of roads with high-traffic volumes and feel less safe when sharing the road with high traffic volumes (Dill and Gliebe, 2008; Broach et al., 2012; Lawson et al., 2013). In Dublin, almost 33% of the respondents in the study of Lawson et al. (2013) would alter their route to avoid these roads. Furthermore, Broach et al. (2012) found that in Portland (USA) despite their sample was primarily made

up of cyclists with a 'road warrior' mentality, they would only use high-traffic volume roads if the alternative were a detour that was twice as long or included steep hills. Yet, also in the Netherlands and Denmark cyclists experience these types of roads as stressful or in a more negative manner compared to other types of roads (Boekhoudt et al., 2017; Vedel et al., 2017). This aversion towards crowding does not only apply to interactions between motorized vehicles and cyclists but also between cyclists.

Since 2005 the number of bicycle kilometers in the Netherlands has increased by 11% and it shows, especially in cities in rush hours large clusters of bicycles and even bicycle traffic jams. Even though studies (Jacobsen, 2003; Schepers et al., 2014) show that cyclists feel safer when the number of cyclists is high, it also causes unsafe situations in big Dutch cities because the bicycle paths and lanes are getting more crowded every year (CROW, 2016). In case of bicycle traffic jams, it can also create unsafe situations outside of the bicycle infrastructure as well due to crossing traffic that is blocked by the number of cyclists that are waiting for a traffic light (Van Duppen and Spierings, 2013).

Also the composition of road users on the Dutch bicycle path has changed a lot over the years. Nowadays, the speed differences on the bicycle path are high since cyclists share their path with e-bikes, light mopeds, carrier tricycles (which take up quite some space) and until recently speed-pedelegs (CROW, 2016; Reurings et al., 2012). In cities with a high intensity of bicycle path users, lane width is therefore also considered an important safety-factor (Reurings et al., 2012). Even in Copenhagen, where bicycle paths are in general much broader than in the rest of the world, people were willing to cycle 1 km longer to avoid high levels of crowding on cycling tracks.

2.7 Social safety

Current research on cycling safety does hardly address the issue of social safety. Even though Sanders (2015) did look into it during the study, the results were never published. Even though the Netherlands is a relatively safe country, cycling (in the dark) alone in some locations can feel unsafe to people. In other countries, this feeling of unsafety can also be present when cycling in some areas during the day; resulting in people avoiding certain streets or neighborhoods. The results of the Safety Monitor of 2016 showed that 16% of all Dutch people sometimes feel unsafe during the evening when they are outside in their neighborhood, and 3% stated that they often feel unsafe. The number of people actually taking a detour because of this feeling of unsafety is much lower: only 2% often detours and 10% state that they sometimes detour; meaning that most people (around 80%) never detour because of a feeling of social unsafety (CBS, 2016c). The effect of social safety might be more relevant in other countries than the Netherlands, but it might also be possible that people detour for this reason without even being aware of it. Van Duppen and Spierings (2013) found in their study some cyclists did feel uneasy in the evening due to malfunctioning street

lights, but it did not deter them from cycling, nor did they found out if people changed their route because of social safety reasons.

Gender can play a role in assessing the social risks of cycling alone in the dark (Rietveld and Daniel, 2004) especially on locations that are poorly lit, quiet and deserted. These locations are often avoided due to safety reasons (especially after dark). An example of these locations are bicycle tunnels. The feeling of unsafety regarding bicycle tunnels might exist when they are located in a deserted place, have a bad reputation and/or the frequency of use is low. The idea of social control plays a significant role, meaning that people will trust nothing will happen when many people are nearby or are passing by (Vis, 1994).

2.8 Illumination

It is commonly known that people use the bicycle less often as a transport mode in case of bad weather conditions and nighttime. However, no research currently exists that looks into changing routes once it is dark outside. Analyzing the registered accidents, Reurings et al. (2012) conclude there is a relatively higher risk of accidents when cycling in the dark when vision is limited. This is especially the case early in the morning before sunrise. Reasons for this can be limited visibility and people having less attention for other traffic. Therefore it is likely that people deviate from their 'normal' route once the visibility decreases to a route that has sufficient street lights. By adding street light on intersections, the risk of accidents significantly decreases (Elvik et al., 2009). In these situations, cyclists might even detour to find such facilities before and after sunrise. This can show contrasting behavior regarding e.g. the use of intersections or the effect of social safety (see section 2.5 and 2.7) during different times of the day.

2.9 Conclusion

Studies regarding bicycle route choice behavior have shown that people do not always travel the shortest route to minimize their travel costs, because the generalized costs for cyclists are made up of more factors than only travel distance. Also factors such as travel motives, travel time and safety play a prominent role in explaining route choice behavior. In figure 2.4 an overview is provided of the safety-related factors of the environment that have shown to influence route choice in previous bicycle research. However, cyclists might show different behavior during different times of the day. Also, not all safety factors might be as relevant to investigate during every part of the day. More information regarding the four timeslots can be found [here](#).

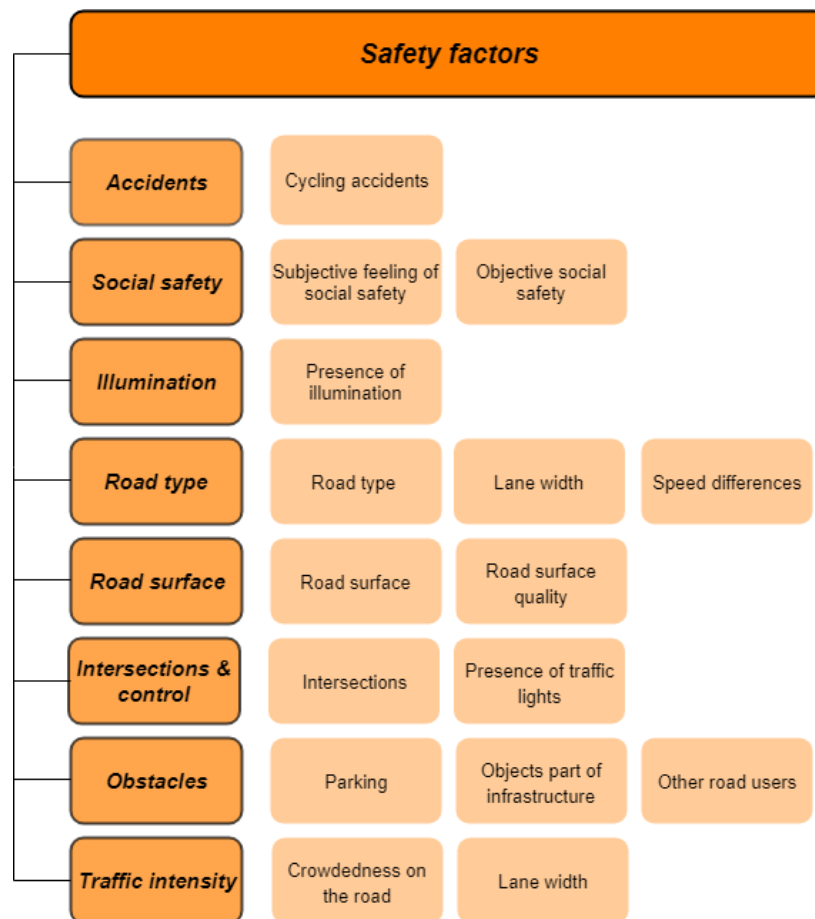


Fig. 2.4.: Summary of safety factors effecting route choice

This chapter will indicate how the research will be carried out step by step and justify the methods that are used for carrying out the study. The mentioned scripts that have been used can be found on Github¹.

3.1 Measuring route choice behavior

3.1.1 Stated preference studies

Route choice behavior has been researched in many different ways. The most commonly used method is *stated preference*. In their simplest form respondents are asked to rank their preferences for different types of facilities. In a more advanced form of stated preference surveys, such as in Tilahun et al. (2007) respondents are given route options from which to choose so that they have to make a trade-off between features such as a separated bike path causing longer travel time.

Even though stated preference research is valuable to analyze the effect of specific parameters without noise, Vedel et al. (2017) argue that there are also limitations to this method. For instance, if the preference for a particular facility depends on parameters that are not included in the research. Yet, the biggest limitation of stated preference studies is that there is no tie to actual behavior because it does not always reflect the reality of the individual's route choice options. Therefore, the preferences often do not manifest in reality (Sanders, 2013; Broach et al., 2012; Winters et al., 2010; Tilahun et al., 2007). Finally, most stated preference studies are small-scale studies. They are often undersampled, biased and are not representative to generalize cycling behavior because these studies deal with many limitations.

3.1.2 Revealed preference studies

Before GPS and Apps were widely available, revealed preference studies asked cyclists to recall their routes. Nowadays, cyclists are easily tracked using either a GPS device or via Apps that use the built-in GPS on their smartphone. Revealed preference studies are now considered as an established way to study route choice behavior (Romanillos et al., 2015). Since GPS is getting more and more accurate and GPS is widely available large-scale studies can easily be set up at low costs (Hood et al., 2011; Garrard et al., 2008). This provides opportunities to gain more insight into the previously undersampled cycling research. Also,

¹<https://github.com/cynthiadevos/ThesisGIMA>

possible gaps and ambiguity are avoided while identifying routes people in reality use (Zhu and Levinson, 2015).

Even though GPS studies are considered valuable, some challenges and limitations still exist. The quality of GPS data is influenced by external conditions, and therefore open space and clear skies are ideal for collecting accurate GPS data (Casello et al., 2011). This might lead to some inaccuracies in case a cyclist rides between or close to large structures, through tunnels or when it is clouded outside. Another challenge in revealed preference studies is the sampling of the cyclists, because it determines to a large extent for what purposes the data can be used and generalized. For instance, the Strava app uses GPS to record (mostly) sportive achievements which makes it unsuitable to research commuting exclusively (Edwards, 2017). Many cycling studies deal with the sampling problem; the targeting of cyclists in Sener et al. (2009) and Broach et al. (2012) lead to a sample of confident cyclists with a road warrior mentality.

Finally, Tilahun et al. (2007) state that revealed preference observes only the final consumer choice and does not take into account how the cyclists came to their final decision or how cyclists would act in case of future- or fictional situations.

3.1.3 Chosen method and study area

To study route choice behavior in relation to safety, the choice has been made to use a GPS based revealed preference method. The way route choice behavior has been investigated in the past, is by comparing cyclists' chosen route to an alternative route.

In this project, the chosen route is compared to the shortest possible alternative. The relationship between safety and route choice behavior, therefore, is investigated by comparing the safety attributes on the chosen and shortest route alternative. Then, it is examined whether the amount of deviation from the shortest route can be explained by the difference between the safety attributes on both routes.

As is mentioned in the previous sections, one of the major challenges in route choice behavior research is the sample size. In this project, the cycling data that has been used is travel (GPS) data from the largest cycling project in the Netherlands: the *Fietstelweek*. For the chosen study area this means that more than 5,000 trips can be analyzed.

A second challenge that has been tackled is the risk of under-reporting. During this large-scale, one-week project of the *Fietstelweek*, all movements of the participants are registered automatically. No actions are required from the participants to start and end recording of their cycling trips.

Finally, a third (still remaining) challenge in this type of research is the sampling process. A huge benefit of using the *Fietstelweek* data is that it is finally possible to investigate cycling for everyday use on a large scale. The data collection method and the promotion of the project among a wide variety of (potential) participants explain why this is the case. However,

even though the Fietstelweek project has made great efforts to reduce the limits and potential biases related to the sampling process, still some remain and have to be taken into account. More information about the Fietstelweek project, the collection method and representativity of the sample can be found in the [next](#) chapter.

Instead of investigating route choice behavior on a national scale, the choice has been made to study a smaller geographical unit to enable fast data-processing. The study area in this research is the municipality of Tilburg, located in the province of Noord-Brabant in the Netherlands. The choice for the study area of Tilburg is based on the high cycling levels during the Fietstelweek and the many available data sources with regard to the safety factors that can be measured. One of the most essential safety datasets is the Fietstersbond (cycling network) data: a rich dataset that not only exists of road locations, but is also enriched with many different safety attributes. Ideally, this data should be known for all streets in the entire study area, but in reality a lot of information is still missing. Compared to other municipalities with high cycling levels, the municipality of Tilburg has the smallest share of missing and unknown data for these safety attributes.

3.2 Research methods

This thesis can be divided into several activities. An overview of these activities, and thus the thesis process, in figure 3.1.

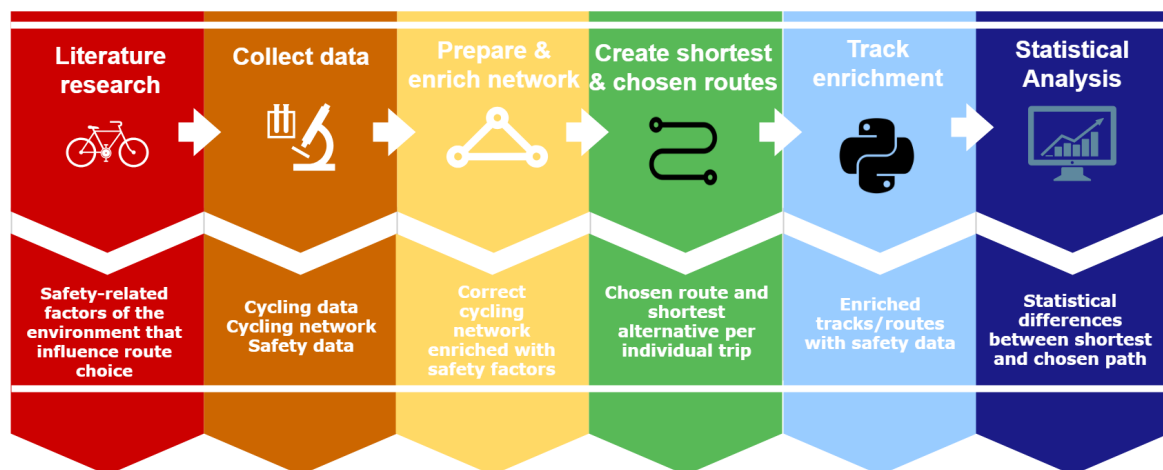


Fig. 3.1.: Conceptual model of the thesis process

3.2.1 Prepare & enrich network

Because there are topological errors in the used cycling network, the cycling network for the research area is altered slightly. This would otherwise result in a not connected network, where in reality roads are actually connected. It is important to correct this, because otherwise the chosen routes do not result in a single line and also it would not be able to create the shortest route alternatives for each route. Furthermore, roads that do not belong in the cycling network are removed and the length of all road segments are re-calculated after the correction of the network. To correct the cycling network, a Python script (*RepairNetwork.py*) is used that manipulates the geometry of the cycling network so that the right road segments connect (Scheider, 2017). This is done based on the connection information between roads that is provided by the Fietzersbond.

To the geometry of the network, *safety* data is added that represent cycling accidents, social safety, illumination, road type, road surface, intersections & control, obstacles and traffic intensity. Because the safety data is available in different formats (i.e., points, lines and polygons) several geo-processing tools and spatial join methods are used to enrich the cycling network with the safety data. More about the data preparation and the enrichment of the network can be found in [chapter 5](#).

3.2.2 Create shortest & chosen routes

To determine the relationship between route choice and safety-factors the shortest and chosen path need to be compared.

Extracting chosen routes in study area

Because the cycling data is provided for the whole country, first the routes that start, end and/or go through the municipality of Tilburg will need to be extracted. Based on the geometry of the boundaries of the study area only routes are selected that at least have one road segment in the study area. To create the chosen routes in the study area, a series of SQL-queries (*SelectionRoutesTilburg.sql*) are used to match the corresponding geometry of those routes that are relevant for to the route information.

Shortest routes

Based on the chosen routes, the start and end locations are determined. Using these locations, the shortest route alternatives are created for all routes. The algorithm that is used to find the shortest route is based on the Dijkstra algorithm (Geertman and Ritsema Van Eck, 1995).

The process of finding the shortest route is explained in figure 3.2. In this figure, the circles represent the nodes and lines represent the edges of the network. In case of roads, the nodes connect the edges where a relationship exists between edges. Also, direction is an important characteristic of a network. In this example all edges are bi-directional, which is visualized by showing two lines, meaning that it is possible to cross the edges (or roads) both ways. In case of unidirectional edges, the network will not allow it to cross the edge in both directions. To determine the cost of traveling over the network, costs will need to be assigned to the edges. The cost of edges can be many things that are considered as an impedance. Standard costs in route analysis are based on travel distance or travel time (Cheng and Chang, 2001).

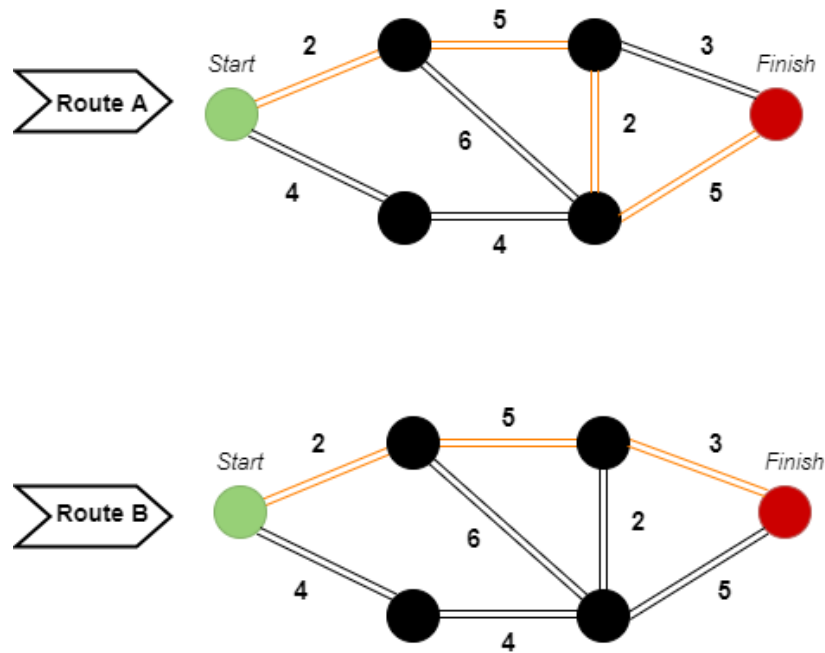


Fig. 3.2.: Shortest path versus alternative path

In this example, the edges are provided with costs (e.g., distance) which means that traveling the network from the start to end location via route A will lead to a summarized cost of 14. The Dijkstra algorithm finds the optimal (e.g., shortest) path between the start and end location; in this example the most optimal route from start to end would be route B with a summarized cost of 10.

To efficiently perform the shortest path analysis for all the routes in the study area a Python script (*Shortest_routes.py*) is used that calls different route analysis tools in ArcGIS.

3.2.3 Track enrichment

After the network has been enriched and both the chosen and shortest route alternatives are created for all the routes in the study area, the routes are enriched with the safety data. Afterwards, the safety data for the traveled roads is aggregated to safety data on route-level.

The list below shows the four different aggregation methods (and the corresponding safety factors) that have been used to calculate the safety data on route-level. More information about the used datasets to measure the safety factors can be found in the [next](#) chapter.

1. Summarize the absolute number of X to route totals

For example: On route A, a total of 10 accidents have taken place

- accidents
- traffic lights

2. Calculate density of X per kilometer

To correct for the total travel distance of the routes, the following factors have been calculated in relative terms.

For example: On route A, 2 accidents per kilometer have taken place

- accidents
- traffic lights

3. Calculate the distance weighted average of X per route

To correct for the total travel distance of the routes, distance has been taken into account while calculating the mean. *For example: Route A has an average lemon score of 7.5*

- number of crimes (per 1000 inhabitants)
- subjective social safety
- speed

4. Calculate the coverage of X per route in %

For example: On route A, 80% of the roads have a good road surface quality

- illumination
- road type
- road surface
- road surface quality
- obstacles
- roads that belong to intersections
- road level (roads that are located next to busy roads)

Missing value problems

However, due to the many missing values in the original data, it is impossible to compare the safety attributes on the chosen route to the shortest route for most of the safety factors

mentioned above. Figure 3.3 shows an example of the problem. In this example, route A (the chosen route) exists out of 30% well-lit roads and 20% poorly lit roads. In total this does not add up to 100% because 50% of the total route has missing values for the degree of illumination. Route B (the shortest route) exists out of 10% well-lit roads and 70% poorly lit roads; resulting in 20% missing values for the degree of illumination. Due to the presence of many missing values in these datasets, comparing percentages of safety attributes is impossible.

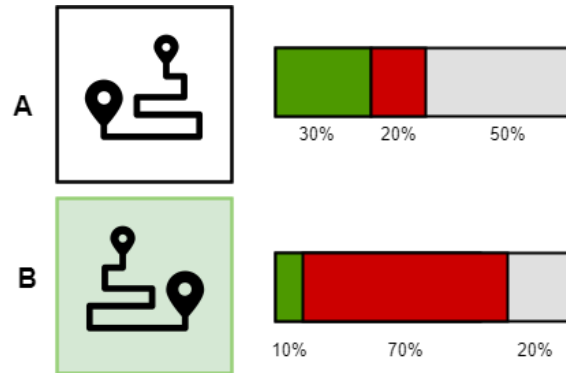


Fig. 3.3.: Comparing safety attributes

Missing value solutions

Three scenarios exist in which the following solutions are used to deal with this many missing values:

1. Less than 50% missing values for categorical and ordinal data

- Apply machine learning methods to estimate missing values
- Used number of routes for analysis: 5,049

By using supervised machine learning (see figure 3.4) a model was created using the known data for training the model. To limit over-fitting also cross-validation has been used. After the model had been trained, responses were predicted for the missing values. The Classification Learner Tool using the Fine Decision Tree method in Matlab is used to train and predict the classes. In the end, the missing values are replaced with the predicted values. More information about the trained models can be found in Appendix A.

It is important to set a lower threshold when determining which factors are safe to predict with machine learning techniques. The amount of data that is known is used for training the model. However, if the amount and variation in the data is low, it is doubtful that the sample is substantial enough to predict the unknown data. The chosen lower threshold is 50%; meaning that at least 50% of the safety data should be known that can be used for training and validating the model.

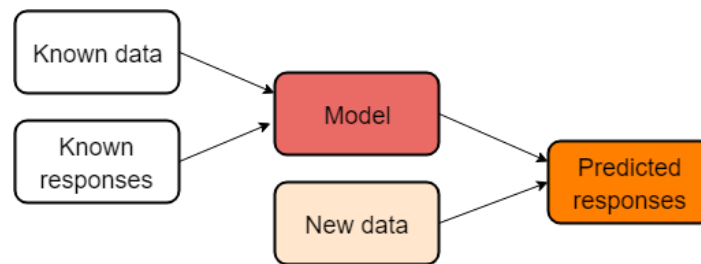


Fig. 3.4.: Supervised Machine Learning proces. Based on Mathworks (2018)

2. More than 50% missing values for categorical and ordinal data

- Only use routes with max 25% missing values and estimate missing values using machine learning
- Used number of routes for analysis: 2,825

The Lemon study, that investigates people's subjective social safety, is only known for 36.5% in the route-level network. This means that still 63.5% needs to be predicted which is more than the 50% lower threshold. However, the subjective social safety is critical to investigate and this data is the only dataset that might directly relate people's subjective feeling of safety to cyclists behavior. The other safety attributes measure this only indirectly and therefore it is decided to investigate this relationship as well. The ratings for subjective social safety will be reclassified to classes and the Fine Decision Tree method will be used to predict the missing values. However, to make sure that the influence of the missing values is not too big, not all routes will be used in the analysis. Therefore, only the routes will be analyzed that have a maximum of 25% missing values per route. By using only the routes that have a maximum of 25% missing values, the effect of estimated values will be limited in the overall distribution.

3. Less than 10% missing values for continuous data

- Estimate missing values using the 'mean'
- Used number of routes for analysis: 5,049

As can be seen in table 3.1 almost 100% of the crime data is known for the part of the cycling network that comprises the routes. Therefore it is safe to say that the missing 0.1% of the data can be estimated using this method. The effect of the estimated values will be limited in the overall distribution.

In general, the problem with this solution is that it uses the simplest possible model to estimate a value; 'the mean'. This means that the traffic intensity (IC-ratio) will not be calculated using this method because less than 90% of the data is known. As a result traffic intensity (using the IC-ratio) is dropped as a possible safety factor from the analysis.

Table 3.1 provides an overview of the relative amount of known data per safety factor and the corresponding used estimation methods. The known data is provided on two scale levels;

first on the total network and secondly on the network on route-level (see Appendix B). The scenarios are based on the known data of the network on route-level that comprises the area of the shortest and chosen routes.

Tab. 3.1.: Methods used for estimating missing data

Safety factor	Type	Known data		Estimation Method
		Network total	Network routes	
Relative amount of crimes (CBS)	continuous	70.7%	99.9%	Average
Road levels (Fietzersbond)	categorical	86.9%	82.4%	Classification Learner (ML)
Road type (Fietzersbond)	categorical	73.2%	75.8%	Classification Learner (ML)
Degree of illumination (Fietzersbond)	ordinal	72.4%	63.6%	Classification Learner (ML)
Road surface (Fietzersbond)	categorical	61.4%	59.4%	Classification Learner (ML)
Road quality (Fietzersbond)	ordinal	61.2%	59.3%	Classification Learner (ML)
Hindrances (Fietzersbond)	ordinal	61.4%	59.3%	Classification Learner (ML)
Speed (Fietzersbond + Traffic Model)	ordinal	59.2%	57.2%	Classification Learner (ML)
Subjective safety (Lemon)	ordinal	2.9%	36.5%	Classification Learner (ML)
Intensity/Capacity Ratio (Traffic Model)	continuous	21.8%	27.8%	None

Creation of delta dataset

To investigate whether the amount of deviation can be explained by the difference in safety factors, the chosen and shortest routes need to be compared. To do this, the final 'delta' (Δ) dataset is created that can be used for the analysis. To measure the deviation both the absolute deviation and relative deviation are calculated in the following way.

Absolute deviation (meters): $\Delta length = length_{chosen} - length_{shortest}$.

Relative deviation (ratio): $Ratio = \frac{length_{chosen}}{length_{shortest}}$

In relative terms, a value of 1.00 indicates an identical length of the shortest compared to the chosen route. Values bigger than 1.00 indicate that the chosen routes are longer than the shortest routes and of course values below 1.00 do not exist.

Finally, the difference in safety factors is calculated as well. For all safety factors this is done in the following manner:

$$\text{Difference in safety factors: } \Delta x = x_{\text{chosen}} - x_{\text{shortest}}.$$

As mentioned before, four different aggregation methods are used to calculate the safety factors on route-level. Therefore, the interpretation of the Δ safety factor also differs. The following list provides an example of the interpretation of the Δ safety factor for each of the four different aggregation methods that have been used.

- *The Δ absolute number of accidents is +5*
This means that on the chosen route more accidents (5) have taken place than on the shortest route alternative.
- *The Δ relative number of traffic lights is -5*
This means that on the chosen route fewer traffic lights (5) per kilometer exist than on the shortest route alternative.
- *The Δ average lemon score is +2*
This means that on the chosen route the average lemon score is 2 points higher than on the shortest route alternative.
- *The Δ asphalt/concrete roads is +30*
This means that on the chosen route the coverage of asphalt/concrete roads is 30% higher than on the shortest route alternative.

3.2.4 Statistical analysis

Finally, the influence of safety on route choice behavior is investigated using a simple single linear regression method. This is a well-established method to investigate the relationship between variables (Field, 2009). The main goal in this research for using regression analysis is to determine whether there is a significant relationship between the dependent and a specific independent variable, and if so, how well this relationship is explained in a linear model. It is important to mention that all models are created using a single explanatory variable.

The dependent and independent variables in this model are:

Dependent: the amount of deviation from the shortest route (in absolute or relative terms).

Independent: the difference between a safety factor on the chosen versus the shortest route.

The independent variables, throughout this research referred to as Δ safety factors, are (depending on the safety factor) the difference between:

- the absolute number of X on the chosen route versus the shortest route alternative, or:
- the density of X per kilometer on the chosen route versus the shortest route alternative, or:
- the distance weighted average of X on the chosen route versus the shortest route alternative, or:
- the coverage of X (in %) on the chosen route versus the shortest route alternative

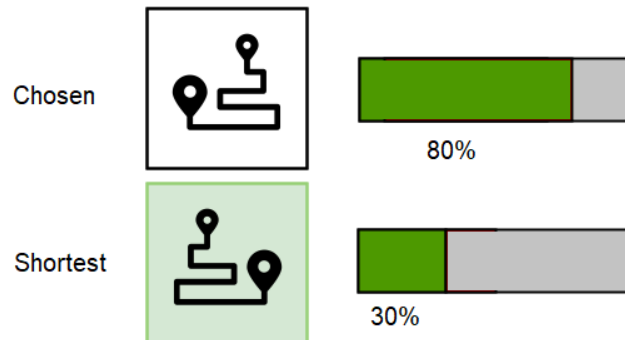


Fig. 3.5.: Difference in safety factors

Figure 3.5 shows an example of the difference in coverage of bicycle paths. The coverage of bicycle path is 80% on the chosen route versus 30% on the shortest route. The difference (Δ) in the coverage of bicycle paths would be +50%. This means that on the chosen route the coverage of bicycle paths is 50% higher than on the shortest route alternative. Comparing travel distances as well shows that this person has traveled 2 kilometers further (or twice as far in relative terms) than the shortest route alternative.

To determine whether this amount of deviation can be statistically explained by the difference in the coverage of bicycle paths, the single linear regression method creates a linear model for all routes in the study area using the amount of deviation as the dependent variable and the difference in the coverage of bicycle paths as the independent variable.

The method of single linear regression -aside from determining whether there is a significant relationship between the two variables- also allows checking how big of an effect the independent variable has on the dependent variable. Furthermore, the strength and the goodness of fit of the model can be investigated as well. The measures that are used for this are: Pearson's correlation coefficient (r) and the R^2 measure (Field, 2009).

The next chapter will describe the datasets that have been used to measure safety and route choice behavior.

4.1 Cycling GPS data

4.1.1 Fietstelweek project

The GPS data that is used to investigate cycling behavior is the open-source Fietstelweek (2016) data (Bikeprint, 2016). The Fietstelweek is a Dutch cycling project that is used to gain and provide insight into movements of cyclists. In order to track cyclists, it uses the Fietstel-App that uses the built-in GPS from the cyclist's smartphone during a one-week period. The Fietstelweek is arguably the biggest project ever done regarding bicycle movements in the Netherlands. The second edition of the Fietstelweek (2016) had 29,000 participants and because of the weather conditions resulted in more data than in the first edition. In general, the second edition of the Fietstelweek is quite representative based on the information gathered from people that have signed up and the main cycling motive is commuting to work (see Appendix C.1). More details about the representativity of the sample in this research is provided later on in this section.

Finally, participating in the Fietstelweek is entirely voluntary and possible to enter for anonymous users; the people that register with their e-mail address and answer some personal questions can win bicycle-related prizes. Everyone with a smartphone can participate and install the app. The app itself does not require active input from its users which is one of the reasons such a wide array of people are reached.

Fietstelweek App

The Fietstelweek App is a smartphone application (available for Android and iOS platforms) that uses the geo-location provided by the smartphone of the user. In the 2016 edition of the Fietstelweek, the movements of people are automatically recorded for a period of one week. To register movements the application does not have to be actively turned on.

Other important things to mention are:

- The geolocation of a person is based on GPS, Wi-Fi and GSM-cells. Also, the tracking data has a time-stamp, and it uses the accelerometer samples from the smartphone to determine the acceleration.
- Because the movements of people are considered as personal data, the Personal Data Protection Act applies to all the data collected during the Fietstelweek, and the data is anonymized before making this data public.

- All trips are anonymized by cutting off 200 meters from the start and end-location of a trip. In addition, a time shift between 0-15 minutes are applied randomly.
- The user can enter the Fietstelweek anonymously; meaning that it is not necessary to provide personal information to participate in the Fietstelweek

Sample representativity

To increase the variety of people participating in the Fietstelweek project, much effort went into targeting the audience. During the first edition in 2015 it showed that the sample was quite representative of the whole country with some exceptions (Edwards, 2017) and also the second edition in 2016 seemed quite representative (Keypoint, 2016). However, this does not necessarily guarantee that the sample is also representative on other scale levels as well.

During the Fietstelweek in 2016 more than 5,000 trips were taken in the study area. Unfortunately, however, it is unknown how many people have cycled during the Fietstelweek because it is possible to enter the Fietstelweek anonymously. In total 129 people signed up and provided their personal information using the location in the municipality of Tilburg as their home location. It is important to note however that even though people signed up using the municipality of Tilburg as their home location that these people might also cycle outside of the municipal boundaries and that other people (not assigned to the municipality of Tilburg) will possibly cycle in this area as well. Since it is unknown how many people in total have entered the Fietstelweek, it is impossible to know how the sample of $N=129$ relates to the total amount of participants. Furthermore, the personal data that is known is not assigned to the trips individually to ensure the privacy of the participants.

From the 129 people in the sample, 52% is male and 48% female. In the municipality these numbers 50% and 50% respectively according to CBS (2016b). Therefore it seems that this sample is highly representative regarding gender. For age, this distribution is provided in table 4.1.

Tab. 4.1.: Age distribution

Age	Sample (N=129)	Municipality (CBS)
<25 years	27%	30%
25-44 years	33%	27%
45-64 years	37%	27%
≥ 65 years	3%	16%

This means that the sample is also quite representative regarding age. The age group 45-64 years, however, is slightly over-represented and the age group of ≥ 65 years is under-represented in the sample.

Also the education level can be checked on both levels. As can be seen in figure 4.2 people with a high education level are highly over-represented in the sample and the low education level is highly under-represented in the sample.

Tab. 4.2.: Distribution of education level

Education level	Sample (N=129)	Municipality (CBS)
Low	9%	36%
Middle	29%	40%
High	62%	24%

Furthermore, a wide array of motives for cycling are found during the Fietstelweek 2016. On a national scale, the main cycling motive is commuting to work, but this is not the case for the study area. The reasons for cycling in the study area are very diverse, but can be characterized as cycling for everyday use. Using this data makes it possible to investigate everyday cycling behavior in the Netherlands. This is often not possible because most available cycling data is collected for sportive or recreational purposes and therefore show different behavior. More information about the cycling purposes in the study area is provided in Appendix C.2.

Even though much effort has been made to reduce the impact of bias, the cycling data from the Fietstelweek is not free from prejudice, and therefore the following should be taken into account.

People that have participated in the Fietstelweek project:

- need to own a smartphone.
People without smartphones are excluded from the project.
- need to (know how to) download the Fietstel App.
People that have not downloaded or do not know how to download the application on their smartphones are excluded from the project.
- should be aware of the project by (targeted) advertisement and news items.
This means that there is no random sample used to collect the data.
- can enter the Fietstelweek anonymously.
This means that there is no real way to check if the people that have entered the Fietstelweek actually are somewhat representative of the whole population.

Hence, it might be safer not to generalize the results of the data analysis.

4.1.2 Fietstelweek (2016) data

The Fietstelweek data exists of one shapefile containing the geometry of the cycling network and one CSV file containing all the route information. Originally, the open source Fietstelweek

data uses the Open Street Map network, but for this research a different cycling network is required. To match the cyclists' movements to the correct cycling network, a map match process is carried out by the NHTV because they are a partner in the Fietstelweek that possess the original raw GPS data (which is not openly available due to privacy reasons).

This map match process matches the original GPS points to both the nearest and most likely road segment that a person would have actually traveled over. It also takes into account if traveling the segments create a connected route over the cycling network, so that the route makes sense in reality. Furthermore, in case of GPS inaccuracy and multiple roads can be matched, it takes into account the types of roads by assigning weights to different kinds of roads. For example, the weight for a car highway is higher than a bicycle path. (Van der Coevering et al., 2014)

After the mapmatch process, the shapefile contains the geometry of the correct cycling network and an ID for the road segment a cyclist has traveled. The corresponding CSV file contains all the route information from the Fietstelweek participants in the Netherlands. This dataset does not contain any geometry but can be linked to the shapefile's geometry because they have the ID for the road segments in common. Furthermore, the CSV file also contains the routeid for each unique route, a sequence code that shows the order in which people have traveled specific road segments for every route, and finally the hour (0-23) in which a person has travelled.

Travelled roads of the cycling network in the study area (Fietstelweek 2016)

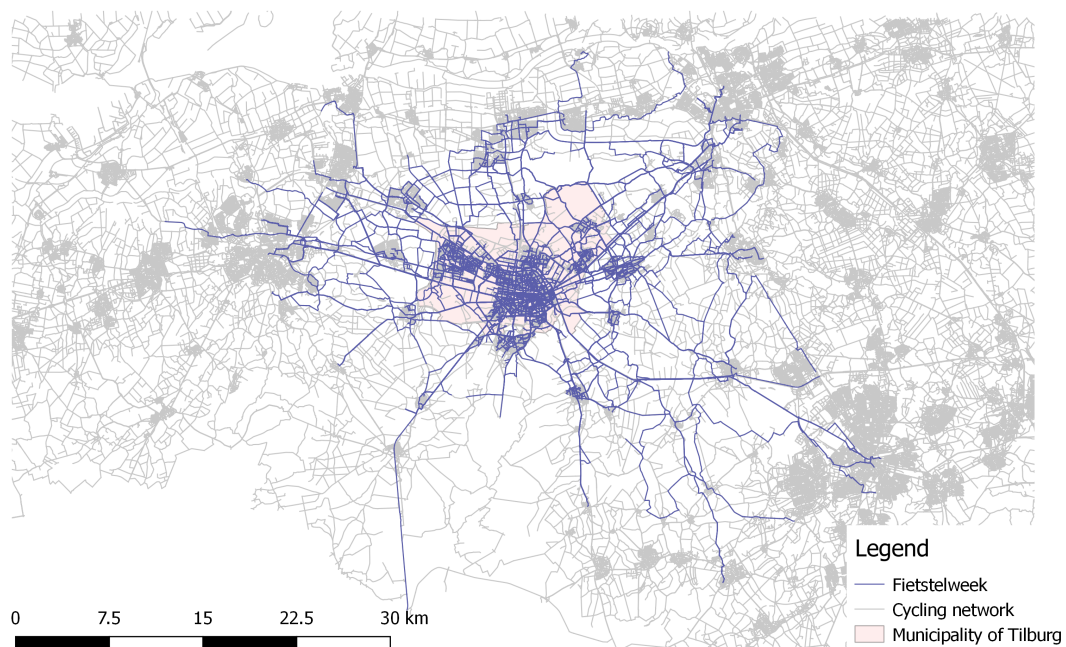


Fig. 4.1.: Travelled roads during Fietstelweek 2016

Figure 4.1 provides a spatial overview of the used and unused parts of the cycling network in the study area. The blue lines in the map represent the roads that have been travelled by at least one person during the Fietstelweek, and the grey lines represent the unused roads.

4.2 Cycling network

The cycling network that is used is created by the Fietzersbond (*English: Cycling Union*). At the moment, this is the most accurate cycling network that exists in the Netherlands. It goes well beyond the main cycling network, and many (safety) attributes of roads are provided. Because in this research the Fietstelweek data from 2016 has been used, it is important in order to map cyclists' routes accurately, to also use the cycling network from 2016 (Fietzersbond, 2016). The reason why the cycling network is so accurate is that it is based on VGI (volunteered geographic information). According to Goodchild and Li (2012) VGI is:

"... a type of crowd-sourcing in which members of the general public create and contribute geo-referenced facts about the Earth's surface and near-surface."

These volunteers help create and update the network from all over the Netherlands, which explains why it is such a rich and up-to-date data source. It is important however to mention that the Fietzersbond data is not open data, but that the metadata is available (Fietzersbond, 2018).

4.3 Safety data sets

In order to measure the impact of safety, various types of datasets are used that are related to safety. In the theoretical chapter, the relevant safety factors have already been identified and in this part, the corresponding used datasets to measure these safety factors are provided in the overview of figure 4.2. More detailed information about the datasets is provided in Appendix D.

Accidents

The data of the accident locations are provided in point features. All accidents that are used are bicycle accidents only for the total year of 2011 until and including 2015 (BRON, 2017). The reason for this is because a person can perceive certain locations as dangerous for quite some time after the accident has happened. To use a relevant time frame within the available data sets that are provided, a time frame of 5 years is chosen. From these locations, **all** available locations are used, because not only cycling accidents that result in

casualties or (seriously) harmed victims are relevant. Also near-misses, one-sided accidents and accidents that result in merely material damage should be taken into account in case there is existing data about it. In this case, as there is almost never, there is no data available regarding one-sided accidents and near-misses. The dataset is as complete as possible, but due to under-registration, the numbers do not represent the reality. It does, however, provide useful insight into dangerous locations.

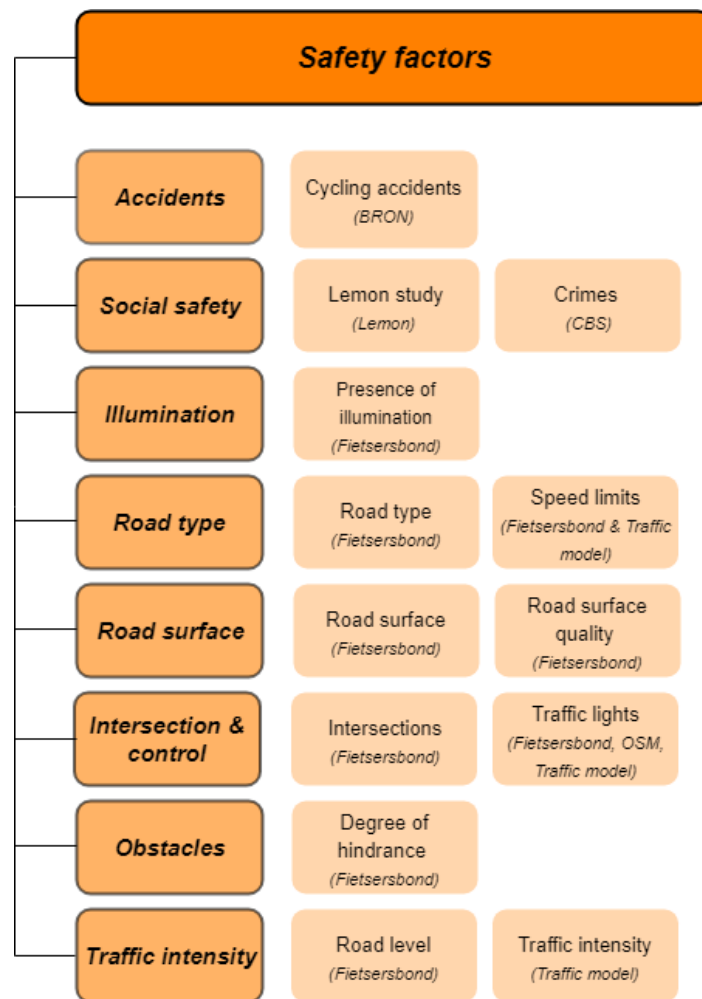


Fig. 4.2.: Overview of safety factors and corresponding datasets

Safety attributes Fietzersbond

Multiple safety attributes are used that are collected by the Fietzersbond initiative. Because the Fietzersbond is a VGI initiative, not all attributes are assigned to all road segments as is mentioned in the previous chapter. The percentage of data that is missing per safety attribute, is provided in table 4.3 for a short overview. Some important things to take into account are the following:

The Fietzersbond describes hindrances as:

"Hindrances are delays and/or dangerous situations as a result of the physical presence of 'other traffic'" (Fietzersbond, 2018).

In this case 'other traffic' can be driving or (incorrectly) parked cars, scooters, other cyclists, pedestrians or any combination of these road users. Also, the degree of hindrances is provided only for rush hour circumstances.

Furthermore, the road speed limits are made up of two different datasets. On its own, the Fietzersbond only provides 53% of the network¹ with known values regarding maximum speeds.

For planning purposes, the municipality of Tilburg has created a traffic model (in collaboration with a company called *Goudappel Coffeng*) that covers the study area (of Tilburg, 2017). This traffic model contains many types of traffic-related information such as data about maximum speeds. Therefore, this traffic model of Tilburg is used to look for additional information regarding maximum speed for road segments that have an unknown speed in the Fietzersbond dataset. This addition slightly increases the coverage to 59%. It is important to mention that the data from the traffic model is not open data.

Tab. 4.3.: Fietzersbond safety attributes

Fietzersbond attribute	Missing data
Intersections	0 %
Road level	13 %
Degree of illumination	27 %
Type of road	27 %
Type of road surface	39 %
Road surface quality	39 %
Hindrances	39 %
Speed limit	47 %

Social safety

Social safety is measured in two ways: 1) the objective social safety and 2) the subjective social safety.

CBS (2015) Objective Safety

The data from CBS is provided per neighborhood (CBS, 2016a). The data that is used is the relative *number of total crimes* (including fraud, destruction of property and violence) and the relative *number of violent and sexually related crimes*. Safety is often expressed per x number of inhabitants to correct for certain factors. Therefore, the absolute number of

¹For an overview of the extent of the cycling network that has been used during this research, see Appendix B

crimes that is provided by CBS is normalized by dividing it by the number of inhabitants per neighborhood.

Lemon (2015) Subjective Safety

Data from the Lemon project is used to measure the subjective safety (Lemon-Onderzoek, 2015). In this project, research has been carried out that investigates the subjective feeling of (un)safety in one's neighborhood. This study is done for the Municipality of Tilburg, and therefore no data outside of the municipal boundaries are known. Also, not all areas in Tilburg are provided with (un)safety ratings. The used data is the rating for 1. *feeling of unsafety in the neighborhood during the day* and 2. *feeling of unsafety in the neighborhood during the night*.

In this case, a high rating means that a person has no/little feeling of unsafety and a low rating means that a person has a high feeling of unsafety in one's neighborhood.

Traffic lights

The location of traffic lights is created by combining three different datasets. First of all, the Fietsersbond dataset has an attribute regarding the type of intersections. The type of intersection that is used to represent the presence of traffic lights is the value *Intersection with traffic regulation installation*. Traffic regulation installations are associated with the presence of traffic lights. Also, the traffic lights that are present on Open Streetmap (OSM, 2017) and in the Traffic model of Tilburg are used, and the locations are matched to nearby road segments.

Traffic intensity

The Tilburg traffic model contains the traffic load and the capacity of roads in line features. The ratio between the two can be used to express traffic intensity (also called the IC-rate). The IC-rate is calculated per hour.

The next chapter will describe how the data has been prepared and will reflect upon the quality of the data mentioned in this chapter.

5.1 Data preparation network

5.1.1 Topology of the cycling network

Before the data can be analyzed, first the data is prepared. First of all the cycling network is set up. This is an important step in order to create the shortest routes for all taken trips later on in the process. One of the most critical aspects of the network is that it is well connected, and general regulations are applied in order for the network to meet the relevant traffic regulations and to create basic rules for cycling behavior in the system (Cheng and Chang, 2001). The second rule for building a network are the specific concerns of the required network.

The general regulations for this network correspond to:

- Connecting relevant line segments that correspond to a road
- Only connecting road segments that are actually connected in reality. For instance; in case of multi-level roads (e.g., tunnels or bridges) the correct nodes need to be connected.
- Costs based on length should be assigned to all edges

The original dataset of the Fietzersbond is topologically incorrect, and therefore the geometry is to be altered to connect the roads that are connected in reality. To correct the cycling network, a Python script (*RepairNetwork.py*) is used that uses the Shapely and Fiona (geospatial programming) tools to manipulate the geometry of the cycling network for road segments to connect. The original dataset of the Fietzersbond does, however, contain an "id" column for every road segment and also the connectivity between these road segments is documented.

After the topology is corrected, one extra column is added to the dataset; a column representing the newly calculated road's length in meters. Travel directions (and accessibility of roads for cyclists) are ignored, because as the mapped (chosen) routes suggest cyclists don't always stick to these rules. However, when calculating the shortest path, one has to make sure that despite not using any restrictions on travel directions, cycling behavior will be somewhat realistic. To make sure no paths are taken that are unrealistic, the 'not accessible to cyclists' roads that are characterized as *highway* are removed from the dataset. The reason for this is because cycling on highways is considered very dangerous and illegal in the Netherlands.

Even though Dutch cyclists might not always stick to traffic rules, it seems unlikely that cyclists would in fact cycle on these roads.

Next, after the data is prepared, the network is built using the Network Analyst toolbox in ArcGIS. It uses the corrected topology to create nodes and edges. Also, during this process the *costs* are assigned to the road's length in meters. Finally, the connectivity is tested and checked to make sure the network is set up correctly.

Data quality

The Fietzersbond dataset is a dataset provided by the efforts of VGI, and all the volunteers can edit the network dataset using a cartographic editor. Because most of the volunteers are not professionally trained as cartographers, the quality of digitizing the roads can vary among the volunteers. This, however, is not necessarily an unsolvable problem since the connectivity of the roads is provided. Because the network is built by volunteers, it is possible that not all roads that exist in reality are digitized. Yet, the roads that are digitized have a large number of attributes relevant for measuring safety and are considered valuable.

5.1.2 Network enrichment

Point features

The data for accidents and traffic lights are provided as point features. These point features are spatially joined to the relevant line segments of the network using an Intersection method in ArcGIS.

Cycling accidents

To match the *cycling accidents* to the roads of the cycling network, a buffer of 10m is created around the point location of the accidents to add more flexibility due to data (GPS and cycling network) inaccuracy. The second reason is that roads that are within 10meters (or the entire intersection) of the accidents can also be considered as dangerous and not only the road segment itself. Especially in the network of the Fietzersbond road segments can be created in various lengths even though they belong to the same road in reality. Therefore it is necessary to use a buffer around these accident locations in order to select multiple road segments that might be relevant.

In the example of figure 5.1 the location of the accident or the location of the roads does not place the location of the accident on the road or intersection. By using the buffer, the relevant road segments can still be selected. However, this method also has some negative outcomes, meaning that it also selects roads that are probably not involved or relevant in the case of this

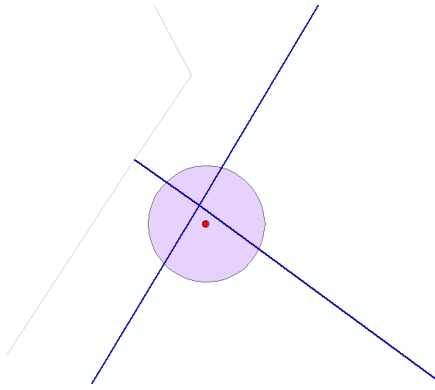


Fig. 5.1.: Inaccuracy of positioning

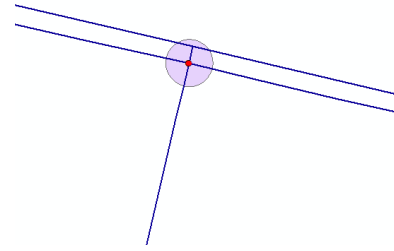


Fig. 5.2.: Inaccuracy of method

accident. Figure 5.2 shows that by using the buffer also other roads that fall within 10 meters are selected. In this case, also a road parallel to the intersection (a service road) is selected.

After careful testing of different methods, however, it showed that (even though there are some flaws) this method scores best in selecting relevant roads. The commonly used Near method does not allow for this type of flexibility that is needed.

Traffic lights

For traffic lights, a similar operation is used.

Traffic lights, however, are provided in three different datasets. Two of these datasets are point data, and they follow the same operation as the accident data. One thing however is different; the addition of the line features of the Fietzersbond dataset. The values of the Fietzersbond attribute *type of intersection* that mention the type 'intersection with Traffic Regulation Installation' are used as well for the presence of traffic lights.

The quality of all three datasets is checked by checking a sample using satellite and Streetview images for the presence of traffic lights. The conclusion is that all three datasets are correct and incomplete on their own.

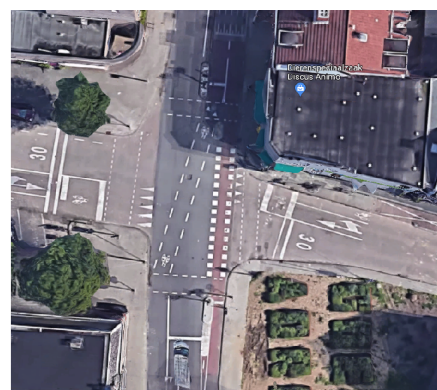
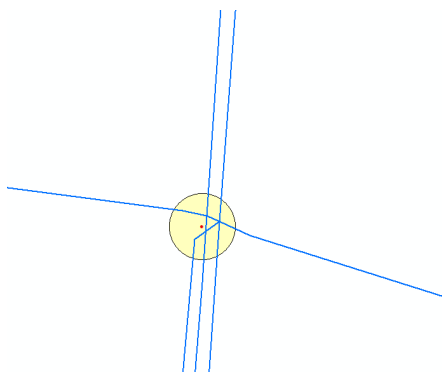


Fig. 5.3.: Validation presence of traffic lights 1

During the validation of this method using satellite and Streetview images, it showed that (despite using several datasets) not all traffic lights are in the dataset. Often more than one traffic light exists in reality at an intersection and yet only one is actually entered. Using the buffer made sure that more relevant roads were taken into account, but sometimes the buffer method felt short as well. Also, again in a small number of cases roads that in reality do not have traffic lights are listed as a road that has a traffic light. For example, figure 5.3 shows that all roads actually have a traffic light, but in reality not all parts have a traffic light. In this case, only before crossing the road one comes across a traffic light and not again once they have crossed the intersection. This, however, is not a problem because these roads in reality are associated with the presence of traffic lights as well. Figure 5.4 however is an example of both over- and underestimation. The grey lines represent the roads that are not associated with the presence of a traffic light. This in fact is not correct on in this case; there should be a traffic light on one of these roads according to the satellite and streetview images, but the point locations do not reach this road. On the other hand, one of the roads is marked as a road with a traffic light, that in reality is a parallel road of one of the roads nearby that does have a traffic light.

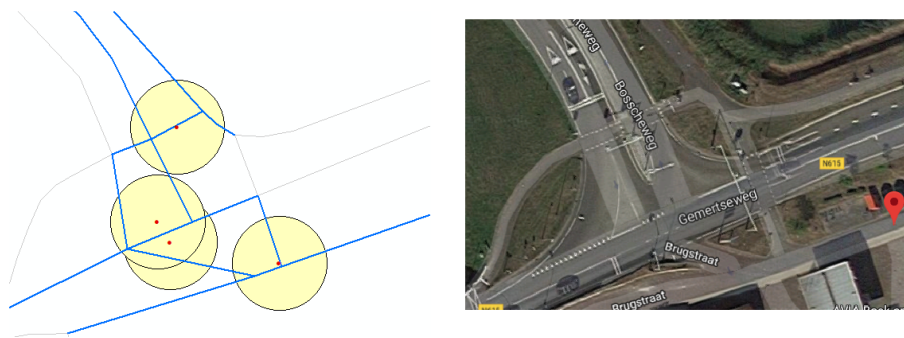


Fig. 5.4.: Validation presence of traffic lights 2

After careful testing of different methods, however, it showed that (even though there are some flaws) this method scores best in selecting relevant roads.

Polygon features

Lemon: Subjective safety ratings

The data from the Lemon project, ratings of subjective social safety, are related to neighborhoods that are provided in polygon features. To match these ratings to the road segments that are located in the neighborhood, the data is spatially joined to the cycling network using an Intersection method in ArcGIS. To make sure that every road has only one neighborhood rating, the roads need to be cut by the neighborhood boundaries and its length recalculated. Then the neighborhood ratings are assigned from the neighborhood that covers the largest part of the road.

Crime data

For objective social safety based on the CBS crime data a similar operation is used.

The data, in this case, however is available for **all** neighborhoods in the Netherlands and it uses the CBS neighborhood format. Therefore other neighborhood boundaries exist compared to the Lemon dataset.

To normalize the absolute crime data, the number of inhabitants (per 1000) is used to express the amount of crimes per 1000 inhabitants (per neighborhood).

Line features

Fietsersbond data

Some of the features of the Fietsersbond network need to be reclassified, such as the *road type*. The reclassification is provided in table 5.1. The most important reclassification is the creation of a new class 'mixed traffic'; this means that these roads are used by different types of road users. The most important thing about these mixed traffic places is that the cyclist does not have a separate or suggested place on the road here.

Tab. 5.1.: Reclassification road types

Original	Reclassified
[null]	<i>[null]</i>
onbekend	
ventweg	
voetgangersdoorsteekje	<i>mixed traffic</i>
normale weg	
voetgangersgebied	
veerpont	
fietsstraat	<i>bicycle boulevard</i>
weg met fiets(suggestie)strook	<i>bicycle suggestion lane</i>
solitair bromfietspad	<i>scooter path</i>
solitair fietspad	<i>bicycle path</i>
bromfietspad (langs weg)	<i>scooter lane</i>
fietspad (langs weg)	<i>bicycle lane</i>

To determine the traffic intensity, the traffic model of Tilburg is used. As figure 5.5 shows, the traffic model is less detailed than the Fietsersbond network. This figure shows that traffic model is coarser than the Fietsersbond network, because the coverage of roads in this dataset is limited to the main road network. The left map shows an overview of differences in coverage for approximately the whole study area. The right map is zoomed in a little further on one of the neighborhoods and shows this in more detail. The reason why the traffic model is limited to the main roads mostly, is because most traffic models focus on motorized traffic and therefore these models do not cover all roads. This dataset can, however, still be useful because cyclists also cycle on the roads that are in the traffic model.

Especially since for some roads data is missing in the Fietsersbond dataset, this can be

complemented by data from the traffic model. Also, the traffic model provides information that the Fietsersbond data does not provide in terms of traffic intensity.

Coverage of roads in Traffic Model vs. Fietsersbond network

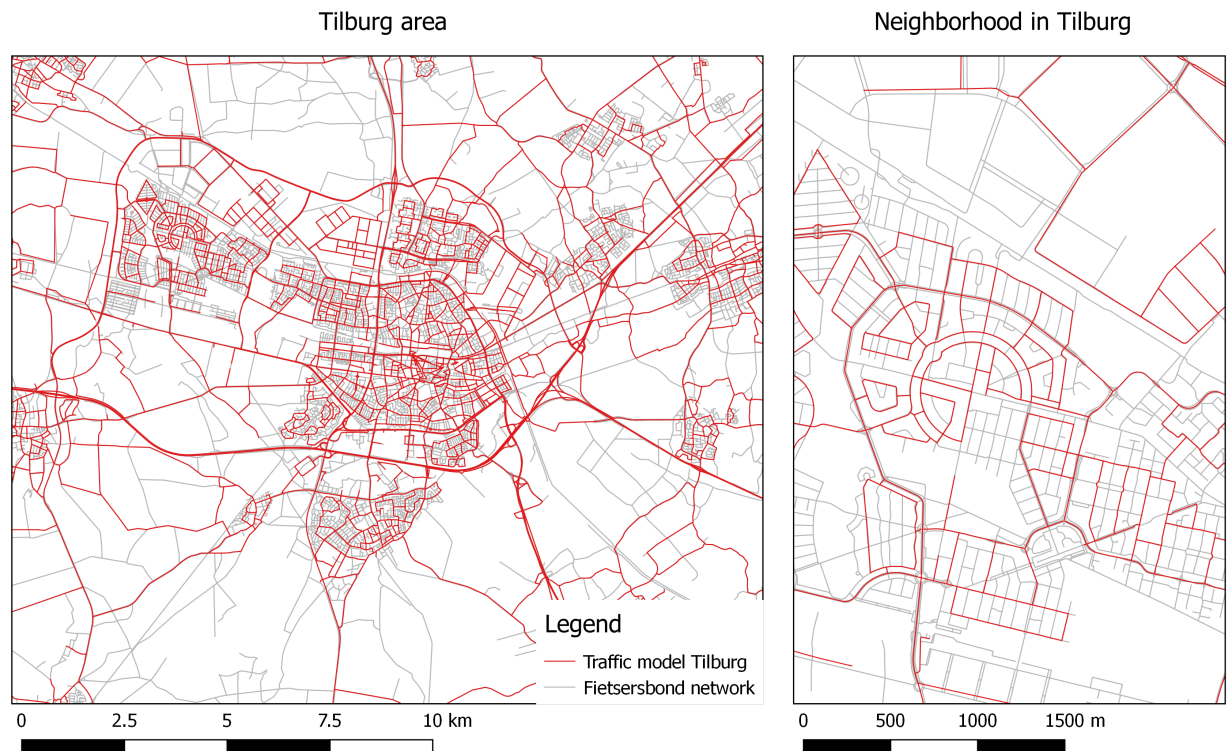


Fig. 5.5.: Coverage of roads included in Traffic Model and Fietsersbond network on two scale levels

The problem of joining both datasets is that no unique key can be used to match the roads in both datasets. Also, because the coverage of both datasets is different, the only way of joining both datasets is by using spatial location. The roads in both datasets are digitized in a slightly different manner. Therefore sometimes a road segment is longer in dataset A than in dataset B. Also sometimes road segments cross, or they are located parallel to the road in the other dataset. To create more flexibility, a buffer of 5 meters around the roads of the traffic model is created. Then, by spatially joining and using the Within method in ArcGIS, the data from the traffic model is matched to the roads of the network that fall within the buffer.

Figure 5.6 shows how the method works. In the figure, the roads in the Fietsersbond network that do not find a match to the roads from the traffic model dataset are colored grey, the roads that do find a match in the traffic model dataset are colored yellow. Additionally, the red lines represent the (slightly differently) digitized roads of the traffic model around which a buffer of 5m (green) is drawn.

Because every road in the Fietsersbond dataset can only be matched to one road in the traffic model dataset, only the roads that have a unique match are joined. By using this method not all roads can be matched (even though they are very likely the same road) because by enlarging the buffer, also parallel and other roads are matched that should not be matched. After careful testing and validating a number of different methods, it showed that this method scores best in selecting relevant roads. The validation of the join is difficult because there is no official way of knowing which road is which due to the lack of an identifier. However, by visually random testing and inspecting the attributes of both datasets it showed that in all tested sample cases the right roads were matched.



Fig. 5.6.: Buffer method to match roads

5.2 Data preparation routes

5.2.1 Creating chosen routes

In this study, the area of interest is the municipality of Tilburg, which means that only the routes that go through or lie entirely within this area are extracted.

To select the relevant routes, first, the roads of the cycling network are selected using the municipal boundaries of Tilburg by means of a Clip operation in ArcGIS. Next a series of SQL queries are used (*SelectionRoutesTilburg.sql*) to extract all the routes (and route information) that are relevant for the study area. This results in 5,063 unique cycling routes that are either fully or partly located in the study area. Then, using the ID of the roads of the enriched network, the geometry and (safety) attributes of the traveled roads are added to the routes. However, 14 routes seem to travel over road segments that are 1. not accessible to cyclists and 2. are characterized as a highway. It seems unlikely that people would have traveled these road segments in reality. Therefore these routes are removed from the dataset. The final total number of routes that are left is: 5,049 routes.

5.2.2 Creating shortest routes

To create the shortest route, first, the start and end location of the chosen routes are determined. This is done by using the 'sequence' information from the Fietstelweek route information. The sequence provides the order in which someone has traveled the different road segments on their route. The start and end location are retrieved by respectively selecting the minimum and maximum sequence value per route. Then, point data is generated for the start and end location for every route.

Using the start and end locations, the shortest path function can be executed for each route. To do this efficiently for all routes at once, a Python script (*Shortest_routes.py*) is written that first creates a route layer where the relevant network, impedance attributes (in this case *length*) and restrictions (*travel direction and accessibility*) are set. Next, the start and end locations for each route are added to this route layer. Then, the shortest routes are created for all start and end locations for each route id, and the result is written to an output file. Finally, the shortest routes are spatially joined to the enriched network using the Within method in ArcGIS to add all attribute information. Both the chosen routes and the shortest alternatives can be found on Github (*chosenroutes_lines.zip* and *shortestroutes_lines.zip*).

5.3 Separation routes in timeslots

Some of the safety attributes are only relevant in certain day conditions, such as the presence of illumination. Also, people can show different behavior in the dark compared to cycling in the daylight, but also people might behave differently during rush hour. Since the roads are more crowded during morning rush hour compared to evening rush hour, a distinction is necessary. During this time of year, four combinations are possible (see table 5.2). More background information and considerations can be found in Appendix E.

Tab. 5.2.: Classification of routes

Type	Conditions	Time	Routes (total: 5049)		Routes (lemon: 2825)	
			<i>abs.</i>	<i>rel.</i>	<i>abs.</i>	<i>rel.</i>
1	Dark & non-rush hour	00:00 - 06:59 20:00 - 23:59	464	9.2%	289	10.2%
2	Light & morning rush hour	07:00 - 08:59	1220	24.2%	609	21.6%
3	Light & evening rush hour	16:00 - 17:59	944	18.7%	507	17.9%
4	Light & non-rush hour	09:00 - 15:59 18:00 - 19:59	2421	47.9%	1420	50.3%

In this section, the results of the data analysis are provided. The first part will statistically investigate the amount of deviation and safety factors individually and in the second part, the influence of safety on route choice behavior is investigated using simple single linear regression models.

6.1 General route characteristics

Before we can investigate what factors influence the deviation of the shortest routes, first the route lengths and deviations itself are investigated to see whether people actually traveled further than the shortest alternatives at all.

6.1.1 Differences in trip length

First of all the routes' lengths are compared to investigate if cyclists actually travelled longer routes than the shortest possibilities. When comparing all descriptive statistics for both type of routes (see table 6.1), it shows at first glance that the shortest routes actually are shorter than the chosen routes in general.

This difference, however, can be due to chance and therefore the difference needs to be proven statistically. To investigate whether there is a significant difference between the average deviation of the two types of routes a paired samples T-test is used (Field, 2009).

Therefore, the following null hypothesis is tested:

H_0 = The average length of the shortest and chosen routes are not different from each other.

The corresponding alternative hypothesis is:

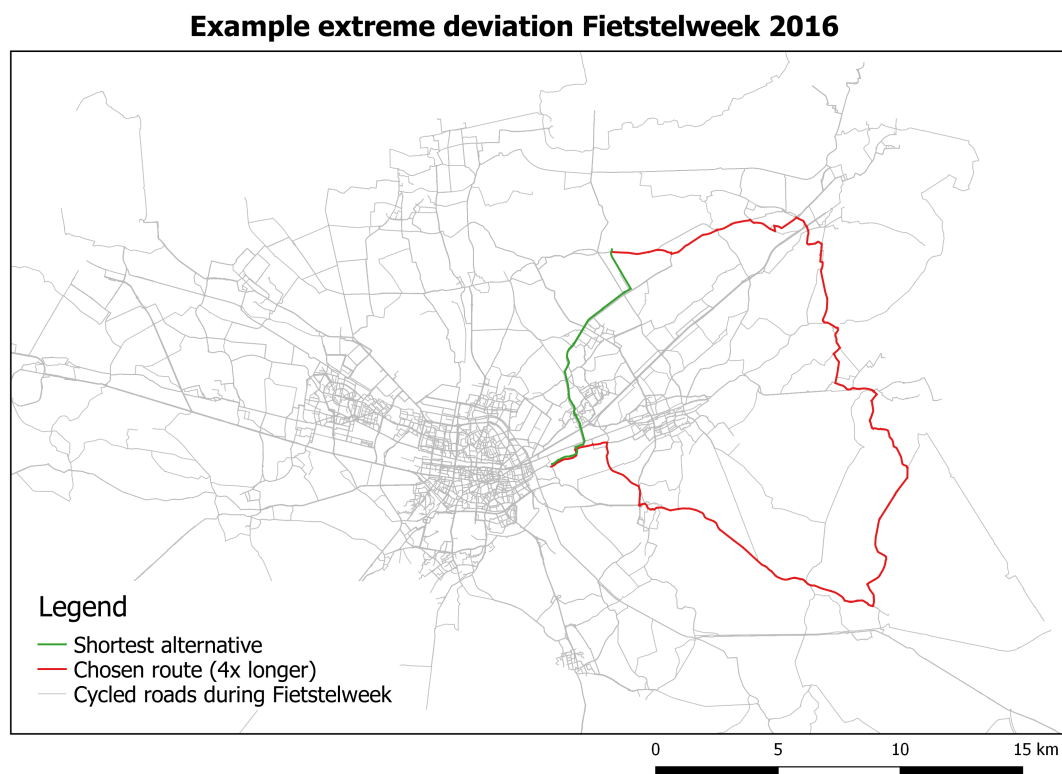
H_A = The average length of the shortest and chosen routes are different from each other.

A paired samples T-test shows that on average, the length of chosen routes ($M = 4,564$) are significantly larger than the shortest routes ($M = 4,226$), $t(5048) = -18.5$, $p < 0.01$, $r = 0.25$. The effect size can be classified as medium.

In general, the deviations from the shortest routes are not large. The first 50% of the routes deviate less than 100meters from the shortest route. Due to a number of outliers and variance in the data, the average deviation is 338 meters or relative terms; the average deviation is 1.07 times as far as the shortest route. Even though many routes are concentrated around small deviations relative to the shortest route, the value of the standard deviation suggests that there is a large variance in the absolute deviation of all routes. Figure 6.1 shows the most extreme deviation from the shortest route alternative which is more than four times longer than the shortest route.

Tab. 6.1.: Descriptive statistics trip lengths

	length chosen (m)	length shortest (m)	abs. deviation (m) (chosen - shortest)	rel. deviation (ratio) (chosen/shortest)
<i>minimum</i>	500	317	0	1.00
<i>maximum</i>	52,783	50,736	35,585	4.14
<i>average</i>	4,564	4,226	338	1.07
<i>st. deviation</i>	4,886	4,278	1,300	0.16
<i>Q1</i>	1,296	1,228	15.5	1.01
<i>Q2</i>	2,940	2,754	96.0	1.03
<i>Q3</i>	6,347	5,965	293.8	1.08

**Fig. 6.1.:** Extreme deviate from the shortest route

When comparing the statistics of the four different timeslots/groups (see table in Appendix F.1), the average deviation among the groups does not seem to differ that much. In terms of absolute deviation (in meters), the second group has the lowest average and standard deviation. This suggests that during the morning rush hour people seem to deviate less from the shortest route than during other timeslots. To rule out if this is due to chance, the Welch test gives more insight into this matter. The Welch test is used as an alternative to the regular ANOVA test when the assumptions of the regular ANOVA are violated (Field, 2009).

For the Welch test the $H_0 = \text{There is no difference in means between the four groups.}$

The $H_A = \text{There is a difference in means between the four groups.}$

The Welch test shows that on average, the absolute deviation between the groups is significantly different. $F = 5.9, p < 0.01$.

This means that it is possible to reject the null hypothesis with a 99% certainty. A Games Howell post hoc test is done to find out which group means significantly differ (Field, 2009). This test showed that the groups that differed significantly are the 2nd (during daylight and morning rush hour, $M = 246.9$) and 4th (during daylight and non rush hour, $M = 395.4$) with a mean difference of -148.5 meters and a corresponding p value < 0.01 . A negative mean difference suggests that on average the absolute deviation of the 2nd group is lower than the absolute deviation of the 4th group.

This is to be expected since trips in group 2 take place during the morning rush hour and people often cycle for commuting purposes and try to minimize their travel time/distance. This, however, is not possible to check using the current data due to lack of cycling motives.

When comparing the statistics for the four groups in relative terms (see Appendix, table F.2) the lowest average deviation is found again in group 2 during the morning rush hour. To see if there is a significant difference between the average relative deviations (and between which groups), again a Welch Test and Games Howell post hoc tests are executed. The Welch test shows that on average, the means of the relative deviation between the groups are significantly different. $F = 20.3, p < 0.01$.

The Games Howell test shows that again the 2nd ($M = 1.05$) and 4th ($M = 1.08$) group are again on average significantly different, but also 2nd and 1st ($M = 1.07$) and the 2nd and 3rd ($M = 1.07$) group are significantly different. This means that on average the relative deviation for 2nd group is smaller than in the other groups; therefore routes during the morning rush hour tend to (on average) deviate less from the shortest route than trips made during the other timeslots.

6.1.2 Difference in safety factors

Before we can investigate what factors influence the deviation of the shortest routes, the factors themselves need some investigation. The hypothesis is that the deviation from the shortest route can be explained by the difference in safety factors on the chosen and shortest routes. Before combining both variables in a model, this section investigates whether the safety factors are actually different on the chosen and shortest routes. A method that is often used to do this is comparing means. To investigate whether the means are the same or significantly different for the chosen and shortest routes, a paired samples T-test is executed for all 35 factors that are investigated.

From the total of 35 safety factors, only 5 factors had no significant mean difference between the chosen and shortest route. Meaning that the null hypothesis of the T-test can be rejected

with a 95% (or higher) certainty for 25 safety factors and the alternative hypothesis ($H_A =$ *The average factor of the shortest and chosen routes are significantly different from each other.*) can be accepted.

The factors that had on average no significant mean difference are:

- percentage of scooter path
- percentage of roads with many hindrances
- absolute number of traffic lights
- number of traffic lights per kilometer
- average lemon day score

The means of the safety factors, in general, do not differ that much. On average, most of them differ from 0.1% to 5% and are most of them considered to have a small (or extremely small) effect size. Yet, three safety factors had slightly larger differences on average and considered to have a medium effect size.

On average there was 7.1% more asphalt/concrete on the chosen routes compared to the shortest routes, 9.2% less mixed traffic and 7.6% less street brick roads. Even though the mean differences are small, the type and quality of road surface differed as well and they match expectations from literature and practice. On average a higher percentage of smooth road surfaces and high road surface quality is used for cycling than the shortest route alternative would be able to provide. Vice versa, on average a lower percentage of less smooth and bad quality road surfaces have been used for cycling than the shortest route alternative would be able to provide. Also, most road types match expectations from literature and practice; a higher percentage of bicycle dedicated infrastructure is used for cycling than the shortest route alternative would provide. Other road types show the opposite effect.

Yet, people also seem to prefer to travel next to busy roads, lower speeds, fewer accidents, more intersections and roundabouts, illuminated roads and they do not prefer lower crime ratings. It must be noted however that these mean differences are all very small and even though they have shown to be significant, the results must be interpreted carefully. For example, the average speeds on the chosen routes 39.9 km/h whereas the average speed on the shortest routes is only slightly higher, 41.0 km/h. Also, the average percentage of other/bumpy road surfaces has shown to be higher on the chosen route, the average percentage on chosen routes is only 0.2% which is 0.1% higher than on the shortest routes. Therefore, the comparison of means alone does not necessarily explain (well) what people actually prefer. All other results can be found on Github (*mean_difference_safetyfactors.xlsx* & *chosen_shortest_delta_stats.csv*)

Until now both variables (safety and deviation) have been analyzed individually. We know that the chosen routes are indeed significantly longer than the shortest routes and that most

safety factors are indeed significantly different on the chosen routes compared to the shortest routes. Yet, we still do not know how these two variables are related. The next section will combine both variables in a simple linear model to find out whether the Δ safety factors (i.e., the difference between safety factors on the chosen versus shortest route) explains the amount of deviation from the shortest route.

6.2 Regression analysis

This section will investigate whether people deviate further from the shortest route for a higher/lower safety attribute on the chosen route. First, this is examined on the scale of the total 5,049 routes. The exception, however, is the safety factor *subjective social safety*. As mentioned in section 3.2.3, due to the many missing values in the original data, only those routes with a maximum of 25% missing values are analyzed. So, for this safety factor a total of 2,825 routes are analyzed instead. Additionally, to take into account the different time periods of the day, the routes are also analyzed on the scale of the four different timeslots as mentioned in section 5.3.

The null and alternative hypotheses of the linear regression are:

H_0 = *There is no relationship between the amount of deviation from the shortest route and the difference between 'safety factor' on the chosen and shortest route.*

H_A = *There is a significant relationship between the amount of deviation from the shortest route and the difference between 'safety factor' on the chosen and shortest route.*

The null hypothesis will be rejected with 95% certainty when the model's probability of the corresponding F -statistic is $p \leq 0.05$.

6.2.1 General outcomes

In general, the difference in safety factors on the chosen and shortest have in many models a significant influence on the amount of deviation from the shortest route. Table 6.2 shows an overview of the results per safety factor. Significant safety factors are marked with an X . The strength of these associations (R^2) are classified as follows:

- no asterisk = R^2 value smaller than or equal to 0.01
- * = R^2 value between 0.01 and 0.05 (incl.)
- ** = R^2 value between 0.05 and 0.10 (incl.)
- *** = R^2 value between 0.10 and 0.20 (incl.)

The exact R^2 values can be found on Github (*regression_r2values.xlsx*).

A positive relationship (*green*) means that people deviate further from the shortest route for a higher share of that specific factor on the chosen route than on the shortest route alternative. A negative relationship (*red*) therefore indicates that people deviate for a lower share of that specific factor on the chosen route than on the shortest route.

Tab. 6.2.: Significant safety factors per group

		Total	1	2	3	4
Accidents	Absolute number of traffic accidents	x	x*	x	x	x
	Relative number of traffic accidents					
Social safety	Relative total number of crimes	x	x*		x**	x
	Relative number of violent/sexual crimes	x	x*			x
	Lemon score day	x		x*		x*
	Lemon score night	x		x*		x*
Illumination	Good illumination	x	x*	x*		x
	Limited illumination	x*	x**	x***	x*	x*
	No illumination	x	x**	x**	x*	
Road type	Bicycle suggestion lane	x	x*	x	x	x
	Mixed traffic	x	x*	x	x	
	Scooter lane	x		x*	x	
	Scooter path	x		x		
	Bicycle boulevard				x	
	Bicycle path				x	
	Bicycle lane					
	Speed limits	x			x	x*
Road surface	Asphalt/concrete	x			x*	x
	Street tiles	x			x*	
	Street bricks					
	Semi-surfaced	x		x	x	
	Unsurfaced	x	x*	x	x*	x
	Other/bumpy road surface	x	x*	x***	x	
Road surface quality	Good road surface quality	x		x		x
	Acceptable road surface quality	x		x		x
	Bad road surface quality	x		x	x	x
Intersections & control	Total intersections and roundabouts	x				x
	Intersections	x				x
	Roundabouts		x*			
	Absolute number of traffic lights	x*	x*	x	x*	x*
	Relative number of traffic lights					
Obstacles	Many hindrances	x	x*		x*	
	Acceptable hindrances		x*	x*		
	Little hindrances	x	x*	x*	x	
Traffic intensity	Next to a busy road					

Only 4 out of the total 35 explanatory variables have shown to be significant (and relevant) in all timeslots during the day, namely: the absolute number of traffic accidents and traffic lights, bicycle suggestion lane and unsurfaced roads. Investigating effects for the routes within a particular timeslot did sometimes lead to the same conclusion as for the general analysis, but also contradicted what had been found on a larger scale or in other timeslots during the day. Furthermore, the effect size and model fit measures are in almost all cases (very) low, meaning that predicting the deviation from the shortest route based on the difference in safety factors, a linear model with only one explanatory variable is not a very good model to do so. The effect size and model fit, however, did increase when investigating the influence of the explanatory variables in the different timeslots, compared to the general analysis based on the total routes. An overview of all significant safety factors are provided in section F.2 and all models and outcomes (including R^2 , coefficients and assumption tests) are available on Github (*significant_regression_outcomes.xlsx*).

6.2.2 Outcomes per safety factor

Illumination

Based on findings in literature and policy, it is expected that people would prefer illumination in darkness and that they might alter their route to travel a better lit route. The regression analysis has shown that the presence of illumination indeed significantly influence how far people deviate from the shortest route. As can be seen in table 6.2 in darkness the coefficient is positive (green colored) which means that if there is a higher coverage of illumination (either limited or good) on the chosen route a person would make a significantly larger detour from the shortest route. The strength of this linear model can be categorized as small, explaining approximately 1% of the variance. The model, however, has also shown to be significant in general and in timeslots where there is no need for illumination. This can suggest that the presence of illumination might be closely linked to other factors that are not taken into account in this model.

Road surface

The road surface and road surface quality are expected to influence route behavior as well. Previous research suggests that people prefer a smooth and well-maintained road surface without (deep) potholes, cracks or wide gaps between bricks or stones that make up the road surface. This is shown in the outcomes of the analysis as well, even though there is no significant relationship for all road surfaces during all timeslots. In general people tend to deviate further from the shortest route for a higher share of asphalt/concrete, but it does not show in the dark or during morning rush hour. Semi- and unsurfaced roads are actually avoided on the chosen routes during most timeslots.

The analysis also showed that people tend to deviate further for good road surface quality and avoid the roads that were in either acceptable or bad conditions. The relationship, however, has not been found throughout the day, but did show up to be significant during the morning rush hour and the non-rush hour part of the day (in daylight). The effect size and model fit, however, are in general and throughout the day considered to be (very) small, explaining less than 1% of the variance.

Surprisingly though, street tiles which are considered of the smoother road surfaces, in general, show a negative relationship; meaning that they make up a smaller part of the chosen route than the shortest route alternative would provide. However, this behavior only occurred in the evening rush when analyzing individual timeslots. Most surprisingly and unexpected is probably that during almost all timeslots people have shown to deviate further from the shortest route for more bumpy road surfaces (semi- and unsurfaced roads excluded) on their route. However, it should be noted that even though the analysis shows that people prefer a higher share of these roads on their route, the coverage of these roads in the area is extremely small. The Δ other/bumpy roads, too, are very small for almost all routes (as is mentioned already in the previous section). The effect size and model fit for most road surface types are small to medium explaining around 1 to maximum 10% of the variance.

Road types

As the literature suggests, we would expect that people would prefer to travel on bicycle specific infrastructure and maybe avoid mixed traffic situations. Some studies have also found that people would even prefer bicycle paths to be completely separated from motorized traffic. This, mostly, however does not show in the analysis. It should be noted however that all found relationships regarding road types are all considered to be (extremely) small.

One thing that does agree with literature is that people tend to avoid mixed traffic which is confirmed by the negative relationship throughout almost the entire day and in general. The only timeslot in which this does not show at all is when the roads are less crowded in non-rush hour periods during daylight. Other than that, bicycle specific road types such as the bicycle path, lane and bicycle boulevard, in general and throughout most parts of the day, did not significantly influence the deviation from the shortest route. Only during the evening rush hour a significant negative relationship occurred which does not necessarily make sense. This might indicate (due to the extremely low model fit) that the relationship is actually not relevant, meaningful or well captured in this linear model with only one explanatory variable (Field, 2009).

The cyclists do however show a preference for bicycle suggestion lanes, because the model found that people tend to deviate further from the shortest route for a higher share of bicycle suggestion lanes on their chosen route. This relationship has been found in general and throughout the entire day. This is quite unexpected, because it contradicts earlier findings.

Yet, it is impossible to know if they prefer bicycle suggestion lanes over another road type so it might just be the case that because they rather avoid mixed traffic, they prefer this type of infrastructure. Also, it might just be the case that in the Netherlands these types of roads are safe enough and that because of the large amount of cyclists throughout the country (compared to other countries) this road type is considered to be valuable cycling infrastructure after all.

Likewise, the analysis showed that cyclists, in general, do not mind sharing their road with only scooters. In fact, in general and especially during the morning rush hour they deviate further from the shortest route for a higher share of these roads. This makes sense since the speed difference on these types of roads (in the built-up area at least) between cyclists and scooters is low.

Finally, high speed limits do not seem to scare off the Dutch cyclists either. The model showed that in general and throughout some parts of the day people deviate further from the shortest route alternative to travel on roads that have a higher average speed. However, it should be noted that average speed on route-level is highly affected by outliers, information that is lost with only investigating the mean as a measure.

Intersections & control

In the general analysis it showed that people, as previous research often has found, actively avoid intersections. They tend to deviate further to travel a smaller fraction of their route on a road that is part of an intersection. However, this only shows up in the non-rush hour periods during daylight. Also, people prefer to travel on roundabouts and they are willing to significantly deviate from the shortest route because of this, but this only seems to occur in the darkness. It might be the case that these roundabouts are closely linked to other factors that are not taken into account in this model, such as the presence of illumination or traffic flow.

Moreover, the number of traffic lights has shown to have a significant impact on the deviation from the shortest route, but not always as expected. Only during the morning rush hour people tend to deviate more to decrease the amount of traffic lights on their route, but during all the other timeslots and in general people actually tend to deviate further to routes with more traffic lights. These relationships however are all considered (extremely) small.

Accidents

Again, the Dutch cyclists are mostly not deterred by the cycling accidents that have happened. In fact, during most of the day, people actually deviate further from the shortest route to routes where more accidents have taken place. Only during the morning rush-hour people seem to avoid these roads, but this might also be linked to other factors that are not taken into account in this model. Generally, accidents happen on roads with larger traffic volumes

(often main roads) that, therefore, automatically results in more accidents. The accidents themselves do not attract people, but rather these roads are closely linked to other factors that are of interest for people to travel these roads. For instance, these roads might be part of the main cycling network. It might therefore be possible that people avoid these types of roads to take shortcuts during morning rush hour to minimize their travel time due to commuting purposes. Roads on the main cycling network can, especially in the morning rush hour, have high crowding levels. To ensure safe crossing of the streets, traffic lights are often placed. This might explain the different behavior during the rest of the day.

Social safety

According to the models, people tend to deviate further from the shortest route to travel on routes with higher crime rates and worse subjective social safety ratings. Even though this relationship is not significant throughout the whole day, it does not make much sense. It is therefore very likely that, similar to the accidents, other factors are closely linked to these roads that are not in the model. Furthermore, the CBS data regarding objective social safety is normalized using the number of inhabitants per neighborhood. This results in creating very high relative crime rates for e.g. park areas where almost no people live but in reality attracts many people to cycle for recreational purposes. In addition, the Δ subjective social safety (Lemon) rating is generally very small and are the model fit and effect sizes of these relationships (extremely) small.

Obstacles

Even though some of the literature states that people are willing to travel significantly further to avoid obstacles, either people or objects, this is not supported by these cyclists. The effect of obstacles can only be analyzed during the rush hour periods due to the nature of the data that has been used. People seem to avoid roads that do not have many obstacles (little hindrances), so there must be other reasons why cyclists are actively avoiding them. People actually rather travel on roads (and they are willing to significantly deviate from the shortest route because of it) that are associated with acceptable or many hindrances. As was found in other research as well, apparently people are not deterred from cycling on these roads and find them safe enough, or they adopt a certain driving style.

It should be noted however that almost all routes have an extremely low coverage of 'many hindrances' and the differences between this factor on the chosen and shortest routes are also extremely small. So, in reality, the obstacles people face will not be that big. Additionally, the effect size and model fit of this relationship are considered to be small.

As is the case with other safety factors, it is very likely that other characteristics of these roads

are more important which explains why people rather deviate from the shortest route to cycle on these roads.

Traffic intensity

Finally, also traffic intensity is investigated. Since the IC-ratio was not possible to use for the analysis anymore due to too many missing values, the effect of the difference in road level alone is investigated but showed no significant relationship that explains the amount of deviation from the shortest route.

Compared to motorized vehicles, the behavior of cyclists is still relatively unknown. At the moment, cycling networks and traffic models are derived from car networks and models which are not at all suited to the behavior of cyclists. It is essential to investigate and understand what factors play a role in route choice behavior to advance our understanding of human mobility and to improve cycling safety and its infrastructure. At this point, cycling behavior is not sufficiently researched on the proper scale to accurately do so.

Mainstream traffic research often assumes that people take the shortest route to minimize their travel costs, but the generalized costs for cyclists are made up very differently from motorized vehicles. For cyclists, factors, such as safety, cannot be ignored because safety can be a factor that is as important (or even more important) than travel time or distance. Therefore, safety can be the reason why a person decides not to cycle at all or why a cyclist travels a particular route.

In this thesis, large volumes of GPS data and GIS methods are used to provide insight into the route choice behavior of Dutch cyclists and its relation to safety. Despite the fact that previous research suggests that safety plays a role in route choice behavior, people often tend to act differently than expected and that might especially be the case in a Dutch context where cycling is very rooted in the culture and in general already very safe.

To investigate whether the data actually supports this, the following main research question has been created and will in this section be answered.

To what extent do safety-factors of the environment influence cyclists' route choice behavior and how can this be empirically measured using GPS tracks and GIS methods?

7.1 Answering research questions

1A. What safety factors of the environment are relevant in terms of cyclists' route choice?

Many studies have found that cyclists significantly deviate from the shortest route, but that there is a limit to the amount of deviation. The cycling circumstances in different countries over the world have a big influence on this. Safety might play a more prominent role in determining one's route in car-dominated countries with fragmented and sparse cycling facilities than in the Netherlands.

Traffic safety is measured by the number of accidents, crashes, deaths etcetera. The problem of cycling accidents, however, is that they are often not registered. Also, accidents only reflect the tip of the iceberg in terms of perceived safety, because near misses happen much more frequently and can have a much bigger effect on perceived safety.

Nonetheless, locations where accidents have taken place are often considered as dangerous and therefore people might want to avoid these locations when choosing their route. Safety should however not only be measured by the results (*accidents*) but also by the causes. One of these potential causes is the *road type*. Safety on the road can be explained by the degree of mixing or separation of road users. Especially the mixing of road users with substantial speed differences is considered unsafe. In general, previous research has found that people are willing to cycle further to cycle on bicycle dedicated infrastructure. Also, the *road surfaces* and the *road surface quality* have had this effect in previous studies; people prefer smooth surfaces that minimize vibrations without any cracks, gaps or potholes in the pavement.

Cyclists are known to avoid *intersections* and especially *traffic lights* because stopping causes one to lose momentum. Traffic lights might however also be preferred to cross intersections safely (or faster) because cycling near an intersection increases the risk of accidents significantly due to the interaction with other road users. Interaction with other road users increases when the streets are crowded. This can be relevant for both bicycle paths with high volumes of cyclists or the shared street with high volumes of diverse road users. Studies have shown that people are willing to cycle longer routes to avoid *high levels of crowding*.

Other reasons for detouring are to avoid *obstacles* and avoiding places because of *social safety*. Finally, route choice behavior can be very much related to day circumstances. People might act differently in rush hour than in non-rush hour. Also in night conditions when it is dark outside, people might want to travel further for the presence of illumination on their route.

1B. How can the safety factors of the environment be measured?

Many different datasets exist that represent these safety factors. In this thesis safety data is combined from: the Fietzersbond, BRON, Open Streetmap, CBS, the traffic model from the Municipality of Tilburg and finally data from the Lemon project is used to measure subjective social safety in the municipality of Tilburg.

Different aggregation methods have been used to express the safety factors on route-level. Depending on the characteristics of the data the safety factors are expressed by:

1. the absolute number of *X* on the route, or:
2. the density of *X* per kilometer on the route, or:
3. the distance weighted average of *X* on the route, or:
4. the coverage of *X* (in %) on the route.

Most of the data (often categorical or ordinal data) are expressed using the 4th method by calculating the coverage of the safety factor over the total route.

The traffic accidents and traffic lights are expressed by the 1st and 2nd aggregation methods. The objective social safety is measured by the number of crimes per 1000 inhabitants and the subjective social safety by scores/ratings that represent how safe people feel in their neighborhood during the day and night. Both the objective and subjective safety data are provided on a neighborhood-level. This data and the data regarding speed limits are aggregated to route-totals using the 3rd method, taking the route distance into account.

2. How can route choice behavior be measured using GIS methods?

Route choice behavior has been researched in many different ways. The most commonly used method to study it is by using stated preference methods. However, the biggest limitation of stated preference studies is that there is no tie to actual behavior because it does not always reflect the reality of the individual's route choice options. Therefore, the preferences often do not manifest in reality.

To deal with this, a GPS based revealed preference method is used in this thesis to study route choice behavior in relation to safety. Using large volumes of GPS data from the Fietstelweek (the largest cycling project of the Netherlands) also tackles the problem of under-sampling that most cycling research deal with and makes it possible to study everyday cycling behavior. Moreover, the sample used for the cycling data seems to be quite representative of the study area (Municipality of Tilburg).

Route choice behavior in this thesis is investigated by comparing the chosen routes to the shortest possible route alternatives in terms of travel distance (*i.e. amount of deviation*) and safety factors (Δ *safety factor*). To create the shortest route alternatives, a cycling network is created in ArcGIS that exists of nodes and edges that are connected. The start and end locations of the routes are determined using several geo-processing tools and spatial join methods. The shortest route alternatives are found based on the Dijkstra algorithm. To efficiently perform the shortest path analysis for all the routes in the study area, a Python script is used that calls different route analysis tools in ArcGIS. Finally, because the safety data is available in various formats (*i.e.*, points, lines and polygons), several geo-processing tools and spatial join methods are used to enrich the cycling network with the safety data.

3. To what extent do these safety-factors statistically influence cyclists' route choice?

Generally, the chosen routes are in fact significantly longer than the shortest routes, but the deviations from the shortest route are on average not that large. The average deviation is 338 meters or in relative terms 1.07 times as far as the shortest route.

Only 4 out of the total 35 explanatory variables have shown to be relevant and to significantly influence the amount of deviation during all timeslots of the day. These factors are: bicycle

suggestion lanes, unsurfaced roads, traffic lights and accidents. Even though all factors do not always significantly explain the amount of deviation, the nature of the relationship (i.e., positive or negative) for all factors does not differ during the day. The only exceptions are traffic lights and accidents. These factors seem to show a different effect on the amount of deviation during the morning rush hour compared to the rest of the day.

The effect size and model fit measures are in almost all cases (very) low. Meaning that the Δ safety factors individually do not explain the amount of deviation from the shortest route that well using a linear model. Interestingly though, the effect size and model fit of the models did increase when the influence of the explanatory variables was investigated delimited by the different timeslots.

Some interesting results that match findings from previous research are:

- In the dark, people tend to deviate further from the shortest route for a higher share of illumination on their route.
- People tend to deviate further from the shortest route for a higher share of smooth (asphalt/concrete) and good quality road surfaces. This behavior however did not show in darkness.
- People are willing to deviate further from the shortest route for a lower share of semi- and unsurfaced and bad quality road surfaces. These effects however did not always show in all timeslots, such as in the dark.
- Throughout most of the day, people tend to deviate further from the shortest route for a lower share of mixed traffic.
- During the rush hour, people tend to deviate further for a higher share of roads where only cyclists and scooters are mixed. This makes sense because the speed differences between the road users are low and still separates cyclists from cars, buses and trucks.
- During the daylight in non-rush hour circumstances people tend to deviate further from their route to avoid intersections.

However, the data also sometimes lacks to support or even contradicts some of the relations that were expected.

- During the evening rush hour, people tend to deviate further from the shortest route for a lower share of street tiles on their route. This is a surprising result because street tiles are considered to be one of the smoother road surfaces.
- Throughout most of the day people seem to be willing to deviate further from the shortest route for a higher coverage of other (i.e. bumpy) roads. However, the coverage of these types of roads in the area is minimal to begin with.

- People tend to deviate further from the shortest route for a higher share of bicycle dedicated infrastructure on their route. People do not seem to be deterred by the perceived safety of these road types or they are in fact safer than has been found so far.
- Most of the Δ bicycle dedicated infrastructure does not show to have a significant effect on the amount of deviation.
- During daylight in the evening rush hour and non-rush hour, people tend to deviate further from the shortest route to travel on roads that have higher speed limits than the shortest route alternative. This does not match the expected behavior because in general cyclists prefer to travel on roads with minimal speed differences between themselves and other road users.
- The number of traffic lights has shown to have a significant impact on the deviation from the shortest route, but not always as expected. Only during the morning rush hour people tend to deviate further to decrease the amount of traffic lights on their route, but during the rest of the day people actually tend to deviate further to routes with more traffic lights.
- Similarly, during the morning rush hour people tend to deviate further from the shortest route to routes where less cycling accidents have taken place, but during the rest of the day people show the opposite behavior. It is clear that cyclists are mostly not deterred by the cycling accidents that have happened. It is very likely that these roads also offer the cyclists some benefits that are not taken into this model.

Most of the relationships that cannot be explained that well, require some nuance. The findings regarding speed limits for instance; people seem to deviate further to travel on routes with a higher average speed limit. Because the mean is used as a measure to investigate the effect of speed limits, it disregards all diversity on the route. By using the mean as a measure for average speed, it results in the loss of information regarding outliers (which highly influence the mean). Furthermore, the average speed limits often are not very high and the average speed limits on the chosen versus the shortest route mostly do not differ that much, e.g. 30km/h versus 31km/h. Even though this relationship is significant, it does not necessarily mean that it is relevant or meaningful. People, in reality, will not base their route decision on a speed difference of 1km/h but instead on considerable speed differences.

In conclusion, route choice behavior is measured by comparing the safety factors and distance on the chosen and shortest route alternative. Investigating the behavior is done by using GPS data from the largest cycling project in the Netherlands. Additionally, GIS is used for combining different datasets and creating the shortest route alternative.

It turns out that several safety factors of the environment do actually influence cyclists' route choice behavior in the way that we would expect based on preferences found in literature. The difference between the presence of illumination, road surfaces and its quality, some road types and intersections on the chosen route compared to the shortest route alternative, explains the size of the detour. However, also a number of safety factors show to influence

route choice in a way that does not match preferences that are found in previous research. For instance, cyclists do not seem to be bothered by the number of accidents on a route, because they deviate further from the shortest route to reach routes with a higher number of accidents. Many effects are difficult to explain because the measured safety factor might be closely linked to other characteristics (more important than safety) of the route that are not taken into account in the simple linear models. Furthermore, some safety factors are measured in a way that disregards all diversity on the route. Despite the fact that many relationships exist that significantly explain the amount of deviation from the shortest route, the effect sizes and model fits are low. Besides, these relationships often do not show up to be significant throughout the day.

7.2 Discussion & limitations

This research project should be considered as a first attempt at data driven research to investigate route choice behavior and its relation to safety using GIS methods.

Route choice behavior in this thesis is investigated using a GPS based revealed preference method. The problem however with this method and the used cycling data from the Fietstelweek is that only the final decision can be observed. Ideally, it would be better to use a mixed methods approach of stated- and revealed preference to gain more insight into the cyclists' preferences or motivations behind their route choices. As is also the case in this research, travel motives are often missing in (large) GPS studies, when in fact this information can be crucial to better understand route choice behavior.

Using open- and freely available cycling data from the Fietstelweek for this project has many benefits, but also a number of limitations have been encountered. The biggest advantage of using this data is that the data is already available and new data collection was not necessary. Also the nature of the project finally made it possible to study cycling for everyday use, which is often not possible with cycling data. Since the Fietstelweek is a large-scale and well-promoted project, large volumes of GPS data were collected. Moreover, the easy-to-use Fietstelweek App does not require any actions from the users to track their movements, making it very easy and accessible for people to join the project. However, using free and publicly available cycling data also has its limits due to privacy.

First of all, there is no raw GPS data available. Instead, the GPS data is map matched to a cycling network so no exact locations were provided. The map match process makes some assumptions as to where people would likely have cycled. Also, all trips are anonymized by cutting off 200 meters from the start and end-location of a trip and a time shift between 0-15 minutes are applied randomly. This means that the derived start and end locations are not the exact start and end locations. Furthermore, because the time of the routes are only provided

in one hour units, and the use of the random time shift it is possible that some routes are assigned to the wrong timeslot.

Even though the sample seems quite representative for the study area, it is impossible to know if this is actually true because one can enter the Fietstelweek anonymously. So, for an unknown fraction of the sample there is no personal data available that can be used for the representativity check. Therefore, it is safer not to generalize the results of the data analysis.

Other limitations of this research are due to safety data availability and quality. Even though the Fietstersbond has the most accurate and rich cycling network that is created and kept up-to-date using volunteers (VGI), many roads still lacked the data for safety attributes. To carry out the analysis despite the many missing values, supervised machine learning methods have been used to estimate these missing values.

Also other datasets were incomplete. As mentioned before, the actual number of cycling accidents is in reality much higher than the ones that are registered in BRON. Furthermore, the traffic model did not cover enough detail, resulting in only being able to match a couple of roads with data from the traffic model. Consequently, one of the safety factors could not be tested in the analysis. Also there does not seem to be a complete dataset regarding the traffic lights, because in the end three different datasets needed to be combined.

The analysis showed that many models exist in which the difference in safety factors could significantly explain the amount of deviation. Previous research about cyclists' preferences supported several of these outcomes. However, in general, the simple linear models did not perform that well and it proves that these relationships are not well captured in simple linear models with only one explanatory variable.

Even though some insight is acquired regarding the effect every (Δ) safety factor individually has on the amount of deviation from the shortest route, these models are not suitable for other purposes such as predicting route choice behavior.

In order to gain more and better insight into the complex relationship between route choice and safety, other (non-linear) methods should be explored and multiple safety factors should be taken into account. To model route choice behavior, also other factors than safety should be considered as well. In the end, people tend to choose a route that is safe enough for their travel motives and available time.

7.3 Recommendations for future research

- For both scientific and social purposes there is a need for (better) models that explain traffic behavior. Especially the behavior of cyclists seems difficult to capture in the current models, because there is still a lot to learn about cyclists' (route choice) behavior. More research is necessary in order to create accurate and flexible models that can provide more insight into the relationship between route choice behavior and safety by using multiple explanatory factors (e.g. purpose and travel time).
- Investigating the relationship between route choice behavior and safety by combining both stated preference and revealed preference methods might provide some new interesting insights. Also other and new analysis methods and techniques (non-linear) such as data mining could offer a lot of new insights into the relationship between route choice behavior and safety.
- The need for safety does not only vary based on cycling motives, but it also differs per person. Safety itself, and the need for safety is different for every person. Therefore it can also be valuable to investigate the relation between route choice behavior and safety for different groups of people (e.g. young adults versus seniors). Also, with the increasingly gaining popularity of bicycle alternatives such as e-bikes and speed pedelecs, the influence of safety on route choice behavior might be as or even more important to investigate. The needs of people using an e-bike or speed pedelec might even be different from regular cyclists.
- Analyzing short- and long-term effects of cycling behavior due to changes to the infrastructure.

The infrastructure is always changing for multiple reasons, resulting in safer and sometimes even less safe circumstances. In reality, this might result in people traveling different routes than before. In order to understand route choice behavior even better, infrastructural changes need to be taken into account as well and analyzed over time.

Bibliography

- AMS (2016). Allegro: unravelling slow mode traveling and traffic. <https://www.ams-institute.org/news/allegro-unraveling-slow-mode-traveling-and-traffic/>.
- Beura, S. K., Chellapilla, H., and Bhuyan, P. K. (2017). Urban road segment level of service based on bicycle users' perception under mixed traffic conditions. *Journal of modern transportation*, 25(2):90–105.
- Bikeprint (2016). Fietstelweek data. <http://www.bikeprint.nl/fietstelweek/>.
- Blokker, B. (2013). Fietzers, die zijn als een zwerm spreeuwen. <https://www.nrc.nl/nieuws/2013/05/14/fietzers-die-zijn-als-een-zwerm-spreeuwen-1246434-a530237>. [Accessed on 20-11-2017].
- Boekhoudt, C., Te Brömmelstroet, M., and Thüsh, M. (2017). De tijd vliegt als je plezier hebt: reistijd op de fiets is persoons-, locatie en tijdsafhankelijk. <http://verkeerskunde.nl/Uploads/2017/10/FInal-Verkeerskunde-Reistijd-en-route-keuze.pdf>.
- Broach, J., Dill, J., and Gliebe, J. (2012). Where do cyclists ride? a route choice model developed with revealed preference gps data. *Transportation Research Part A: Policy and Practice*, 46(10):1730–1740.
- BRON (2017). Bron accident data 2011-2015. <https://www.via.software/>.
- Buehler, R. and Dill, J. (2016). Bikeway networks: A review of effects on cycling. *Transport Reviews*, 36(1):9–27.
- Casello, J., Nour Omar, A., Rewa Cyril, K., and Hill, J. (2011). Analysis of Stated-Preference and GPS Data for Bicycle Travel Forecasting. *Transportation Research Board, 90th Annual Meeting*, 5(January):18p.
- Caulfield, B., Brick, E., and McCarthy, O. T. (2012). Determining bicycle infrastructure preferences—a case study of dublin. *Transportation research part D: transport and environment*, 17(5):413–417.
- CBS (2016a). Geregistreerde criminaliteit per gemeente, wijk en buurt, 2010-2015. <https://www.cbs.nl/nl-nl/maatwerk/2016/45/geregistreeerde-criminaliteit-per-gemeente-wijk-en-buurt-2010-2015>.
- CBS (2016b). Population data per municipality. <http://www.cbsinuwbuurt.nl>.
- CBS (2016c). Veiligheidsmonitor 2016.
- CBS (2018). In 2017 meer verkeersdoden op de fiets dan in de auto. <https://www.cbs.nl/nl-nl/nieuws/2018/17/in-2017-meer-verkeersdoden-op-de-fiets-dan-in-de-auto>.

- Cheng, M.-Y. and Chang, G.-L. (2001). Automating utility route design and planning through gis. *Automation in Construction* 10 (4), pages 507–516.
- CROW (2015). Fietsstroken: de maat genomen. *Fietsverkeer*, 37:10–14.
- CROW (2016). Drukke op fietspaden. *Fietsverkeer*, 39:9–10.
- Dill, J. and Gliebe, J. (2008). Understanding and measuring bicycling behavior: A focus on travel time and route choice.
- Dozza, M. and Werneke, J. (2014). Introducing naturalistic cycling data: What factors influence bicyclists' safety in the real world? *Transportation research part F: traffic psychology and behaviour*, 24:83–91.
- Edwards, D. (2017). Selftracking: Het nut en de valkuilen. <http://www.emergent.city/selftracking-het-nut-en-de-valkuilen/>. [Accessed: 2017-10-15].
- Ehrgott, M., Wang, J. Y., Raith, A., and Van Houtte, C. (2012). A bi-objective cyclist route choice model. *Transportation research part A: policy and practice*, 46(4):652–663.
- Elvik, R., Vaa, T., Høy, A., and Sørensen, M. (2009). *The handbook of road safety measures*. Emerald Group Publishing.
- Field, A. (2009). *Discovering Statistics Using SPSS*. Sage.
- Fietsersbond (2016). Fietsersbond data routeplanner. <https://routeplanner.fietsersbond.nl/pagina/handleiding>.
- Fietsersbond (2018). Metadata database fietsersbond. <http://docplayer.nl/7295641-Metagegevens-database-fietsrouteplanner-fietsknooppunten-poi-s.html>.
- Fishman, E., Washington, S., Haworth, N., et al. (2012). Understanding the fear of bicycle riding in australia. *Journal of the Australasian College of Road Safety*, 23(3):19.
- Garrard, J., Rose, G., and Lo, S. K. (2008). Promoting transportation cycling for women: the role of bicycle infrastructure. *Preventive medicine*, 46(1):55–59.
- Geertman, S. C. and Ritsema Van Eck, J. R. (1995). Gis and models of accessibility potential: an application in planning. *International journal of geographical information systems*, 9(1):67–80.
- Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120.
- Harms, L., Bertolini, L., and Brömmelstroet, M. T. (2016). Performance of municipal cycling policies in medium-sized cities in the netherlands since 2000. *Transport Reviews*, 36(1):134–162.
- Heinen, E., Van Wee, B., and Maat, K. (2010). Commuting by bicycle: an overview of the literature. *Transport reviews*, 30(1):59–96.
- Hendriks, R. and Bussche, D. (2016). Fietsdata: wat kun je er nu al mee? *Fietsverkeer*, 38:16–23.
- Hölzel, C., Höchtel, F., and Senner, V. (2012). Cycling comfort on different road surfaces. *Procedia Engineering*, 34:479–484.

- Hood, J., Sall, E., and Charlton, B. (2011). A gps-based bicycle route choice model for san francisco, california. *Transportation letters*, 3(1):63–75.
- Jacobsen, P. L. (2003). Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury prevention*, 9(3):205–209.
- JeKoPhoto (2016). Twilight calculator. <http://jekophoto.eu/tools/twilight-calculator-blue-hour-golden-hour/index.php>.
- Joolink, H. (2016). Routekeuze fietsers enschede: vergelijking van de routekeuzevoorkeur van fietsers in enschede, met de afgelegde route. B.S. thesis, University of Twente.
- Keypoint (2016). Age and cycling purposes fietstelweek 2016.
- Krizek, K. J., El-Geneidy, A., and Thompson, K. (2007). A detailed analysis of how an urban trail system affects cyclists' travel. *Transportation*, 34(5):611–624.
- Langley, J. D., Dow, N., Stephenson, S., and Kypri, K. (2003). Missing cyclists. *Injury prevention*, 9(4):376–379.
- Lawson, A. R., Pakrashi, V., Ghosh, B., and Szeto, W. (2013). Perception of safety of cyclists in dublin city. *Accident Analysis & Prevention*, 50:499–511.
- Lemon-Onderzoek (2015). Lemon de leefbaarheidsmonitor: Leefbaarheid volgens bewoners. <http://lemon-onderzoek.nl/index.php/gemeenten/tilburg/>.
- Mathworks (2018). Train regression models in regression learner app. <https://nl.mathworks.com/help/stats/train-regression-models-in-regression-learner-app.html>. [Accessed: 2018-03-19].
- Menghini, G., Carrasco, N., Schüssler, N., and Axhausen, K. W. (2010). Route choice of cyclists in zurich. *Transportation research part A: policy and practice*, 44(9):754–765.
- Methorst, R. (2007). Shared space: veilig of onveilig? een bijdrage die er op gericht is om een populaire ontwerpfilosofie te objectiveren. Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk 2007, 22 en 23 november 2007, Antwerpen.
- of Tilburg, M. (2017). Traffic model municipality of tilburg. <https://www.tilburg.nl/actueel/gebiedsontwikkeling/mobiliteitsplan-2040/>. [Accessed on 08-03-2018].
- OSM (2017). Stoplichten in nederland. <https://data.overheid.nl/data/dataset/stoplichten-in-nederland>.
- Pinder, D. (2001). Ghostly footsteps: voices, memories and walks in the city. *Ecumene*, 8(1):1–19.
- Pucher, J. and Buehler, R. (2008a). Cycling for everyone: lessons from europe. *Transportation Research Record: Journal of the transportation research board*, (2074):58–65.
- Pucher, J. and Buehler, R. (2008b). Making cycling irresistible: lessons from the netherlands, denmark and germany. *Transport reviews*, 28(4):495–528.
- Pucher, J. and Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: lessons from the netherlands and germany. *American journal of public health*, 93(9):1509–1516.

- Reurings, M., Vlakveld, W., Twisk, D., Dijkstra, A., and Wijnen, W. (2012). Van fietsongeval naar maatregelen: kennis en hiaten. *SWOV, Ed*, 203.
- Rietveld, P. and Daniel, V. (2004). Determinants of bicycle use: do municipal policies matter? *Transportation Research Part A: Policy and Practice*, 38(7):531–550.
- Rijksoverheid (2008). Strategisch plan verkeersveiligheid 2008-2020 : Van, voor en door iedereen. <https://www.rijksoverheid.nl/documenten/beleidsnota-s/2008/07/10/strategisch-plan-verkeersveiligheid-2008-2020-van-voor-en-door-iedereen>. [Accessed: 2017-10-10].
- Rijksoverheid (2017). Vertrouwen in de toekomst: regeerakkoord 2017-2021. <https://www.kabinetsformatie2017.nl/documenten/publicaties/2017/10/10/regeerakkoord-vertrouwen-in-de-toekomst>. [Accessed: 2017-10-10].
- Romanillos, G., Zaltz Austwick, M., Ettema, D., and De Kruijf, J. (2015). Big data and cycling. 36:1–20.
- Sanders, R. L. (2013). *Examining the cycle: how perceived and actual bicycling risk influence cycling frequency, roadway design preferences, and support for cycling among bay area residents*. University of California, Berkeley.
- Sanders, R. L. (2015). Perceived traffic risk for cyclists: The impact of near miss and collision experiences. *Accident Analysis & Prevention*, 75:26–34.
- Scheider, S. (2017). Python script: Correction of cycling network. <https://github.com/cynthiadevos/ThesisGIMA/blob/master/RepairNetwork.py>.
- Schepers, P., Hagenzieker, M., Methorst, R., Van Wee, B., and Wegman, F. (2014). A conceptual framework for road safety and mobility applied to cycling safety. *Accident Analysis & Prevention*, 62:331–340.
- Schepers, P., Twisk, D., Fishman, E., Fyhri, A., and Jensen, A. (2017). The dutch road to a high level of cycling safety. *Safety science*, 92:264–273.
- Sener, I. N., Eluru, N., and Bhat, C. R. (2009). An analysis of bicycle route choice preferences in texas, us. *Transportation*, 36(5):511–539.
- Shankwiler, K. D. (2006). *Developing a Framework for Behavior Assessment of Bicycle Commuters: A Cyclist-Centric Approach*. PhD thesis, Georgia Institute of Technology.
- Spinney, J. (2009). Cycling the city: Movement, meaning and method. *Geography Compass*, 3(2):817–835.
- Standen, C., Crane, M., Collins, A., Greaves, S., and Rissel, C. (2017). Determinants of mode and route change following the opening of a new cycleway in sydney, australia. *Journal of Transport & Health*, 4:255–266.
- Stinson, M. and Bhat, C. (2003). Commuter bicyclist route choice: Analysis using a stated preference survey. *Transportation Research Record: Journal of the Transportation Research Board*, (1828):107–115.
- SWOV (2017). Factsheet: Wegwijzer verkeersveiligheidscijfers. <https://www.swov.nl/feiten-cijfers/datasheet/wegwijzer-verkeersveiligheidscijfers>. [Accessed on 2017-11-13].

- Tilahun, N. Y., Levinson, D. M., and Krizek, K. J. (2007). Trails, lanes, or traffic: Valuing bicycle facilities with an adaptive stated preference survey. *Transportation Research Part A: Policy and Practice*, 41(4):287–301.
- Tingvall, C. and Haworth, N. (2000). Vision zero: an ethical approach to safety and mobility. In *6th ITE International Conference Road Safety & Traffic Enforcement: Beyond*, volume 1999, pages 6–7.
- Tirry, D. and Steenberghen, T. (2014). Een uitwisselmodel voor verkeersveiligheidsindicatoren. steunpunt verkeersveiligheid 2012-2015, ra-2014-004.
- Van der Coevering, P., Kruijff, J., and Bussche, D. (2014). Bike print: Policy renewal and innovation by means of tracking technology. https://www.cvs-congres.nl/cvspdfdocs_2014/cvs14_033.pdf.
- Van der Knaap, P. (2017). Verkeersveiligheid moet een nationale prioriteit worden”. 2.
- Van der Schaaf, T. W., Lucas, D. A., and Hale, A. R. (2013). *Near miss reporting as a safety tool*. Butterworth-Heinemann.
- Van Duppen, J. and Spierings, B. (2013). Retracing trajectories: the embodied experience of cycling, urban sensescape and the commute between ‘neighbourhood’ and ‘city’ in utrecht, nl. *Journal of Transport Geography*, 30:234–243.
- Van Genugten, W. and Van Overdijk, R. (2016). Het optimaliseren van fietsgedrag in verkeersmodellen. <https://www.royalhaskoningdhv.com/nl-nl/nederland/nieuws/nieuwsberichten/royal-haskoningdhv-en-tu-eindhoven-tonen-gedrag-van-fietsers-aan/5682>. [Accessed: 2017-11-28].
- Vandenbulcke, G., Dujardin, C., Thomas, I., de Geus, B., Degraeuwe, B., Meeusen, R., and Panis, L. I. (2011). Cycle commuting in belgium: spatial determinants and ‘re-cycling’ strategies. *Transportation research part A: policy and practice*, 45(2):118–137.
- Vedel, S. E., Jacobsen, J. B., and Skov-Petersen, H. (2017). Bicyclists’ preferences for route characteristics and crowding in copenhagen—a choice experiment study of commuters. *Transportation research part A: policy and practice*, 100:53–64.
- Verkaik, Y. (2017). Verkeersdeskundige: ‘registratie verkeersongevallen onvolledig’. <https://demonitor.ncrv.nl/verkeersveiligheid/verkeersdeskundige-registratie-verkeersongevallen-onvolledig>. [Accessed on 2017-11-13].
- Vis, A. (1994). Ontwerp en uitvoering van veilige fietsvoorzieningen. *SWOV: Stichting Wetenschappelijk Onderzoek Verkeersveiligheid*.
- Wardrop, J. G. (1952). Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362.
- Wegman, F., Zhang, F., and Dijkstra, A. (2012). How to make more cycling good for road safety? *Accident Analysis & Prevention*, 44(1):19–29.
- Wijlhuizen, G. and Aarts, L. (2014). Monitoring fietsveiligheid: Safety performance indicators (spi’s) en een eerste opzet voor een gestructureerd decentraal meetnet.

- Winters, M., Teschke, K., Grant, M., Setton, E., and Brauer, M. (2010). How far out of the way will we travel? built environment influences on route selection for bicycle and car travel. *Transportation Research Record: Journal of the Transportation Research Board*, (2190):1–10.
- Zhu, S. and Levinson, D. (2015). Do people use the shortest path? an empirical test of wardrop's first principle. *PloS one*, 10(8):e0134322.
- Zimmermann, M., Mai, T., and Frejinger, E. (2017). Bike route choice modeling using gps data without choice sets of paths. *Transportation research part C: emerging technologies*, 75:183–196.

List of Figures

1.1	Numbers of recorded cyclist fatalities per billion bicycle kilometers. Source:Schepers et al. (2014)	2
2.1	Relative amount of road users in traffic accidents 2001-2015 (BRON)	8
2.2	Relative development of bicycle accidents in % over period 2001-2015. Baseline = 2001	9
2.3	Types of bicycle infrastructure	11
2.4	Summary of safety factors effecting route choice	18
3.1	Conceptual model of the thesis process	22
3.2	Shortest path versus alternative path	24
3.3	Comparing safety attributes	26
3.4	Supervised Machine Learning proces. Based on Mathworks (2018)	27
3.5	Difference in safety factors	30
4.1	Travelled roads during Fietstelweek 2016	35
4.2	Overview of safety factors and corresponding datasets	37
5.1	Inaccuracy of positioning	42
5.2	Inaccuracy of method	42
5.3	Validation presence of traffic lights 1	42
5.4	Validation presence of traffic lights 2	43
5.5	Coverage of roads included in Traffic Model and Fietstersbond network on two scale levels	45
5.6	Buffer method to match roads	46
6.1	Extreme deviate from the shortest route	51
A.1	Number of observations per class 'road surface quality'	80
A.2	True positive rates per class 'road surface quality'	80
B.1	Extent of the used cycling network during this research	81
C.1	Age distribution National Fietstelweek 2016 (Keypoint, 2016)	82
C.2	Cycling purposes National Fietstelweek 2016 (Keypoint, 2016)	82
D.1	Safety datasets part 1	84
D.2	Safety datasets part 2	85

List of Tables

2.1	Location of cyclists on categorized roads. Source: (Schepers et al., 2017)	13
3.1	Methods used for estimating missing data	28
4.1	Age distribution	33
4.2	Distribution of education level	34
4.3	Fietsersbond safety attributes	38
5.1	Reclassification road types	44
5.2	Classification of routes	48
6.1	Descriptive statistics trip lengths	51
6.2	Significant safety factors per group	55
A.1	Trained models using Decision Tree method	80
C.1	Main cycling motive Fietstelweek 2016 (Municipality of Tilburg), Source: (Keypoint, 2016)	83
C.2	Second cycling motive	83
C.3	Third cycling motive	83
F.1	Descriptive statistics delta route length in meters	87
F.2	Descriptives ratio length	87

Appendices

Machine learning

By using supervised machine learning a model was created using the known data for training the model. To limit over-fitting also cross-validation has been used. The following table shows the performance of the trained models. For instance, the model created by a Fine Decision Tree method was 81.0% correct in predicting the classes.

Tab. A.1.: Trained models using Decision Tree method

Safety factor	Correctly predicted
<i>Road type</i>	81.7 %
<i>Road level</i>	92.1 %
<i>Road surface</i>	84.1 %
<i>Road surface quality</i>	81.0 %
<i>Illumination</i>	89.0 %
<i>Speed</i>	85.0 %
<i>Hindrances</i>	84.9 %



Fig. A.1.: Number of observations per class 'road surface quality'



Fig. A.2.: True positive rates per class 'road surface quality'

The map below provides an overview of the full extent of the cycling network that has been used during this research. In reality, of course, the original cycling network includes the whole of the Netherlands. Yet, since only the southern part of the Netherlands is travelled by the participants in the study area this subset has been chosen for processing purposes. For calculating the missing values for the network level, the extent of the whole network (in grey) is used. To calculate the missing values for the route-level only a portion of this network is used (in red). The network on route-level is a combination of the spatial extent of both the chosen routes and the shortest routes.

Coverage of cycling network versus coverage of route network

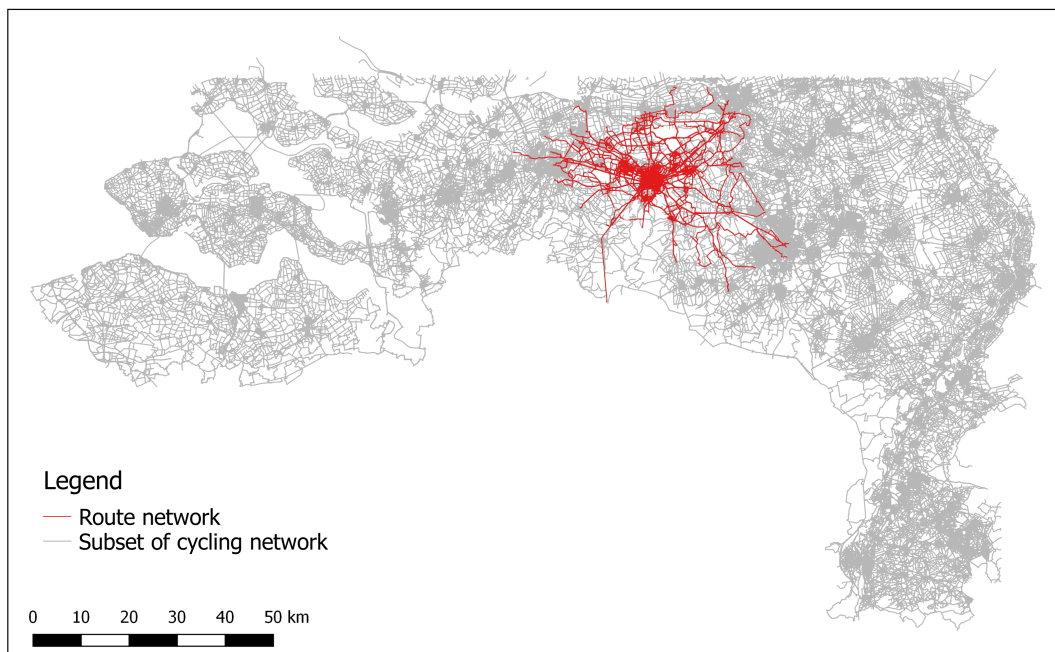


Fig. B.1.: Extent of the used cycling network during this research

C.1 Age and cycling motives on national scale

These figures represent the age and most important cycling motive of all the participants in the national Fietstelweek 2016 that have provided the Fietstelweek organization with their personal information. Since it is also allowed to participate anonymously, it does not take into account all participants but only those who have registered.

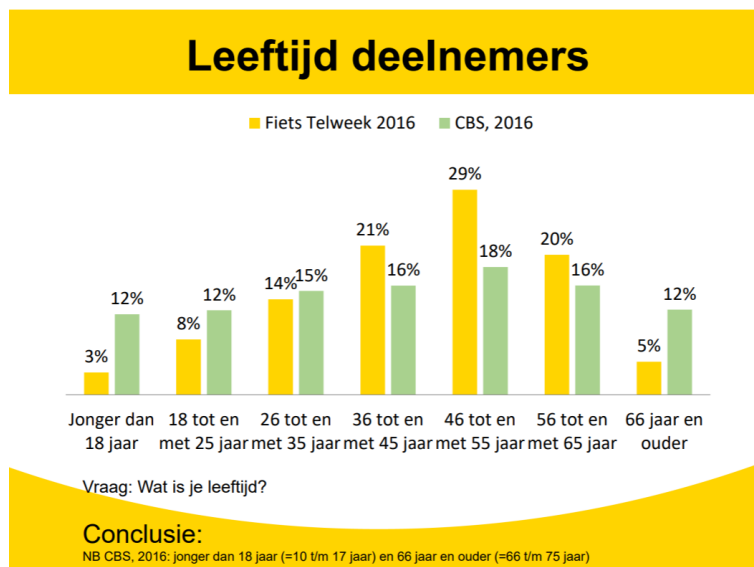


Fig. C.1.: Age distribution National Fietstelweek 2016 (Keypoint, 2016)

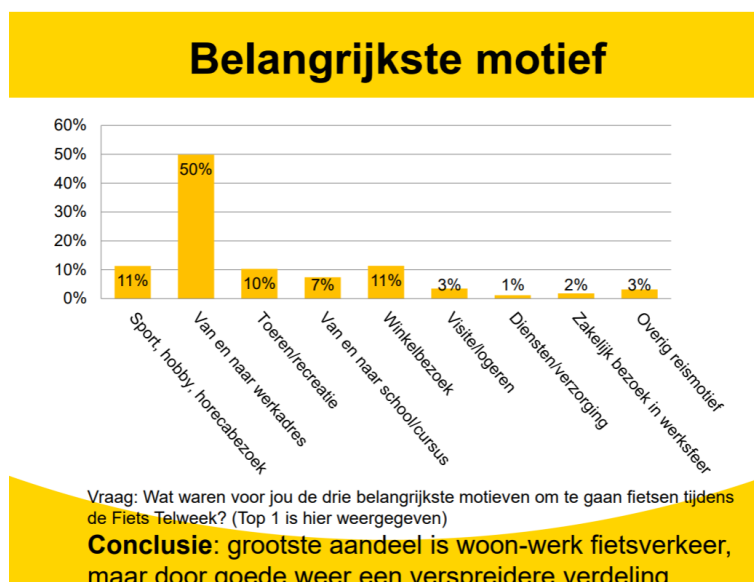


Fig. C.2.: Cycling purposes National Fietstelweek 2016 (Keypoint, 2016)

C.2 Cycling purposes sample Tilburg

In the questionnaire held after the Fietstelweek people were asked to enter their most important (main) cycling motive, followed by their second most important cycling motive and their third most important cycling motive. It seems that only providing 1 cycling motive was mandatory in this questionnaire since the second and third cycling motive have not always been answered by all participants. Again, it is important to mention that this overview does not reflect the cycling motives of the total number of participants in the study area but only of those who have registered with their personal information.

Tab. C.1.: Main cycling motive Fietstelweek 2016 (Municipality of Tilburg), Source: (Keypoint, 2016)

Purpose	Absolute number of people	Percentage of total sample (N = 129)
<i>Eductional</i>	78	60.5 %
<i>Sport/hobbies/cafes</i>	18	14.0 %
<i>Shopping</i>	15	11.6 %
<i>Recreational</i>	13	10.1 %
<i>Visiting people</i>	5	3.9 %
Total	129	100.0 %

Tab. C.2.: Second cycling motive

Purpose	Absolute number of people	Percentage of total sample (N = 129)
<i>Shopping</i>	36	27.9 %
<i>Sport/hobbies/cafe</i>	27	20.9 %
<i>Educational</i>	24	18.6 %
<i>Recreational</i>	11	8.5 %
<i>Visiting people</i>	9	7.0 %
<i>Business purposes</i>	7	5.4 %
<i>Services</i>	3	2.3 %
<i>Other</i>	3	2.3 %
Total	120	93.0 %

Tab. C.3.: Third cycling motive

Purpose	Absolute number of people	Percentage of total sample (N = 129)
<i>Shopping</i>	24	18.6 %
<i>Sport/hobbies/cafe</i>	16	12.4 %
<i>Recreational</i>	13	10.1 %
<i>Commuting to work</i>	12	9.3 %
<i>Visiting people</i>	8	6.2 %
<i>Educational</i>	7	5.4 %
<i>Other</i>	6	4.7 %
<i>Services</i>	1	0.8 %
Total	87	67.4 %

Overview of safety datasets

The tables below provide the used datasets and corresponding values present for each safety factor.


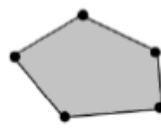
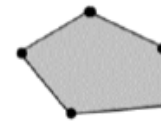




Factor	Subfactor	Elements	Type of data	Source
Accidents	Cycling accidents	Locations of all registered cycling accidents		BRON (2011-2015)
Social safety	Objective social safety	Objective social safety; <ul style="list-style-type: none"> • Number of total crimes per year (per 1000 inhabitants) • Number of violent/sexual crimes per year (per 1000 inhabitants) 		CBS (2015)
	Subjective social safety	Subjective social safety (grade between 1-10) <ul style="list-style-type: none"> • During the day • During the night 		Lemon (2015)
Illumination	Presence of illumination	Degree of illumination; <ul style="list-style-type: none"> • present • partly present • not present 		Fietsersbond (2016)
Road type	Road type	Type of road reclassified to: <ul style="list-style-type: none"> • Mixed traffic with no bicycle dedicated infrastructure • Bicycle boulevard • Bicycle lane • Lane for bicycles and scooters • Bicycle path • Path for bicycles and scooters • Road with bicycle lane suggestion 		Fietsersbond (2016)
Road surface	Road surface	Type of road surface <ul style="list-style-type: none"> • Street tiles • Asphalt/concrete • Unsurfaced • Semi-surfaced • Crushed shell path • Other (uneven, rough roads) 		Fietsersbond (2016)
	Road surface quality	Road surface quality <ul style="list-style-type: none"> • Good • Acceptable • Bad 		Fietsersbond (2016)

Fig. D.1.: Safety datasets part 1








Factor	Subfactor	Elements	Type of data	Source
Intersection & Control	Presence of traffic lights	1. Location of traffic lights 2. Intersections <ul style="list-style-type: none"> Intersection with traffic regulation installation 3. Type of intersections <ul style="list-style-type: none"> Intersection with traffic regulation installation 	1 & 3  2: 	1. OSM (2018) 2. Fietzersbond (2016) 3. Traffic model Tilburg(2015)
	Presence of intersections	1. Type of intersections <ul style="list-style-type: none"> Intersection Intersection with traffic regulation installation Roundabout Does not apply 2. Navigation <ul style="list-style-type: none"> Part of an intersection Part of a roundabout Standard 		Fietzersbond (2016)
Hindrances	Traffic intensity & Parking	Traffic hindrance/obstacles <ul style="list-style-type: none"> Many Acceptable Little 		Fietzersbond (2016)
	Traffic intensity	Heavy traffic <ul style="list-style-type: none"> Ratio between road capacity and traffic load. 		Traffic model Tilburg (2015)
	Road level	Road level <ul style="list-style-type: none"> Next to a busy road 		Fietzersbond (2016)
	Speed limits	Road speed limits <ul style="list-style-type: none"> Slow driving (15 km/h) 30, 40, 50, 60, 70, 80, 100 		Fietzersbond (2016) Traffic model Tilburg (2015)

Fig. D.2.: Safety datasets part 2

During the Fietstelweek 2016 (19/09/2016 - 25/09/2016) the sunrise was between 07:20 and 07:30 and the sunset between 20:00 and 20:15 (JeKoPhoto, 2016). Since the original data only provides the hour (and not the exact time) a person cycled, some assumptions are made. To determine the routes in the dark I wanted to make sure that the person actually cycled in the dark, therefore all routes that were cycling within the hours 0-6 and 20-23 (inclusive). Meaning that the routes between 7 and 19 (inclusive) are considered as trips cycled during daylight. Yet, it is important to mention that when the sun comes up or goes down that it does not happen instantly. This process, called twilight, happens gradually and during the Fietstelweek 2016 it took 33 minutes. Meaning that while interpreting the results of the analysis the following should be taken into account:

- Twilight morning: Starts around 06:45
- Sunrise: Occurred around 07:15
- Twilight evening: Starts around 19:45
- Sunset: Occurred around 20:15

By making these assumptions it means that actually during the morning rush hour there can be some people cycling between 07:00 and 07:15 that are cycling in the twilight still. Also, during the timeframe of 18:00 - 19:59 which is considered *light & non-rush hour* there might be some people cycling in twilight as well.

F.1 Descriptive statistics absolute and relative deviation

Tab. F.1.: Descriptive statistics delta route length in meters

	Abs. deviation (total)	Abs. deviation (group 1)	Abs. deviation (group 2)	Abs. deviation (group 3)	Abs. deviation (group 4)
<i>count</i>	5,049	464	1,220	944	2,421
<i>mean</i>	338	332	247	313	395
<i>standard deviation</i>	1,301	1,231	569	872	1,665
<i>minimum</i>	0	0	0	0	0
<i>Q1</i>	15	12	28	12	14
<i>Q2</i>	96	87	132	82	88
<i>Q3</i>	294	275	274	317	299
<i>maximum</i>	35,585	14,829	14,180	12,659	35,585

Tab. F.2.: Descriptives ratio length

	Rel. deviation (total)	Rel. deviation (group 1)	Rel. deviation (group 2)	Rel. deviation (group 3)	Rel. deviation (group 4)
<i>count</i>	5,049	464	1,220	944	2,421
<i>mean</i>	1.07	1.07	1.05	1.07	1.08
<i>standard deviation</i>	0.16	0.14	0.09	0.13	0.19
<i>minimum</i>	1	1	1	1	1
<i>Q1</i>	1.007	1.007	1.009	1.007	1.006
<i>Q2</i>	1.031	1.031	1.031	1.029	1.032
<i>Q3</i>	1.075	1.075	1.060	1.076	1.085
<i>maximum</i>	4.14	2.29	2.72	2.22	4.14

F.2 Overview significant positive and negative relationships

In this section an overview is provided of the safety factors that significantly influence the deviation from the shortest route alternative are provided. A distinction is made between positive and negative coefficients. A positive coefficient means that people deviate further from the shortest route for more 'safety factor' on their route compared to what the shortest alternative would provide. A negative coefficient means that people deviate further from the shortest route for less 'safety factor' on their route compared to what the shortest alternative would provide.

General analysis

Positive impact on amount of deviation

- Δ Number of accidents
- Δ Number of traffic lights
- Δ Bicycle suggestion lane
- Δ Scooter lane
- Δ Scooter path
- Δ Speed limits
- Δ Asphalt/concrete
- Δ Other/bumpy road surface
- Δ Good road surface quality
- Δ Rel. total number of crimes
- Δ Rel. number of violent/sexual crimes

Negative impact on amount of deviation

- Δ Mixed traffic
- Δ Street tiles
- Δ Semi-surfaced roads
- Δ Unsurfaced roads
- Δ Acceptable road surface quality
- Δ Bad road surface quality
- Δ Total intersections and roundabouts
- Δ Intersections
- Δ Lemon score day
- Δ Lemon score night

Group 1: Dark & non-rush hour

Positive impact on amount of deviation

- Δ Number of accidents
- Δ Number of traffic lights
- Δ Good illumination
- Δ Limited illumination
- Δ Bicycle suggestion lane
- Δ 'Other/bumpy' road surface
- Δ Roundabouts
- Δ Rel. number of violent/sexual crimes
- Δ Rel. total number of crimes

Negative impact on amount of deviation

- Δ No illumination
- Δ Mixed traffic
- Δ Unsurfaced roads

Group 2: Light & morning rush hour

Positive impact on amount of deviation

- Δ Bicycle suggestion lane
- Δ Scooter path
- Δ Scooter lane
- Δ 'Other/bumpy' road surface
- Δ Good road surface quality
- Δ Acceptable hindrances

Negative impact on amount of deviation

- Δ Number of traffic lights
- Δ Mixed traffic
- Δ Semi-surfaced roads
- Δ Unsurfaced roads
- Δ Acceptable road quality
- Δ Bad road surface quality
- Δ Little hindrances
- Δ Lemon day score

Group 3: Light & evening rush hour

Positive impact on amount of deviation

- Δ Number of accidents
- Δ Number of traffic lights
- Δ Bicycle suggestion lane
- Δ Scooter lane
- Δ Asphalt/concrete roads
- Δ 'Other/bumpy' road surface
- Δ Relative total number of crimes
- Δ Many hindrances
- Δ Speed limits

Negative impact on amount of deviation

- Δ Mixed traffic
- Δ Bicycle boulevard
- Δ Bicycle path
- Δ Street tiles
- Δ Unsurfaced roads
- Δ Semi-surfaced roads
- Δ Bad road surface quality
- Δ Little hindrances

Group 4: Light & non-rush hour

Positive impact on amount of deviation

- Δ Number of accidents
- Δ Number of traffic lights
- Δ Bicycle suggestion lane
- Δ Asphalt/concrete roads
- Δ Good road surface quality
- Δ Speed limits
- Δ Rel. number of violent/sexual crimes
- Δ Relative total number of crimes

Negative impact on amount of deviation

- Δ Unsurfaced roads
- Δ Bad road surface quality
- Δ Acceptable road surface quality
- Δ Total intersections & roundabouts
- Δ Intersections
- Δ Lemon day score
- Δ Lemon night score

