GIMA
Geographical Information Management and Applications

Master thesis

Built environment and cycling speed:

Investigating built environment influences on cycling speed in the Netherlands

Lieuwe Arend Visser

Supervisor

dr. Kees Maat

Professor

prof.dr.ir. Peter van Oosterom

Universiteit Utrecht
TU Delft
UNIVERSITY OF TWENTE.
WAGENINGEN UR

Date: 01/03/2019

Acknowledgements

Cycling is part of the Dutch DNA. I was raised in the small town of Houten (the Netherlands), which is a town in the vicinity of Utrecht. Houten has one of the best cycling infrastructures in the world, while car use is being discouraged by infrastructural design. It is fascinating how designs of infrastructure and built environment can actively influence travel behaviour. After six months of hard work I am proud to present my thesis that addresses this very subject: *Investigating built environment influences on cycling speed in the Netherlands.*

This master thesis would not have been finished without the support of a number of people, whom I would like to thank.

Firstly, I would like to thank my supervisor Kees Maat for the amount of hours spent on debating, giving feedback and answering my questions. It was very helpful to have guidance along the process. Furthermore I would like to thank Paul van de Coevering and Dirk Bussche who provided me with a cycling speed dataset. Their aid in answering questions was crucial in getting an understand of the dataset. Also, thanks go out to the Fietsersbond for providing me valuable road network datasets.

Secondly, I am grateful to have friends, family and my partner who supported me along the way. They were listening and supportive in any way. They endured not only my frustrations, but helped me by discussing on subjects that mattered and by giving feedback.

Summary

Cycling speed research is scarce and limited. Especially the emergence of electrical supported bikes is important to consider consequences with respect to higher cycling velocities. While higher cycling speeds lead to higher risks (Schleinitz, Petzoldt, Franke-Bartholdt, Krems, & Gehlert, 2017), there are opportunities as well (Plazier, Weitkamp, & van den Berg, 2017). Previous cycling speed research was mainly focussed on aerodynamics or personal characteristics (El-Geneidy, Krizek, & Iacono, 2007; Schantz, 2017). Two dominant aspects which influence physical activity are infrastructure and built environment (Handy, Boarnet, Ewing, & Killingsworth, 2002; Moudon et al., 2005).

This study aims to find factors of infrastructure and built environment that affect cycling speed. Infrastructure factors which have proven to affect cycling behaviour are: the road surface and road quality, type of road and intersections. The built environment consists of a number of different components. Saelens & Handy (2008) divide built environment into: *land use, transportation system* and *urban design.* Ewing & Cervero (2010) outline *density, diversity* and *design* as three measures of the built environment. These components serve as a basis for the construction of variables in this study.

Four infrastructure variables have been constructed: road quality, road surface type, type of road and traffic lights per km. Two built environment variables have been built: density of addresses and land use type. Additional variables that have been created during this research are: standard deviation of speed and type of bike. Furthermore this study used age and gender as control variables.

This study concluded that built environment affects cycling speed strongly. Built environment variables have overlap with infrastructural variables as they could be seen as a part of the built environment. Infrastructure variables are more detailed and explain only a marginal part of cycling speed each. On the other hand there are other underlying built environment aspects influencing cycling speed that are not part of infrastructure. A final model has been built to express these relationships in one model. Further studies should include built environment factors when researching cycling speed. In addition, it would be interesting to elaborate on these factors.

Content

Chapter 1: Introduction

Contemporary governments, especially in North European countries, try to increase the modal split share of cycling (Krizek, El-Geneidy, & Thompson, 2007; Larsen & El-Geneidy, 2011; Schepers, Twisk, Fishman, Fyhri, & Jensen, 2017). Cycling is seen as an important contributing factor for solving several problems, such as traffic congestion and climate change. Cycling has positive effects on health, reduces car-use and is cheaper than other modes of transportation (Jensen, Rouquier, Ovtracht, & Robardet, 2010; Larsen & El-Geneidy, 2011). Despite these positive effects, there are several reasons people do not make their trips by bicycle. Cycling is less comfortable than using a car, takes physical effort and in a lot of countries cycling is often more dangerous than using the car or public transport. Many studies have been looking into cycling conditions, with the aim of improving them. Cycling is far less researched than other modes of transport and motorized mobility is much better researched than non-motorized mobility (Schantz, 2017). Cycling speed is a motivational factor for commuters to use the bike. In this research cycling speed will be examined within the region of Noord-Brabant in the Netherlands. This study focusses on commuter traffic, because cycling speed is less important for recreational cycling (Moudon et al., 2005).

1.1 Relevance

There has been a 12% increase of cycled kilometres in the Netherlands since 2005 (*Kennisinstituut voor Mobiliteitsbeleid*, 2017). Still only a third of the trips shorter than 7.5 kilometres, were made by bike (Harms & Kansen, 2016). Speed is an important factor determining the efficiency of a transportation system (Jensen et al., 2010). Moreover, the 'constant travel time budget theory' states that people have a certain average time they are willing to spend on travelling. An increase of transportation speed results in an increase of locations that can be reached in the same time (Marchetti, 1994; Metz, 2008). It is possible that more people are willing to cycle if the same destination can be reached quicker by bike. However, the relationship between trip distance, duration and speed is more complex. There are three basic characteristics of mobility in a mode of transport: distance, duration and speed (Schantz, 2017). Duration is the result of trip distance and speed. Distance, duration and speed are constantly influencing each other. However, there are more factors influencing this relationship.

Previous research on cycling speed is scarce and largely limited to subjects like aerodynamics within race cycling and safety issues raised by higher speeds (Schepers et al., 2017). But this is not the focus of this study. The current study aims to give insight into specific factors that affect cycling speed of commuters at *specific locations*, and aims to answer the question why cycling speed varies so much at different locations. The social relevance of the current study is to give suggestions for building new cycle paths that are better focused on cyclists. A considerable effort has already been made in The Netherlands by building cycling highways which are introduced in order to offer better connection between urban areas (Harms & Kansen, 2016). However, there is still little known about the effect of this type of infrastructure.

The introduction of e-bikes and speed pedelecs creates new opportunities for cyclist. However, it also leads to disadvantages for other cyclists (Plazier, Weitkamp, & Van den Berg, 2017). A new type of rapid cyclists is emerging and is making use of the same infrastructure. Nederlandse

Omroep Stichting [NOS] (2019) reported a sales growth of electrical bikes of 40% in the Netherlands. Higher velocity has disadvantages regarding safety. Crashes and accidents have a stronger impact, and crashing risks grow because other road users are not aware of the higher speeds (Schleinitz et al., 2017). On the other hand higher speeds could offer new opportunities for choosing bicycles over cars. Comparable distances can be cycled in a shorter period of time. This implies a decreased travel time for cyclists and thus increasing the opportunity to travel longer distances in the same average time cyclists are willing to travel (Plazier et al., 2017).

1.2 Scientific relevance

Although researching cycling speed is relevant, only a few factors are well studied. A couple of studies refer to cycling speed determinants, mostly pertaining to gender and age. Male gender and younger age have a positive relationship with cycling speed (El-Geneidy et al., 2007; Schantz, 2017). Other factors that show a positive relationship with cycling speed are, health (lower BMI and higher body weight) and trip-duration (Schantz, 2017). Schantz (2017) calculated cycling speed by both measuring travel time and travel distance between origin and destination. Furthermore evidence is shown of a positive relationship between off-road cycle facilities and cycling speed (El-Geneidy et al., 2007). This study only has eight participants however, which means that the generalizability is limited. Another determinant named by El-Geneidy and colleagues is having a certain level comfort while cycling on on-road facilities through traffic. On- and off road facilities are respectively non separated and separated cycling facilities (cycling lane on the road/cycling lane separated from road).

Many cycling behaviour studies are focussed on route choice. Subjects of study were for example, infrastructure, safety, degree of illumination, on/off-road facilities and physical barriers (e.g. train tracks) (Broach, Dill, & Gliebe, 2012; Caulfield, Brick, & McCarthy, 2012; de Vos, 2018; Stigell & Schantz, 2011; Van Genugten & Van Overdijk, 2016). Although not all subjects are considered a relevant determinant of cycling speed, some of these subjects are. Asphalt and concrete slabs are proven to be more comfortable road surfaces than self-binding gravel and cobblestones (Hölzel, Höchtl, & Senner, 2012). Therefore the quality of the road surface might influence cycling speed. Moreover different average cycling speeds between on and off-road facilities were found (Schleinitz et al., 2017). This was a survey study including the use of a Data Acquisition System which measured actual speed and trip distance.

1.3 Literature gap

A factor that will be examined in the current study is the built environment, because it is expected to be a determining factor influencing cycling speed. The built environment of a road has on the one hand a relationship with travel behaviour in general (Ewing & Cervero, 2010). On the other hand there is also a relationship between the built environment and physical activity. Two dominant aspects of built environment emerge as influential on physical activity, namely infrastructure and land use (Handy et al., 2002; Moudon et al., 2005; Saelens & Handy, 2008). Examples of influential aspects are the quality of roads, street patterns and the type of neighbourhood. The direct built environment influence on cycling speed has not been researched yet.

This study focusses on cycling speed measurements using GPS methodologies. Only recently large scale cycling GPS datasets have become available for research. Also, it has become more accurate for correctly measuring physical activity. Using contemporary GPS data is a proper method for measuring physical activity. Before, cycling speed at specific GPS locations is

researched poorly. The current study will examine cycling speed factors using GPS based methodologies. It will include factors that are studied otherwise and factors that were not researchable before due to a lack of GPS data.

1.4 Scope

This research will examine factors that are expected to have influence on cycling speed. Factors explaining cycling speed are not all well researched. Three factors, which are expected to be of influence on cycling speed, are: personal characteristics, infrastructure and built environment. Previous research provides evidence of a relationship between both personal characteristics and cycling speed, and infrastructure and cycling speed. Since a relationship between physical activity and the built environment has been proven, there is reason to believe that built environment is an influential factor regarding cycling speed.

The three factors will be expressed as variables in a statistical model that will be developed as part of this research. The research will be carried out in the province of Noord-Brabant (the Netherlands) for two reasons. Firstly, the Netherlands is one of the countries with the highest shares of bicycle use in the modal split, and therefore eligible for this study. Moreover, the province of Noord-Brabant is at the forefront in promoting cycling. The second reason is the availability of data. This study will use the so-called B-riders dataset which covers the province of Noord-Brabant.

According to Handy et al. (2002) geographical scale is important when researching the built environment. Handy et al. (2002) explains that car trips are more influenced at region scale, while walking trips are more influenced by the characteristic of a neighbourhood. In order to focus better on commuting cyclists the smaller MRE21 (Metropool Regio Eindhoven – Metropolitan Area Eindhoven) is defined as research area. This region has been chosen because of its geographical scale and the commuter interactions between Eindhoven, Helmond and surrounding municipalities.

This research is focused on commuters since cycling is a good alternative for transportation during rush hours. An increase in commuting cyclists would contribute to sustainability of the transportation system during rush hours. It is pre-supposed that cycling behaviour is more dependent on the built environment than other transportation modes (Moudon et al., 2005), since both car and public transport users have, in comparison to cyclists, less sensory interaction with the surrounding built environment.

Factors that are not being discussed within this research are type of cyclist and weather. The used dataset does not provide data on the type of cyclists and is therefor not included. Although weather is seen as an important factor for cycling behaviour in general, it is expected that weather is less relevant with respect to cycling speed. Moreover other research covers weather influences on cycling behaviour.

1.5 Research questions

The research questions are as follows:

In what way do infrastructure and the built environment influence commuting cycling speed, controlled for personal characteristics?

- Sub question 1: *How are personal characteristics related to cycling speed?*

Gender and age are proven to have effect on cycling speed and travel time(Schantz, 2017; Shafizadeh & Niemeier, 1997). Therefor these variables cannot be neglected in this research.

- Sub question 2: *How do infrastructural variables, such as traffic lights, road quality and type of road, affect cycling speed?*

The factors traffic lights, road quality and type of road are known factors to be influencing the route choice of cyclist (Hölzel et al., 2012; Van Genugten & Van Overdijk, 2016). However two of these factors have not yet been tested for their influence on cycling speed. The research of El-Geneidy, Krizek and Iacono (2007) concludes that people cycling on a on road facility ride significant slower than cyclists riding on an off-road facility. Moreover, cyclists that have cycling experience and are comfortable with riding in heavy traffic tend to have higher speeds. However this study used only 8 respondents. Therefor the current study can contribute to broader confidence in determining infrastructural factors that influence cycling speed.

- Sub question 3: *To what extent does the built environment affect cycling speed?*

The operationalization of the built environment is more difficult than other factors. Three elements influence each other: built environment, travel behaviour and physical activity. This sub question will examine the assumption that cycling speed, as part of travel behaviour and physical activity, has a relationship with the built environment too. The scope of the built environment is limited to the following variables: Land use, population density, building density and building proximity. Firstly the proximity of buildings will be measured. Secondly the building density will be calculated per neighbourhood. Thirdly the population density of an area (neighbourhood) will provide better insight of the built environment. Lastly land use data will be reclassified and used to determine the directly surrounding environment of a road.

1.6 Limitations

The outlined research questions represent the scope of this research. However there are relevant cycling speed factors which will not be discussed within this research. Personal characteristics are important to consider, however these will be restricted to gender and age for the reason that it is difficult to further investigate personal characteristics. For example the relationship between health and cycling is already investigated by survey research. However, in The Netherlands is no GPS cycling dataset available that contains health information. Moreover the attractiveness of a bicycle road or path is also important considering bicycle speed. Although a part of the researched factors also play a role in attractiveness, the concept of attractiveness is not researched in this research for the reason that attractiveness is a ‘fuzzy’ concept to define. There is too little data today available to create an attractiveness variable based on GPS data.

1.7 Area of research

The area of study is the MRE21 which is a metropolitan area containing 21 municipalities in the vicinity of Eindhoven. Two large municipalities are Eindhoven and Helmond. In Figure 1 a map is shown of the MRE21 region to visualize the geographical scope of this study.

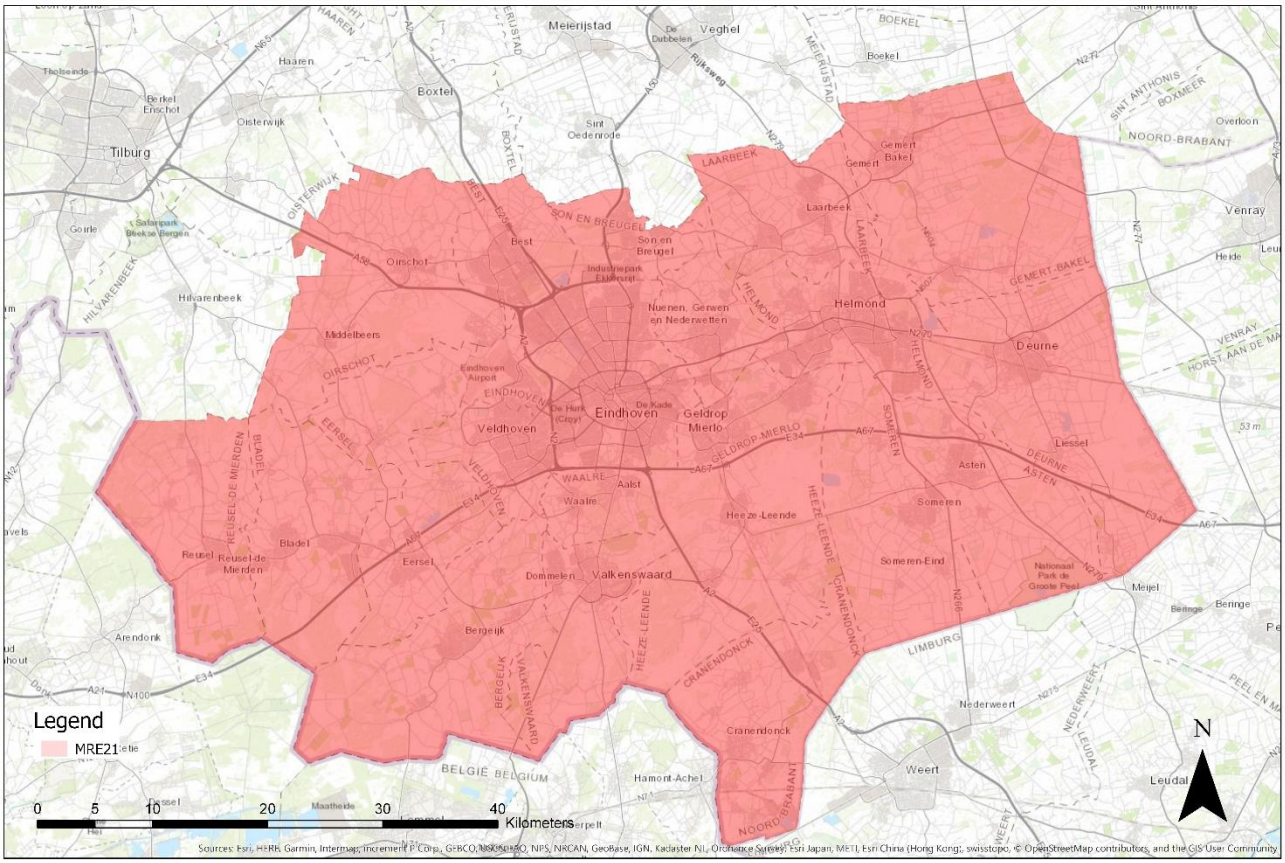

Figure 1: Map of MRE21 region, located in the south-western part of the province of Noord-Brabant (the Netherlands)

Chapter 2: Theoretical framework

In this chapter theory relevant for this study is outlined. The theoretical framework contains constructs, definitions and assumptions made by previous studies. Furthermore it will provide a basis for answering the main research question of how infrastructure and the built environment influence cycling speed. Within the scope of this research there are a couple of terms that are important to outline. This chapter begins with discussing cycling speed. Then the terms infrastructure and built environment will be defined within the scope of this study. Chapter two will be concluded with a conceptual model of relationships considered of importance regarding cycling speed.

2.1 Cycling speed

Cycling speed divergence in daily traffic is growing due to the introduction of e-bikes and speed pedelecs. Cycling speed in previous research is often measured as average trip speed (e.g. Caulfield et al., 2012; Jensen et al., 2010; Schantz, 2017). Survey based techniques are often used by these studies. Only recently GPS-based techniques are used for measuring physical activity. There are some innovative cycling studies that use GPS-based techniques. Examples of these studies are covering cycling safety and route choice (de Vos, 2018), a cycling route choice model (Broach et al., 2012) and e-bike commuters' travel behaviour (Plazier et al., 2017). Although cycling speed is not neglected within these studies, the main aim was not to explain cycling speed. Recent technological improvements with respect to GPS based methodologies this study aims to measure cycling speed by using available road sections. Thus the measured GPS points containing cycling speed information will directly be used in relating cycling speed determinants with the actual cycling speed.

The relation between personal characteristics and physical activity has often been researched. Of these personal characteristics, gender, age and health are seen as most important variables influencing cycling speed (El-Geneidy et al., 2007; Schantz, 2017). As health is out of the scope of this study it will not be discussed furthermore.

Besides personal characteristics, there is reason to believe the direct surrounding environment influences physical activity as well. Firstly the infrastructural factors influence cyclist behaviour and secondly built environment factor have influence on the physical activity on that location. These factors and variables will be outlined in the next sections.

2.2 Infrastructure and cycling speed

An important cycling speed factor is infrastructure. The infrastructure factor consists of several determinants of which is proven they effect cycling speed. The most important determinants are: road surface and the quality of the road, bicycle facility character and intersection control.

<u>Road surface and road surface quality</u>
The road or bicycle path itself is highly explaining cycling behaviour (Hull & O'Holleran, 2014). Road surface and road surface quality are related to cycling comfort (Hölzel et al., 2012). Moreover Hölzel et al. (2012) prove that the type of road surface influences cycling speed, as the amount of rolling resistance determines for a major part the amount of energy needed to cycle at a certain speed. This study outlines six types of surfaces: new asphalt, battered asphalt, new concrete slabs, battered concrete slabs, cobblestones and self-binding gravel. The study of

Hölzel et al. (2012) shows a dependency of surfaces with respect to the velocity. The higher the cycling speed the more the cyclists speed depends on the surface. Moreover, it is showed that cycling speeds are relatively much more influenced by a cobblestone surface than any other surface. The study concludes that there are two 'best' surfaces for bicycle paths: concrete slabs and asphalt. The first one performs better with respect to rolling distance and the second one performs better with respect to cycling comfort (Hölzel et al., 2012).

Besides the type of surface, its quality also influences cycling behaviour, and thus cycling speed. Poor road surface quality has a negative influence on cycling speed (Hull & O'Holleran, 2014). When cyclists need to avoid cracks or gaps in the road, or have a bumpy ride, they are likely to cycle at lower speeds. The first reason is that cyclists want to ride comfortable. Low quality roads effect cycling comfortability more at higher speeds than at lower speeds. Secondly cyclists try to avoid crashes or accidents which could be caused by road damages.

Type of bicycle facility

The second variable of influence is the type of road facility. Literature divides road facilities in three types: non-separated cycling facilities, separated cycling facilities and no facilities at all (cyclist on conventional road). Cyclists try to avoid roads with mixed use. Often longer routes were taken in order to avoid other traffic (De Vos, 2018; Fishman, Washington, & Haworth, 2012; Heinen, van Wee, & Maat, 2010; Winters, Teschke, Grant, Setton, & Brauer, 2010). Caulfield et al. (2012) say that cyclists like to stay within the cycling network and therefore avoid interactions with other road users. Since the type of road influences cycling behaviour it could also be possible that cycling speed is influenced as well. This relationship was proven by a study which used a Virtual Reality environment to test whether type of cycling facility influenced cycling speed. This study concluded that there is significant relationship, although measured in VR environment (Nazemi, Mohsen; van Eggermond, Michael A.B.; Erath, Alexander; Axhausen, 2018).

Intersections and intersection control (traffic lights)

Intersections are the next discussed infrastructural influencers of cycling speed. Cyclists try to avoid intersections and intersection control (Broach et al., 2012; de Vos, 2018; Van Genugten & Van Overdijk, 2016). Firstly intersections come with safety risks, since crossing other traffic increases the chance of accidents. Secondly cyclist are annoyed by being stopped. Especially gaining speed after being stopped, costs physical effort. Moreover in the study of El-Geneidy et al. (2007) intersection control is considered to influence cycling speed within a road segment. However, the measured effect is not significant and the study is only based on eight participants. It is still doubtful whether and how intersection control influences cycling speed at longer road segments. Therefor intersection control is an important variable within this study.

2.3 Built Environment

Definition of built environment

The built environment is a complex concept to define. In general the built environment could be defined as the by human constructed part of the physical environment (Saelens & Handy, 2008). Moreover, Handy et al. (2002) and Saelens & Handy (2008) define the built environment based on three elements:

1. Land use patterns: contains the activity distribution across space alongside with buildings in which the activities take place.
2. Transportation system: contains the physical infrastructure such as roads or other networks.
3. Urban design: contains the ordering and visual characterisation of physical elements.

The first two elements are taken into account within this study. Expected is that speeds outside urban areas will be higher in general. Within built-up areas the number of intersections is higher and it is expected that the traffic load within built-up areas is also higher. Interaction with other traffic causes cyclist to pay attention to other road users, which lowers the cycling speed. Moreover the openness of the landscape might influence cycling speed as well. Cycling through a forest is different from cycling through open fields. For those reasons the land use relationship with cycling speed will be studied. The *transportation system* is the second element which influences cycling speed. The physical infrastructure mainly consists of the roads itself and the cycling network as a whole. The content of this has already been discussed earlier in this chapter. The third element is *urban design. Urban design* focusses more on the appeal of the physical environment. Although this effect influences travel behaviour, it is not expected to influence cycling speed of commuters. Therefore it will not be taken into account within this study.

When studying built environment an important bias should be considered. A National Research Council report (Transportation research board institute of Medicine, 2005) states: *"If researchers do not properly account for the choice of neighbourhood, their empirical results will be biased in the sense that features of the built environment may appear to influence activity more than they in fact do. (Indeed, this single potential source of statistical bias casts doubt on the majority of studies on the topic to date…)" (pp. 134–135).* An example within the subject of this study could be that cyclists participating in this study probably have knowledge about the built environment they cycle in. Cyclist that cycle with high velocities might choose for routes with a high level of comfortability. For example they might take a route which consists mostly of asphalt roads. The influence of asphalt on cycling speed could be overestimated. This effect should be taken into account regarding all built environment variables. Moreover cyclists in the MRE21 region might behave differently in comparison to cyclists in other parts of the Netherlands or in other countries.

Travel behaviour, physical activity and the built environment

When researching built environment in relation with travel behaviour it is important to consider the geographical scale. Handy et al. (2002) provides six dimensions of the built environment with respect to the geographical scale:

1. Density and intensity: amount of activity in a given area
2. Land use mix: proximity of different land uses
3. Street connectivity: directness and availability of alternative routes through the network
4. Street scale: three-dimensional space along a street as bounded by buildings

5. Aesthetic qualities: attractiveness and appeal of a place
6. Regional structure: distribution of activities and transportation facilities across the region

Three dimensions seem relevant within the scope of this research: density and intensity, street scale and street connectivity. The *density and intensity* dimension focusses on population density and density of addresses within a specific area. High number of addresses in a specific area means a high number of origins or destinations from or to that area. It is thus likely that the intensity of trips made in that area will be higher. Furthermore, non-motorized trips are encouraged by short trip distances, which are created by high densities in a mixed-use land use pattern (Klinger, Kenworthy, & Lanzendorf, 2013). This study aims to investigate whether this relationship has effect on cycling speed. The second dimension, *street scale* covers the area that is bounded between street and street surrounding features, such as walls, trees or buildings. The third dimension is *street connectivity*, which covers within the scope of this research the road infrastructure (e.g. intersections).

Travel behaviour is defined as "a trip, which is a movement from one address location to another". It is consists of a number of elements: trip frequency, trip destination, trip length, and mode of travel. An additional element could be the purpose of the trip (e.g. recreational or commuting trip) (Handy et al., 2002). Ewing and Cervero (2001) provide similar elements. In addition they outline derivatives of the previous given elements, namely 'person miles travelled', vehicle miles travelled', and 'vehicle hours travelled'. Within this study the element speed will be researched, which is in theory a product of these elements.

Handy et al. (2002) explain in their article the link between the built environment and travel behaviour. Studies with the aim of explaining this link were mostly focussed on automobile travel behaviour until the year 2000. Nowadays the focus has shifted towards other topics, such as the relationship between physical activity and the built environment . Automobile focussed studies, which discuss the link between the built environment and travel behaviour, can be used in contemporary cycling focussed studies if they have a similar scope. Whereas studies often try to express this relationship within an economical model, this research aims to build a statistical model of cycling behaviour related to the built environment. Handy et al. (2002) conclude that built environment improvements (e.g mixed-use development (mixed use of activities), street connectivity and good design) can improve the pedestrian' or cyclists travel perception.

<u>Density, diversity and design</u>

In the article of Ewing and Cervero (2010) the three D variables are outlined as measures of the built environment. The theory of the three D's (density, diversity and design) was originally conceived by Cervero and Kockelman (1997) in order to describe travel demand. The *density* measure describes a variable per unit of area (Ewing & Cervero, 2010). It means that data is summarized or calculated for each specific area. The second D, *diversity*, aims to summarize the differentiation in land uses in a specific area. A low value represents a single-use environment and a high value represents a varied land use (Ewing & Cervero, 2010). The third D, *design*, consists of the network characteristics within a specified area. This could be a number of intersections, trees alongside a street, pedestrian crosses, etcetera (Ewing & Cervero, 2010). Later on other D's were added: destination accessibility, distance to transit, demand management and demographics (Ewing & Cervero, 2001; Ewing & Cervero, 2010). When comparing the D-measures to the six dimensions of the built environment similarities can be

found. Although elements and definitions are different, underlying theoretical assumptions are equivalent to one another.

2.4 Conceptual model

In Figure 2 the conceptual model is shown. The conceptual model outlines the relationship already explained within this chapter. Important is the dotted line that represents the scope of this study. Although the concepts outside the scope of this study are relevant they are for different reasons not included. Within this model the built environment consists of three parts: network, density and land use. These three concepts are based on the built environment definitions of Ewing and Cervero (2001), Ewing and Cervero (2010), Handy et al. (2002), and Saelens and Handy (2008). Infrastructure comprises the *network* of roads and its characteristics. *Density* represents the built environmental characterization in terms of travel activity. And the third one, *land use*, represents the character of the environment surrounding roads. These three built environment concepts are together with personal characteristics the most important concepts influencing cycling speed.

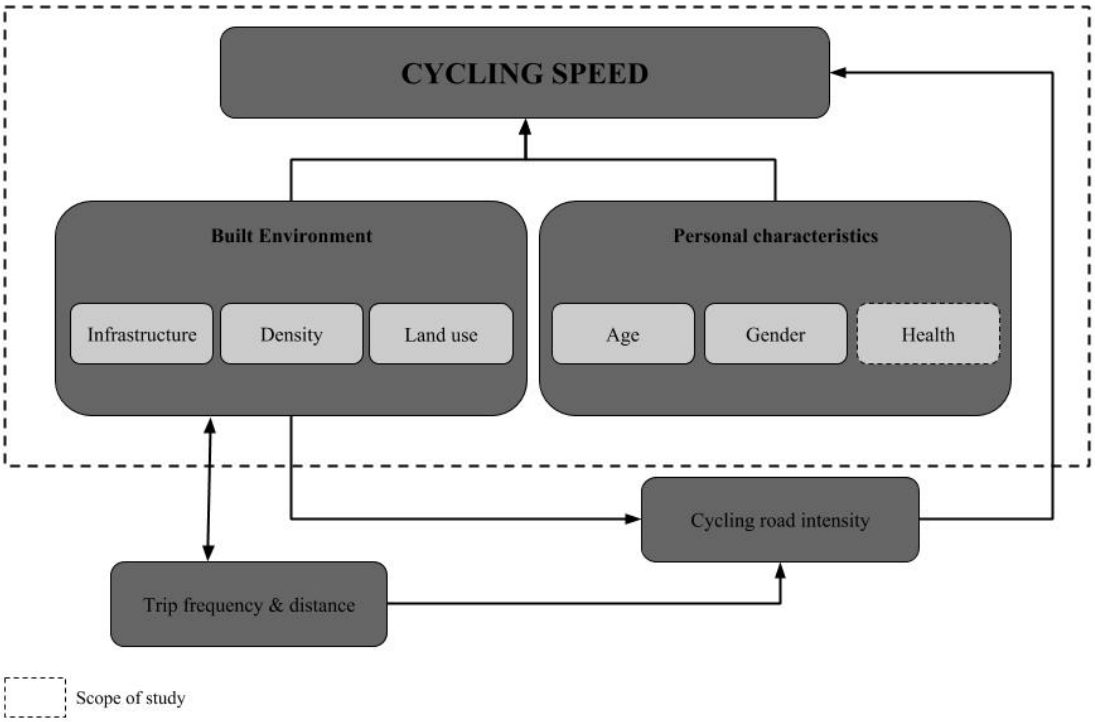

Figure 2: Conceptual model including the scope of this study

Chapter 3: Methodology

3.1 Measuring cyclists travel behaviour

Studying travel behaviour of cyclists is not new. Cyclist travel behaviour is often studied using survey methodologies. Survey questions can be aimed to reconstruct cyclists routes or to investigate cycling experiences. Schantz (2017) researched cycling speed of commuters by conducting surveys. Respondents were asked to report distances, routes, and durations of their trips. Other survey research focussed on the relationship between urban trail systems (cycling facilities) and route choice (Krizek et al., 2007). More recently, research makes use of GPS based methodologies. Until 2000 it was hardly impossible to measure speed of outside activities with GPS, due to technical degradation by US military (El-Geneidy et al., 2007). Nowadays GPS data is proven to be quite accurate and adequate for measuring outside activities (Stigell & Schantz, 2011). Furthermore, cycling behaviour studies are increasingly using GPS data since it is an easy method for acquiring reliable data (Broach et al., 2012; Stigell & Schantz, 2011). Plazier et al. (2017) researched the emergence of e-bikes and speed pedelecs. There is still very little known about travel behaviour of this emerging group of cyclists. The newest methodology trend is to make use of large scale datasets. An example is a study of Jensen et al. (2010), in which datasets of a shared bicycle system were used to map cycling routes and speeds throughout the city of Lyon. The advent of smartphones enabled possibilities to record travel behaviour at larger scale. Respondents can be recruited more easily since participating is more approachable and costs less effort.

3.2 Cycling speed factor by factor

In order to study factors affecting cycling speed several steps have to be taken. The study is broken down to three sub questions, as discussed in the research objectives. These questions embody all factors influencing cycling speed. Each factor contains a number of variables whereof influence on cycling speed is expected. Each variable will be tested for its influence within a framework of hypothesises.

Step-by-step each hypothesis will be tested. Therefor the first step is to assess and edit available datasets before use. Secondly GIS analysis will be performed, which means that datasets are geographically combined in order to compare factors with cycling speed at a specific location. Lastly a statistical model will be built, wherein each hypothesis is tested.

3.3 Workflow

In Figure 3 the workflow of this study is showed. The workflow is divided in three main parts: datasets, data enrichment and analysis/statistics. The left block gives an overview of all datasets needed for this research. The middle block provides information on what information the dataset should contain. The final block provides one with the outcome of synergy between datasets. It will lead to a regression model as a result of this research. The following section explains how each variable will be calculated and what software will be used for these calculations. Although each variable will be summarized shortly, the dataset descriptions will follow up in chapter 4.

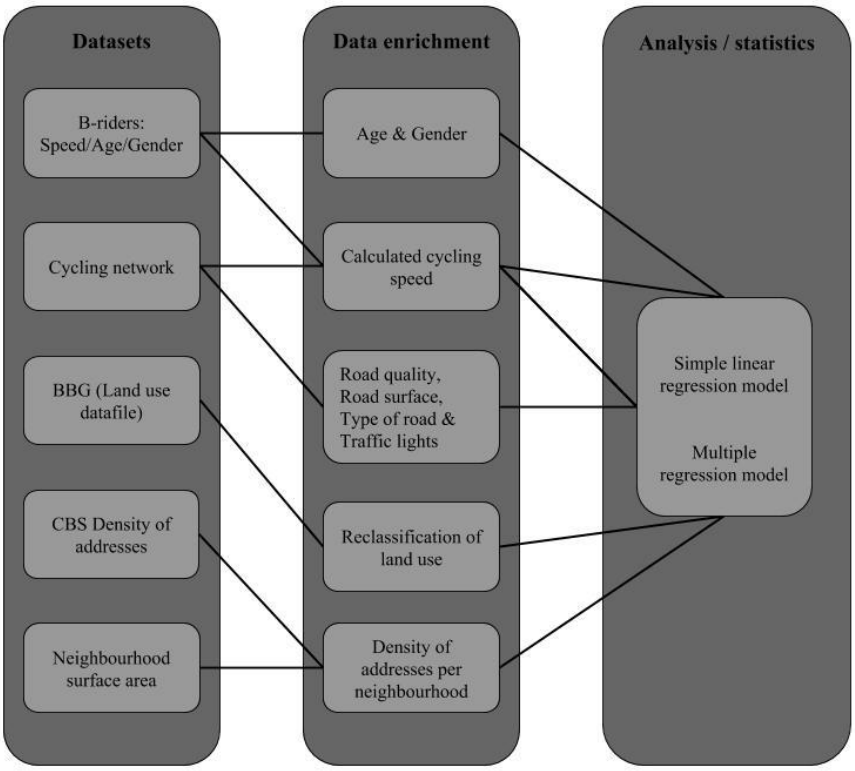

Figure 3: Workflow of data

3.4 Variables

Each variable will be explained step by step. Calculations made in effort of each variable will be supported by an example figure. Figure 4 shows a fictive route that will be used throughout the following sections. The line shows of four line segments with length of respectively 4, 5, 4 and 1 kilometre. The total route length from A to B is thus 14 km.

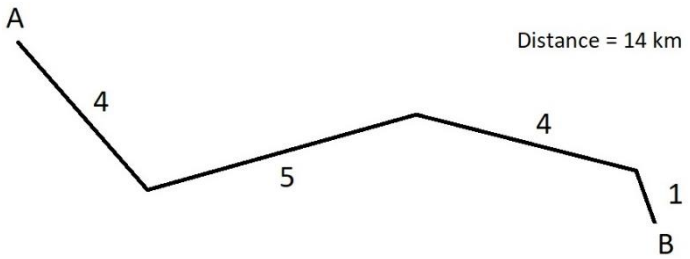

Figure 4:Example route taken by a participant: 4 line segments connecting location A to location B.

3.4.1 Cycling speed

Cycling speed is measured in two ways because calculating average cycling speed can be ambiguous due to wrong measurements in data and wrong assumptions made. The two methods will be explained and discussed accordingly within the result chapter.

Cycling speed method 1:
Method 1 measures cycling speed by recording GPS locations during the trip made by the participant. Each measured GPS point contains speed information. Moreover the recorded GPS points are linked to each individual person as well as to every individual trip made. GPS points are spatially connected to a road network. Each segment (polyline) in the road network can be linked to a part of a route taken by a cyclist.

The first step in calculating cycling speed is to pick one random trip per participant. Since participants did not all cycle the same amount of trips, a random selection will suffice in making sure the sample is sampled randomly. This study aims to investigate behaviour of commuting cyclists. Therefor the sample only contains trips that are made either from home to work or from work towards home. The average cycling speed will be calculated for each trip made by a participant. The well-known formula used is: s (distance) = v (velocity) * t (duration)

In Figure 5 the calculation of average speed is visualized. Firstly the cycling duration per line segment is calculated. Secondly the durations of each line segment is summed. And lastly the total trip distance is divided by the total trip duration, resulting in the average trip speed. This average cycling speed for each trip will be the dependent variable within this research. The

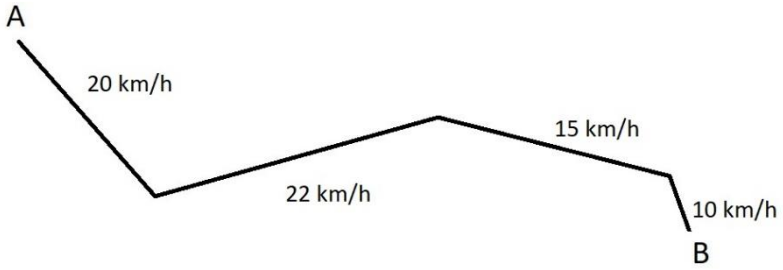

Figure 5:Calculating average speed of one trip of a participant

example Figure 5 shows average speeds per road segment. As visualized in Figure 4 the road segments have distances of respectively 4, 5, 4 and 1 kilometre. The average speed of the total trip is calculated as follows:

$$v_{total\ avg} = \frac{\left(s_{trip\ total}\right)}{\sum\left(\frac{v_{avg\ per\ segment}}{s_{segment}}\right)} \qquad v = 17.63\ km/h = \frac{(4+5+4+1)}{\frac{4}{20}+\frac{5}{22}+\frac{4}{15}+\frac{1}{10}}$$

The main advantage of this method is that speed data already has been matched to road segments in a realistic way. Furthermore, average speeds of trips in total seem to have a realistic variance. Maximum average speeds do not exceed 35 km/h and minimum speeds are not any lower than 10 km/h. The disadvantage of this method is that there are missing values in some parts of trips. More than 65% of all routes cycled by participants have at least one or more missing values. On the other hand the average percentage of segments with missing speed values is only 2.72%. The average total number of segments per trip is 117 segments (only

routes containing at least one missing value). This suggests that on average only 3 segments of these 117 segments are missing. When analysing where these missing segments did occur there was not one answer to give. Missing values are caused by limitations of the measuring method used. GPS locations were measured every 30 seconds. While map matching GPS points to a network the whole route is taken into account. However, there are cases of road segments that could not directly linked to a GPS point. (e.g. three short segments which have only one nearby GPS point). Whenever that happened a cycling speed of 0 was measured. This does not mean that the actually cycled slowly or not at all. As segments vary in length it is difficult to say how much impact missing values exactly have within the sample. Moreover the location of missing values occur at different locations. Sometimes missing values occur at the beginning of a trip, sometime on longer segment parts in the middle of a trip and sometimes missing values could be found at very short road segments (often part of intersections). Since there is a big variance in location occurrence and segment length, the choice has been made to leave missing values out of the speed calculation. In that respect the assumption is made that missing values have the same value as the total average speed of the trip (calculated by using all road segments that had no missing speed value). Because method 1 raises some questions and thoughts on speed calculation a second method is used to calibrate the eventual results.

Cycling speed method 2:

The second method measures cycling speed by using timestamps created within the B-riders dataset. Unlike method 1, which uses pre research map matched speeds, the second method tries to avoid these pre-fabricated speeds. The B-Riders dataset contains raw GPS data points as well. The raw points are locations measured by an application on participants' phone. Each point gives thus an x and y location and a corresponding speed. These points are measured at specific moments, which give additional time information. Each time a point is measured an associated timestamp is recorded as well. This second method find the minimum and maximum timestamp of each route within the sample. The minimum and maximum timestamp provide information of total trip duration of each participant. Total trip distances are known as well. GPS points that are matched to a road network give a specific route through a network. The total distance of all trip segments can be calculated. The speed formula can thus be used again: s = v * t.

Trip duration:	Start (minimum)	End (maximum)
	1-1-2014 12:00:00	1-1-2014 12:47:00

Δ duration = 47 minutes = 0.7766 hours
Distance = 4+5+4+1 = 14 kilometres
Speed = 18 km/h

$$v = 18\ km/h = \frac{(4 + 5 + 4 + 1)}{0.7766}$$

The advantage of this method is that there are no missing values with respect to speed. It seems a more reliable way of measuring average speed than method 1, which includes estimations

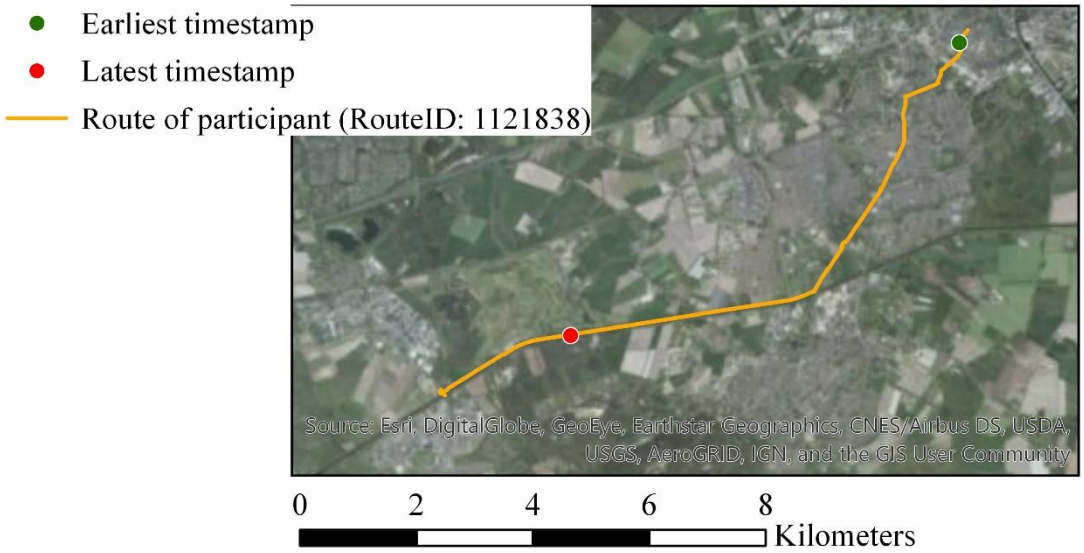

Figure 6: Discrepancy of location between measured starting point and end point and the actual cycled route. The starting and ending points contain timestamps.

along routes. Despite method 2 seems reliable it is not. Timestamps do not exactly overlap with the route taken by a participant, as shown in the extreme example in Figure 6. The earliest (starting point) timestamp of the route and the latest (ending point) timestamp give the maximum trip duration of a participant. The discrepancy between the route and measured points differs among different routes. This implicates that speeds measured with this method are difficult to interpret. Especially at shorter trip distances.

Moreover timestamps have only been recorded in minutes and not in seconds. The shorter the total route distance the larger the negative effect of a wrongly measured timestamp. This effect is shown in Table 1. Table 1 shows the margin of error related to rounding trip durations to minutes. The speeds in bold show the big difference between average cycling speeds at shorter trip distances. This study assumes that differences of 5.5 km/h or more are too big. Differences of 2.6 km/h seem more acceptable considering an average speed of 20 km/h in the total sample.

Table 1:Difference in speed calculations due to imprecise time data (example)

Trip duration (min)	**Trip distance (km)**	**Avg. speed (km/h)**
1.5	0.5	20
1.1	0.5	27.3
1.8	0.5	16.7
3	1	20
3.4	1	17.6
2.6	1	23.1
6	2	20
6.4	2	18.8
5.6	2	21.4

Short routes are thus much more influenced by measuring errors than routes with larger total distances. For that reason every route with a shorter total distance than 2000 meters has been excluded out of the sample.

Speed pedelec, e-bike and cycling speed

In order to get insight in what realistic speeds are a new variable is calculated. The new variable aims to explain cycling speed better. A separation in the sample data has been made between speed pedelec users and e-bike users.

By separating the type of bike results will change. Since different bike types give a different level of support to cyclists it is important to consider these differences when looking into average speeds. In general e-bikes in the Netherlands only support up to the maximum speed of 25 km/h. Higher speeds are possible, however only by physical effort. Speed pedelecs have a speed limit of 45km/h for supporting cyclists, however in practice speeds up to 50 km/h can be reached. Nevertheless, it can be assumed that higher speeds than 50 km/h will not be reached. Although the aim of this study is not to find differences between e-bikes and speed pedelecs the research sample has been split up accordingly. It is assumed that cycling behaviour differs between users of both groups. Higher speeds lead for example to higher risks of severe injuries when crashing (Schleinitz et al., 2017). Although the B-Riders dataset did not contain any information on what type of bike was used to make a trip, this information could be derived from maximum speeds within the trips. Each segment of a trip contains an average speed of that segment. By looking at the maximum speed of each trip assumptions could be made on what type of bike had been used. A threshold value of 27 km/h has been used in order to separate e-bike users from speed pedelec users. If a cyclist has not exceeded the average speed on any segment with 27 km/h or higher, the cyclist is considered a e-bike user. If in any segment of a trip a higher speed (>27 km/h) is measured the cyclist is considered a speed pedelec user. Since there is an assumed difference in cycling speed behaviour, the groups have been analysed as well separately as well as together. Figure 7 shows two histograms of the frequencies of speed values. Splitting the dataset between e-bike and speed pedelecs results for both groups in a relative normal distribution.

Figure 7: Two histograms showing the distribution of average speed values for both e-bike users and speed pedelec users.

Although the criteria distinguishing e-bike users from speed pedelec users can be set, the results cannot be 100% certain about a participant being an e-bike user. This method does correctly sample the speed pedelec users (in the sense that the speed pedelec group only consist of real speed pedelec users, however the sample could lack speed pedelec users that are cycling at lower speed than 27 km/h). As long as the maximum speeds are correct, you can assume that speed pedelec users are correctly classified. However the group of e-bike users could contain participants that are in fact speed pedelec users. Speed pedelec users who did not exceed a

maximum speed of 27 km/h are wrongly classified as e-bike user. Despite this ambiguity most of cases within the sample will be correct. Moreover this ambiguity is of most importance while drawing conclusions about differences between those two groups. Wrongly classified speed pedelec users behave (while explaining cycling speed) probably more or less similarly to e-bike users while riding at similar speeds. Within this study is assumed that this group will be affected minorly because of this ambiguity.

Standard deviation of cycling speed per participant

Another cycling speed variable is created to get insight in the different speeds within one trip. For each trip/participant a standard deviation of cycling speeds is calculated. The variable gives information about how much variance occurs within one trip. When a cyclist varies a lot in cycling speed the standard deviation will be high. When a cyclist cycles more or less at a similar speed throughout his trip the standard deviation will be low. Cyclists with a high standard deviation passed more traffic lights (high correlation between traffic lights/km), probably had to make more turns or came across something else that influenced their cycling speed. This measure is especially interesting when comparing different groups within the sample. For example when comparing gender or type of bike. Moreover it could be interesting to find out whether there is a relationship between the standard deviation and average speed of trips.

3.4.2 Road infrastructure

The road infrastructure consists of four variables: road quality, road surface, type of road and a traffic light ratio. The variables road quality, road surface and type of road are difficult to express within a statistical model. Firstly these variables have an ordinal or nominal scale. Secondly for each trip a ratio per value should be calculated since this research aims to explain cycling speed through locational factors. Therefore one value (aspect) of each variable will be chosen to construct this ratio measure.

Furthermore road infrastructure variables need to be extracted from a road network dataset. The road network dataset does not exactly geographically match with the route cycled by a participant. In order to select the correct road segments of the network dataset a buffer operation is used. The buffer is a flat buffer with a buffer distance of 0.3 meters. Figure 8 shows the buffer operation. Network line segments are selected if the centroid of the segments is within the buffer.

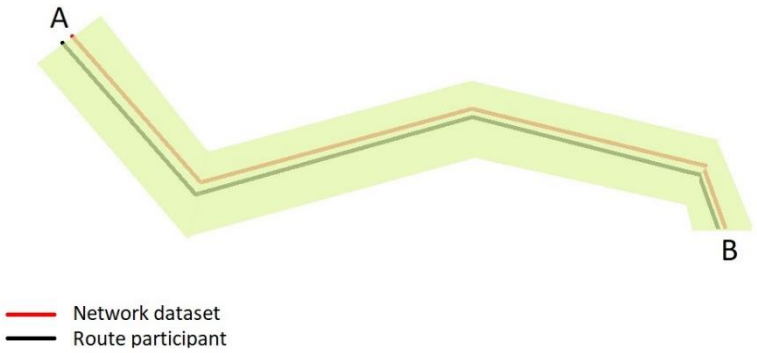

Figure 8: Buffer operation for selecting network road segments that are aligned with the participants route.

The selection of line segments which represent the route on the road network dataset will be exported as a new table. For each route a table is created. Each table consists of segments that are part of the route taken by a participant. Each segment contains information on infrastructure variables: road quality, road surface and type of road. Moreover the length of each segment is stored in the table as well. Further calculations will be performed in Python software. As will be explained in this chapter, each variable will be expressed in a percentage. In general the length of a variable value has been added up for each route and divided by the total distance of each route. In Table 2 an example calculation is performed. Each variable will be explained accordingly within this chapter as well.

Table 2: Example of calculating infrastructure variables for participant X

Segment	**Road surface (reclassified)**	**Road quality**	**Length (km)**
1	Asphalt (Smooth)	High	4
2	Cobblestones (Not smooth)	Low	5
3	Concrete slabs (Smooth)	Medium	4
4	Unpaved road (Not smooth)	High	1
Total percentage	Smooth: 8 / 14 * 100 = 57% Not smooth: 6 / 14 *100 = 43%	High: 4 + 1 / 14 *100 = 36% Low: 5 / 14 *100 = 36%	14

Road quality

As stated by Hölzel et al. (2012) cycling speed is lower on battered roads. The dataset of the Fietsersbond (cycling union) contains road quality. The road quality in this dataset is measured in ordinal scale. There are three values for road quality defined: high, medium and low. As there are ordinal values only the variable is reconstructed. The values ‘high’ and ‘low’ are expected to be most influential with respect to cycling speed. Both values will be tested separately within this study. Eventually the most reliable predictor will be expressed in the road quality variable as a part of the final statistical model. Each value will be expressed as a percentage of the total trip made by a respondent. For example, a trip distance from location A to B is 14 km, and 7 km of the trip has been made on a road qualified as ‘low’, the road quality value will be 50%. The qualification ‘high’ will be assessed in the same way. In Figure 9 examples are provided of high quality and low quality roads.

Figure 9: Reference pictures: High quality (left) roads and low quality roads (right)

Road surface

The dataset of the Fietsersbond also contains information on road surfaces. This variable is also measured in ordinal scale. Although the Fietsersbond dataset provides more than three values, this research reclassifies all values into three values (outlined in chapter 4). These values are: Asphalt, cobblestones and unpaved. As discussed by Hölzel et al. (2012) asphalt and concrete slabs provide the highest comfort level and cobblestones the least. It is expected that less comfortable trips have a lower average speed. The most important distinction could be made between smooth road surfaces and less smooth road surfaces. As is expected that average speeds will be higher at smooth road surfaces, a percentage of smooth road surfaces will be calculated. Thus, if 9 km out of a 14 km trip has been made on a smooth road surface the variable value for that respondent will be 64%. Figure 10 gives reference pictures of 'smooth' surfaces and surfaces that are considered 'not smooth'.

Figure 10: Reference pictures: smooth surface (left) and unsmooth surface (right)

Type of cycling facility

The type of road variable is also derived from the Fietsersbond dataset. This variable is originally measured in ordinal scale. The dataset of the Fietsersbond contains a list of types of facilities. In Figure 11 the most important types of facilities are outlined. The statistical analysis of this study will make use of two ratio variables in the end. Every facility will be classified either as 'separated cycling facility' (no interaction with other road users) or as 'non separated facility' (interaction with other road users, such as car drivers). The reclassification containing all types of facilities is shown in

Table 6 in chapter 4. The variable will be expressed into a percentage of the route either separated or not separated for each participant.

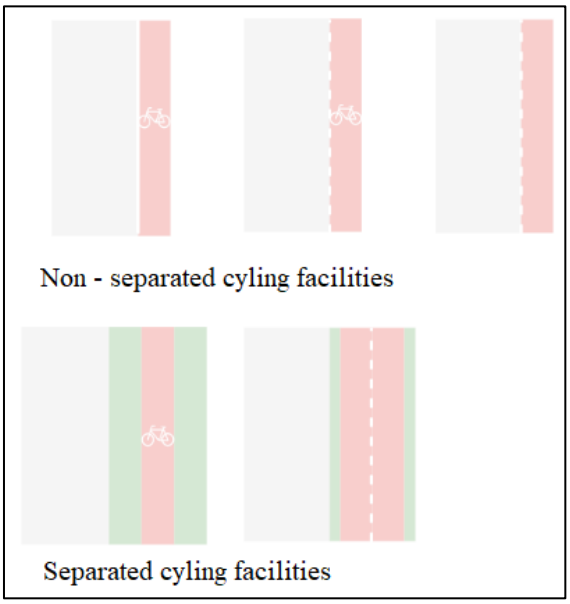

Figure 11: Types of cycling facilities and the difference between separated and non - separated facilities; adapted from De Vos (2018).

Intersection control (intersections with traffic lights)

Intersection control, which is defined as an intersection controlled by traffic lights, is a ratio variable. Intersections with traffic lights slow cyclists down as they are obliged to stop for crossing traffic. Moreover physical effort is needed to gain speed again. The traffic light dataset will be merged with the network dataset. Each route a participant has taken contains a number of traffic lights he/she had to cross. Since cycled distances vary between participants the number of crossed traffic lights will be divided by the total trip distance. The result is a ratio of traffic lights per kilometre. This number will be used in finding a relationship between intersection control and cycling speed. An example of a calculation is shown in Figure 12.

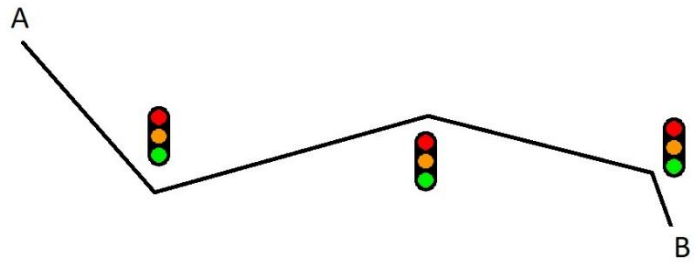

Figure 12: Calculation number of traffic lights:

$$number\ of\ traffic\ lights\ per\ kilometre = \frac{number\ of\ traffic\ lights}{total\ trip\ distance\ (km)}$$

$$0{,}214 \text{ traffic lights per km} = \frac{3}{14}$$

A side note for this variable is that within the MRE21 area (area of study) 580 of the 632 traffic lights are located within built-up areas (as classified within this research). Moreover most traffic lights that are outside urban areas are located near edges of built-up areas (631 traffic lights within 250 meters of built-up area). Therefor it is important to consider possible correlations between this variable and other variables, such as the built-up area and density of addresses.

3.4.3 Built environment

Density of addresses

The first built environment variable is the density of addresses [DoA]. This variable is chosen because the density of addresses gives an indication of locations of origins and destinations. Although this study only focusses on commuting, the number of addresses can provide valuable information on how crowded a location could possibly be. As well Ewing and Cervero (2010) as well as Handy et al. (2002) state that density is a correct parameter when studying travel behaviour. The density of addresses is a dataset derived from CBS (2014). It contains the average number of addresses that are situated within a radius of 1 km. CBS (2014) aggregated this information into neighbourhoods. This study uses the neighbourhood scale. Figure 13 and Table 3 explain the use of the density of addresses within this research.

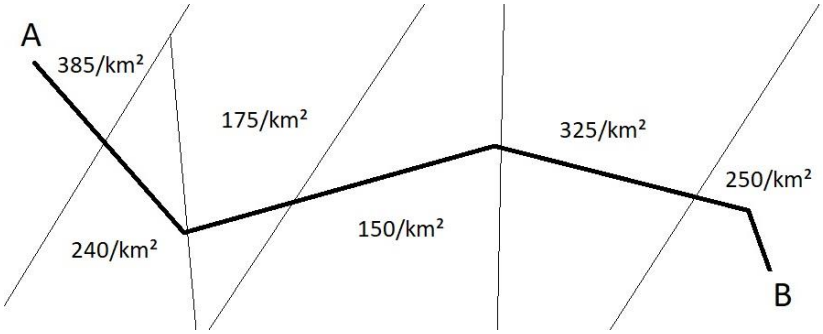

Figure 13: Example calculation of density of addresses variable. For each route an average density of addresses is calculated.

$$Average\ density\ of\ address(avg.DoA) = \frac{\sum(DoA * segment\ distance)}{total\ trip\ distance}$$

Table 3: Example Arcgis Pro output table

Segments	**Segments cut by DoA Area (Intersection)**	**DoA/km²**	**Original Segment distance**	**Cut segment distance (intersection)**
1	1a	385	4	2
	1b	240		2
2	2a	175	5	2
	2b	150		3
3	3a	325	4	3
	3b	250		1
4	4	250	1	1

$$Avg.DoA = \frac{(385 * 2 + 240 * 2 + 175 * 2 + 150 * 3 + 325 * 3 + 250 * 1 + 250 * 1)}{(2 + 2 + 2 + 3 + 3 + 1 + 1)} = 251{,}79$$

Firstly the distance of segments through each density of addresses has to be determined. The intersection tool in Arcgis Pro calculates the distance in which one feature is overlapped by another one. The tool results in a table, as is shown in Table 3, containing information about density of addresses for every segment of the route cycled by a participant. By dividing the sum of density of addresses per segment by the total length, an average DoA for the total trip is calculated.

Land use

The land use dataset consists of various land uses which will be reclassified into three types: built-up area, open landscape (agriculture), and less open landscape such as (forests) (see dataset chapter 4). Since the average land use of a trip cannot be calculated three variables will be derived from three land use types. For example the percentage of urban area within trip A to B will be calculated. This results in three variables containing percentages for each environment per trip per participant.

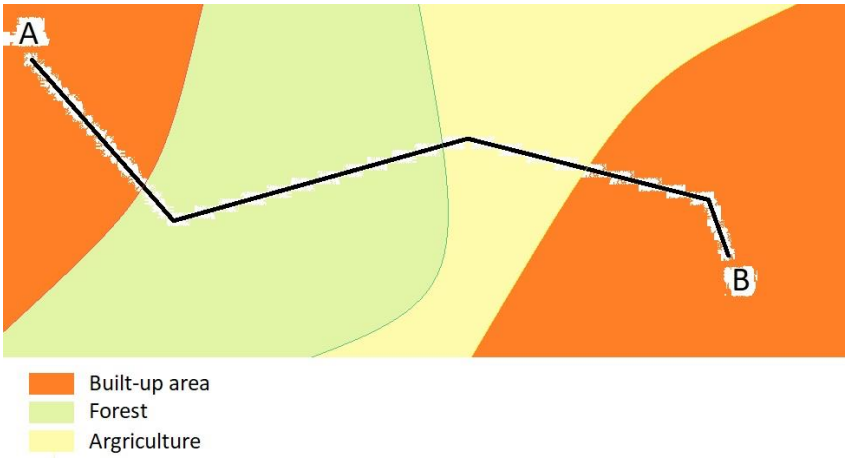

Figure 14: Example calculation of land use variables. For each route a percentage of total trip distance through each land use type is calculated.

$$Percentage\ land\ use = \frac{total\ distance\ through\ land\ use\ type}{total\ trip\ distance} * 100$$

As is shown in Figure 14 the percentage of land use can be calculated with the intersection tool of ArcGIS Pro. The procedure of the DoA variable is similar for this variable. After calculating intersection distances for each land use type the results will be exported into a table. The only difference with example

Table 3 is the 'DoA/km^2' column which should be replaced with land use type. The values in the table would be either, 'Built-up', 'Forest' or 'Agriculture'.

3.5 Statistics

After analysing and combining datasets, statistical tests are necessary to prove suggested relationships. A statistical model will be built that tries to explain cycling speed/velocity through the researched variables.

Hypotheses

The statistical method that fits best to build this model is multiple regression, as speed is measured on ratio scale. However, speed influencing factors may not be completely linear. Therefore variables will be studied one by one, before building a statistical model. Hypotheses for each variable are:

1. Personal characteristics:
   - $H1_A$: Males tend to cycle at higher average speeds than women.
   - .
   - $H2_A$: The younger the participant, the higher the cycling speed.
   - $H3_A$: Speed pedelec users have a higher average cycling speed than e-bike users. users.
2. Infrastructure:
   - $H4a_A$: The higher the percentage of high quality roads, the higher the average speed.
   - $H4b_A$: The higher the percentage of low quality roads, the lower the average speed.
   - $H5_A$: The higher the density of traffic lights, the lower the average cycling speed.
   - $H6a_A$: Higher percentages of separated cycling facilities lead to higher average cycling speeds.
   - $H6b_A$: Higher percentages cycling on non-separated facilities lead to lower average cycling speeds.
   - $H7_A$: Higher percentages of cycling on surfaces that are not smooth lead to lower average cycling speeds.
   - $H8_A$: Higher percentages of cycling on smooth surfaces lead to higher cycling speeds.
3. Built environment:
   - $H9_A$: Higher percentages of forest environment lead to higher average speeds.
   - $H10_A$: A higher density of addresses leads to lower average cycling speeds.
   - $H11_A$: Higher percentages of urban areas lead to lower average cycling speeds.
   - $H12_A$: Higher percentages of agricultural area lead to higher average speeds.

Independent samples t-test

The independent samples t-test will be used in order to find differences between means across groups within the dataset. Hypothesis 1 and 3 will be tested in this way. Because gender and type of bike are nominal variables the t-test is used to test whether their means of average cycling speed differ significant. When tested significant the direction of this relationship is examined. The hypotheses already suggest a relationship, nevertheless a two-sided test will be performed.

Linear regression

Linear regression will be used for testing the other hypotheses. A simple model will be created for every variable within the hypotheses. The simple model will provide insight in the relationship between independent variables and the dependent average speed variable. Before a linear model can be build the following requirements need to comply to general rules (Field, 2013):

1. Variables are preferably measured on interval/ratio scale.
2. Values are measured independent of each other.
3. The relationship between variables have to be theoretically causal.
4. The relationship is linear:
   - In order to determine the degree of linearity scatterplot of all independent variables will be created. Average cycling speed is influenced by a lot of factors. For that reason the average speed data contains a lot of noise. Nevertheless linear relationship can be found.

5. Residuals are equally divided (homoscedasticity):
    - The residuals of a statistical model should vary randomly. When the residuals change by following a pattern (e.g. growing residuals when the dependent variable gets a higher value) the model is heteroscedastic. When this assumption is violated the significance value of your model is not reliable, which could lead to a type 1 error.
6. Multicollinearity ($r \geq 0.9$; Multiple Regression only):
    - In order to understand multicollinearity each variable will be tested within hypotheses. Checking for multicollinearity possibly leads to variables that have overlap within the explanation of the independent variable. Firstly this will be tested by looking at the correlation matrix of variables. Furthermore it will be tested for in by using the VIF factor as is recommend by Braun & Oswald (2011) and Field (2013).

Multiple linear regression

Multiple linear regression analysis will be performed as a final product of this study. The model should explain the average cycling speed by using a number of independent variables. The focus lies on infrastructure and built environment since these subjects have not been researched well before. Since there are variables that have not been studied before in relation to cycling speed regarding methodologies used in this study, it is important to outline the methodology of the multiple regression model. Linear regression and exploratory research could give some problems when analysing the impact of independent variables (Braun & Oswald, 2011). Firstly one has to be sure of the linearity and secondly one should be aware it is not sure whether a linear model fits the data the best. Braun and Oswald (2011) and Field (2013) discuss multiple linear regression methodologies quite well. A common used way of comparing variable effects within multiple linear regression is making use of the Beta. In order to describe studied phenomena it is good to get insight in the relative importance of predictors of a regression model (Braun & Oswald, 2011). Additionally these authors state that determining relative importance of predictors is always ambiguous when predictor variables are correlating. Indices can be computed in order to reflect on the relative importance. Braun and Oswald (2011) provide three indices: Incremental R^2, General dominance weights and Relative importance weights. Field (2013) uses incremental R^2 to validate predicator explanation of the dependent variable. In SPSS software it is relatively easy to use the incremental R^2 method. Hence this method will be used within the multiple regression analysis of this study.

For some part this research consists of some exploratory elements. Although it seems logical that built environment and infrastructure influence cycling speed is has not been studied often. To put it even more strongly, variables such as density of addresses and land use have never been research before using similar methodologies and amounts of data. However some variables have been researched using other methodologies such as road surface type. There are several ways of entering predictors to one's multiple regression model (Field, 2013). Although stepwise methods in combination with automatic linear modelling seem very helpful when trying to find the best fitting model when studying exploratory predictors, there are too much downsides of using this method. Model estimations are not reliable because it results in missing predictors that are significant and more important it results in too high estimates of significance and R^2 values. Because this study has some exploratory elements the enter method is not ideal as well. Using the hierarchical method is a compromise between the two methods. In several steps predictors are added to the model, starting with the literature based predictors and ending with more exploratory predictors.

Chapter 4: Datasets and pre-processing

4.1 Introduction

Several datasets are necessary in order to execute this research. Different datasets are acquired with the aim of constructing variables that are outlined in previous chapters. Data of speed, personal characteristics of respondents, a road network and the built environment variables are essential. This chapter outlines the contents and the way the datasets will be used.

4.2 Cyclists travel behaviour dataset

Cycling speed is derived from GPS-based tracking data of cycling behaviour. This dataset contains cycling speed as well as personal characteristics of participants. By recording GPS stamps during the route a cyclist rides, the speeds can be calculated at specific locations. The dataset of B-riders is a datasets that meets those demands. B-riders data is data that originates from a project that stimulates people to use the bike as an alternative for the car. Participants earn points according to the distance they rode with their bike. Moreover during rush hour they get bonus points, in order to encourage commuters to use the bike. The more points the participant gains, the higher the chance of winning a price. Although the dataset provides a huge amount of data (45 million points) about/with, specific route and speed information, the dataset is biased. Participants are not randomly recruited and are being stimulated to use the bike. Therefor this study results should be reviewed with caution. The dataset covers a part of the Noord-Brabant province of the Netherlands. The B-riders dataset contains a set of files, which are summed in Table 4. Table 4 provides information on the contents of each particular file.

Table 4: Contents of B-Riders dataset

File name	Content	Linkage	Topology
gpspunten.csv	This is a csv file containing as well as the GPS location as the speed at that location per participant and trip. Moreover timestamps for each record are included.	USER_ID ROUTE_ID	None-Table
gpspunten.shp	This is a shapefile containing as well as the GPS location as the speed at that location per participant and trip. Moreover timestamps for each record are included.	USER_ID ROUTE_ID	Points
Links.shp	This is a shapefile containing a network of roads which have been used by participants. Each road segment has an unique number which can be linked to trips that have been made.	LINKNUMMER	Polylines
Gebruiker-leeftijd-geslacht.csv	This csv file contains personal characteristics of participants, namely age and gender.	USER_ID	None-Table

Gps-match.csv	This csv file contains speed, time and heading of each road segment within a trip made by a participant.	ROUTE_ID LINKNUMMER	None-Table
User_routes.csv	This csv file contains a list of every trip being made by bike per participant.	ROUTE_ID USER_ID	None-Table

As shown in Table 4 all files in the provided dataset can be linked to one another. Routes that have been cycled by participants can easily be linked to a network, since the linkage is based on a unique number. Each road in the network is given a unique id: LINKNUMMER. Each part of a route has a unique id for the route itself (ROUTE_ID) and which road was used (LINKNUMMER). Furthermore is each unique route (ROUTE_ID) only cycled by one participant (USER_ID).

Determination of routes

The gps-match.csv file contains the speed per LINKNUMMER (road segment) per participant. Table 5 shows an example of the gps-match.csv file. The LINKNUMMER column can be joined to a shapefile containing geographical locations of LINKNUMMER (road segments). Links.shp, is a shapefile consisting of a network of segments. Each segment of the network file has an attribute LINKNUMMER. Because every segment of the route of each participant has a unique LINKNUMMER the routes can be derived by looking at the join between Links.shp and gps-match.

Table 5: Example table of route determination (gps-match.csv)

ROUTE_ID	**SPEED**	**LINKNUMMER**	**SEQUENCE**
001	25	101	1
001	19	102	2
001	22	103	3
002	18	100	1
002	22	102	2
002	19	104	3

4.3 Network dataset

The second type of dataset that will be used is the cyclist network dataset of the Fietsersbond. This dataset consists of all digitized roads accessible for cyclists, including bicycle paths. The roads are stored in vector format. This dataset contains attributes such as road quality, road surface, traffic lights and road/cycle path type, which are also shown in *Table 6.*

Table 6 gives an overview of all attributes present in the Fietsersbond dataset, that are within the scope of this study (next page).

Table 6: Contents of Fietsersbond dataset (cycling union)

Attribute name	**Content**	**Facility value**	**Reclassification**
WEGTYPE	Type of road	Bromfietspad (langs weg)(sep) Fietspad (langs weg)(sep) Fietsstraat(non-sep) Normale weg(non-sep) Solitair bromfietspad(sep) Solitair fietspad(sep) Veerpont(un) Ventweg(non-sep) Voetgangersdoorsteekje(un) Voetgangersgebied(un) Weg met fiets(suggestie)strook(non-sep) Unknown	Separated cycling facility (sep) Not-separated facility (non-sep) Unknown (un)
WEGDEKSRT	Road surface	Asfalt/beton halfverhard Klinkers Onverhard Schelpenpad Tegels Overig (hout/kinderkopjes) Unknown	Smooth surface (Asphalt / concrete slabs) Less smooth surfaces (other) Unknown
WEGKWAL	Road quality	Goed Redelijk Slecht Unknown	High quality Medium quality Low quality Unknown

All values in the three attributes in

Table 6 will be reclassified into only three values. Firstly because the defined values have little differences. Moreover, previous research has already made use of other classifications. By classifying in a similar way previous studies can be compared with this current study.

4.4 Density of addresses dataset

Another dataset that is required is the CBS Statline dataset of density of addresses. The density of addresses is a calculated statistic that returns the number of addresses per square kilometre for each neighbourhood. Per address is counted how many addresses are present within a radius of 1 kilometre. This study uses a derivative of this calculation, since is it not necessary to use

calculations used for each house. Therefor the data is aggregated on neighbourhood level. The data is updated yearly by CBS (CBS, 2014).

4.5 Land use dataset

The BBG (Basis bestand bodemgebruik- Key registry of land uses) will be used for determining the land use effect on cycling speed. The dataset is derived from CBS (2012). This dataset contains 13 classes of land use. These classes will be reclassified. A simplistic land use map will be used to find relationships between cycling speed and land use, because a high number of land uses would make the statistical model too complex. Moreover various land uses have a similar effect on the cycling speed relationship and could lead to multicollinearity.

A couple of pre-processing steps have to be performed before the land use data can be used for statistical analysis. The first step is creating a raster dataset out of the vector land use file. Thereafter, each land use cell can be reclassified. In Table 7 the reclassification of the land use dataset is outlined.

Table 7: Contents of Bestand Bodem Gebruik (Land use) dataset

Old land use	**New land use**
Industrial area/Business park	Built-up
Semi-built up	Proximity wise
Forest	Forest
Recreation	Proximity wise
Built-up	Built-up
Dry natural terrain	Agriculture
Horticulture	Agriculture
Roads	Proximity wise
Wet natural terrain	Null
Agriculture	Agriculture
Water	Null
Airport	Null
Railway	Null

This reclassification results in three types of land use: Agriculture, Built up and Forest. Figure 15 visualizes the reclassification of the land use types. Since cyclists do not ride through water, an airport or a railway those land uses have a null value. Semi-built up, recreation sites and roads will be reclassified by proximity to another land use in the following order:

1. Roads: A lot of cyclist routes will be on or follow a road. Since this study is interested in the land use influences on cycling it is not interesting to have a land use class which covers roads. Participants cycling on a road will statistically not be influenced by land use when the land use ‘roads’ would not be reclassified.
2. Recreational sites: Recreational sites can differ in reality from little parcs in urban areas to golf courses outside urban areas. Therefor the nearest cell with another land use type will be assigned to this cell. When recreation sites are near agriculture, it is logical that cyclists perceive this area more as agricultural than for example as built up area.
3. Semi-built up: Similar to 2. Recreational sites, semi-built up is reclassified according to proximity to other land use types. Since semi-built up covers often the shift from built-up to agriculture or forest, the proximity measure is a good tool for reclassifying.

After reclassifying the raster dataset will be converted back to a vector dataset in order to calculate the route percentages (variables). The second simplified land use map is used to calculate percentages for each participant. Each participant has its own shapefile and with that its own route. This route has been intersected (intersection tool ArcGIS Pro) with the simplified land use map. The result is a set of tables (for each participant one) containing distances through each land use type. In python the segment distances that were intersected, are being summed and divided by the total trip distance (as explained in the methodology chapter).

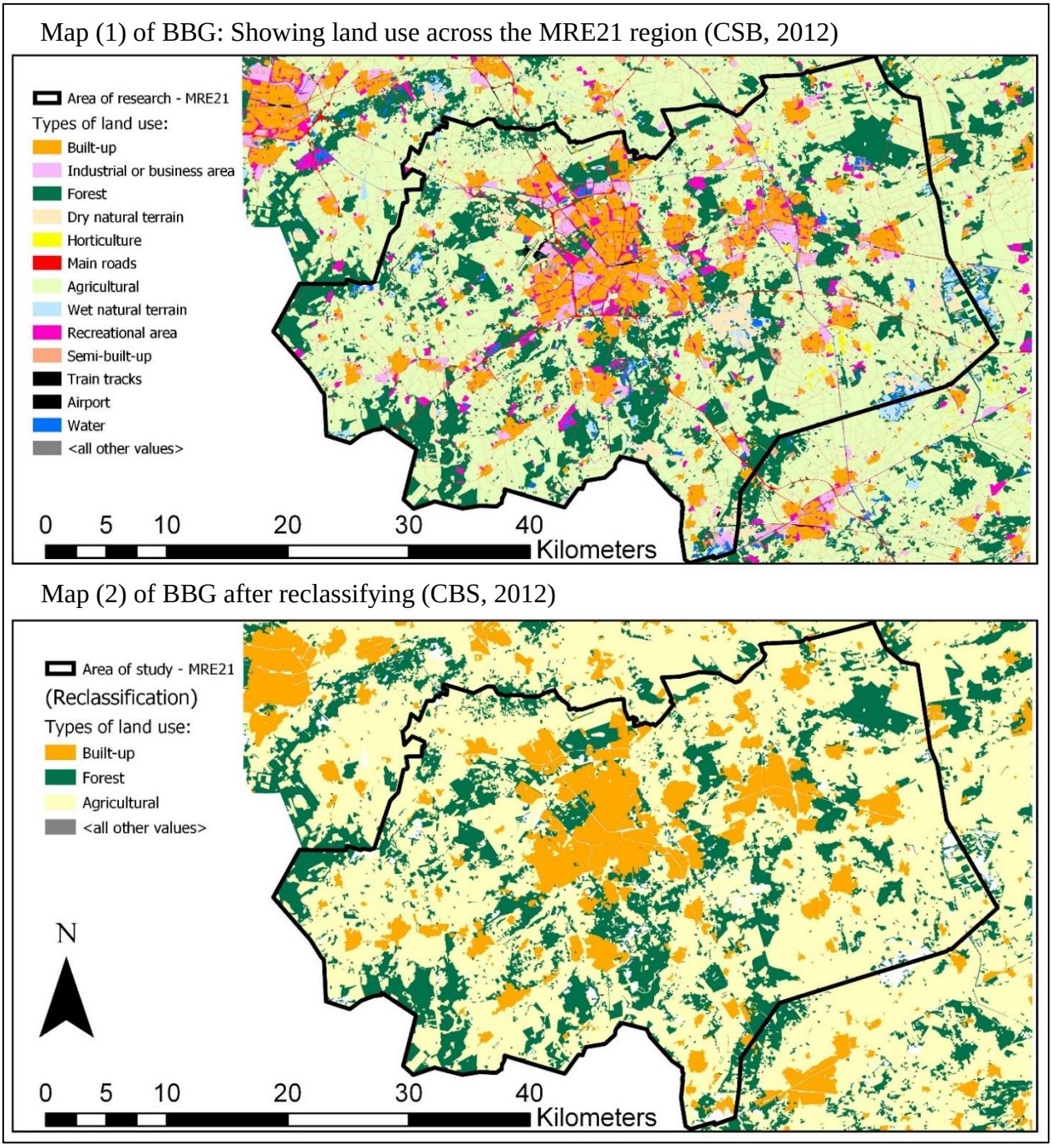

Figure 15: Two maps showing the reclassification of the BBG

4.6 Data sample B-riders dataset

The B-riders dataset contains thus information on speed, age and gender per respondent. The dataset is sliced in order to derive the information necessary for this study. Since the MRE21 is the area of research, the dataset will be narrowed down according to this area. In Table 8 the respondent statistics of the sample taken are outlined.

Table 8: Age and gender, B-riders Sample and CBS statistics (CBS, 2014)

Number of respondents (sample)	**Average age commuters in years (sample)**	**Gender (sample)**	**Average age in the Netherlands in years (CBS)**	**Gender in MRE21 (CBS)**
734	50,4s	Male 53,4% Female 46,6%	41,0	Male 50,4% Female 49,6%

Each respondent has cycled at least one or more routes within the MRE21 region. A random sample is taken for each respondent. The sample is taken as follows:

- Since this study is only interested in commuters the dataset has been queried on routes that were either from 'home' to 'work' or from 'work' to 'home'.
- The dataset is grouped by participant and unique route number thereafter.
- Per participant one random route has been picked. This means that when a participant only cycled one trip that trip always has been chosen.

Within the sample some cases will be excluded according to the nature of the data. All routes that were shorter than 2000 meter have been excluded due to larger margins of error of speed calculations. As explained in speed calculation method 2, as trip distances become shorter the measured or calculated speed becomes less reliable. Moreover when looking at discrepancies between the two speed calculations the discrepancies increased as the total trip distance decreased. 69 out of 734 routes were shorter than 2000 meters and were excluded. All statistical calculations will be done with 665 routes. Furthermore the first method of speed calculation is used while performing the analysis' within chapter 5, as the first method appeared to be most reliable one. Furthermore it would be interesting to compare different measuring/calculation methodologies, this is however not the goal of this study and also not lying within its scope. Additional research could further investigate the differences between cycling speed measuring methodologies.

Chapter 5: Statistical analysis

This chapter tries to explain which infrastructural and built environment factors influence cycling speed. The results will consists of a statistical analysis explaining the relationships between variables and cycling speed.. Firstly all hypotheses that were set up in chapter 3 will be discussed. The result will be analysed for its significant effect on cycling speed or its lack of effect. Moreover, will be discussed for which reasons this effect is present or absent. Secondly, a similar analysis will be performed wherein a separation is made between e-bike users and speed pedelec users. Lastly a statistical regression model has been built containing most important variables regarding this research. The length of use of this model will be discussed as well.

5.1 Analysis 1

Analysis 1 is a basic analysis. It uses the original dataset which is not yet split per type of bike.
Hypotheses
Hypotheses will be tested by modelling variables one by one. The results provide insight in the dataset, the sample and the variables that should be chosen for modelling average cycling speed. Most hypotheses will be tested with a simple linear regression test. Since most variables have a ratio scale a simple linear regression model is built for these variables. The model tries to explain a part of cycling speed. Expected is that most variables explain average cycling speed only for a small part. As discussed in the theoretical framework there are a lot of different variables that explain cycling speed. For that reason it is difficult to determine a clear relationship with cycling speed. Despite the fact that effects can be little, this chapter tries to explain the size of effects or the lack of effect. Furthermore it is important to consider that these simple linear models have not been controlled for personal characteristics. The final multiple regression model will control for these variables. Two variables are tested with an independent samples t-test: H1 and H3. Those independent variables contain both two groups, respectively male & female and speed pedelec & e-bike. The t-test will conclude whether average speeds are similar or not.

Table 9 shows 8 simple linear models which try to explain the average cycling speed. Not every hypothesis can be found within this table. Table 9 only gives relevant models. All models that did not fit are not included, but will be discussed thereafter. Table 9 gives extent of significance, a direction of the relationship between the predictor variable and outcome and the R^2.

Table 9: Simple linear models of 8 predictors; built to test for hypotheses (chapter 3).

	R^2	Predictors	B	Standard Error	Beta(β)	p-value
Model 1	0.2%	Constant	21.068	0.715		p=0.000
		Age	-0.014	0.014	-0.039	p=0.315
Model 2	5.3%	Constant				p=0.000
		Traffic lights/km	-1.589			
Model 3	0.7%	Constant	19.344	0.476		p=0.000
		Smooth	0.015	0.007	0.084	p=0.031

Model 4	0.4%	Constant	20.644	0.213		p=0.000
		Not Smooth	-0.013	0.008	-0.065	p=0.096
Model 5	10.7%	Constant	22.407	0.255		p=0.000
		Perc. built-up	-0.038	0.005	-0.327	p=0.000
Model 6	6.5%	Constant		0.190		p=0.000
		Perc. agriculture	0.034	0.005	0.256	p=0.000
Model 7	4.5%	Constant	19.682	0.161		p=0.000
		Perc. forest	0.048	0.009	0.213	p=0.000
Model 8	5.6%	Constant	21.421	0.204		p=0.000
		Avg. DoA	-0.001	0.000	-0.236	p=0.000

Interesting is that 8 out of 12 variables significant influence the average cycling speed. As discussed by Schantz (2017) important determinants of cycling speed are gender and age. This relationship has been tested with the independent samples t-test. This test showed that there is a significant difference between males (M=20.76) and females (M=19.88). Although this analysis shows that indeed males have a higher average cycling speed than females, the difference is small. The mean difference is only 0.89 km/h (p=.000). Furthermore this analysis shows that there is no linear relationship between age and cycling speed. Furthermore when recoding age into two groups (young and old; 50 years old or younger and 50 years and older) an independent samples t-test concluded that there are no significant differences between the means of the average speed either. An explanation for differences with the study of Schantz (2017) could be that this study only uses participants riding an e-bike or speed pedelec whereas other studies have participants using non-electric bikes. This explanation could also explain a relative small difference in speed between males and females. When electric aided bikes are used, cycling speed is less influenced by the amount of physical effort of a participant. A possible explanation of difference in sexes could be that males are more willing to take risks than females by cycling at higher speeds.

As expected speed pedelec users (M=21.45) have a higher average speed than e-bike users (M=19.38). The t-test concluded a difference of means of 2.07 km/h (p=.000). Since maximum speeds are being used for determining whether a participant is an e-bike or a speed pedelec user this conclusion should be observed with caution.

Built environment

Another interesting result is the difference in explanation of cycling speed between DoA and percentage Built-Up area. Expected is that there are similarities between those variables. Important to consider is the difference in measurement unit. DoA values have a range from 23 to 3996 (within this dataset) and a percentage of route through built-up area has a range from 0 to 100. An increase of DoA from 3000 to 4000 would lead to an increase of average cycling speed of approximately 1 km/h. And an increase of Built-up area from 75% to 100% would lead to an increase in speed of 0.95 km/h, which is a similar result. Still, DoA explains only 5.6% of average cycling speed while built-up area explains 10.7%. DoA seems a more accurate measure because of the level of detail of the data, as the DoA is calculated by using data at neighbourhood scale. An explanation could be that built-up area overlaps with other possible variables that influence cycling speed, which are not explained by the DoA.

Lastly, it is interesting that most infrastructural variables do not influence the average cycling speed within this sample. The road quality and the type of road (separated facilities or not) are not affecting cycling speed. Further analysis will follow later on in this chapter.

Summarizing the first results, there are some variables that show a significant relationship with the average speed. This means that the following variables within analysis 1 affect cycling speed:

- Gender
- Type of bike (e-bike or speed pedelec)
- Percentage of route through smooth surfaces (asphalt or concrete slabs)
- Percentage of route through all three land use types (built-up, agriculture and forest)
- Average DoA

Further analysis will show whether this conclusions are correct or not.

5.2 Analysis 2: separating bike types

Speed pedelec versus e-bike

The B-Riders dataset sample only consists of e-bike users and speed pedelec users. As discussed within the methodology section those different types of bikes are able to operate at very different speeds. E-bike provide only aid up to 25 km/h and speed pedelecs provide aid up to 45 km/h. This difference results noise in analysis 1. The result of hypothesis 3 in Table 9 shows that speed pedelec users have a significant higher average speed than e-bike users. As users of both groups possibly behave in a different way a second analysis has been performed. Every hypothesis will be tested once again while split by type of bike.

Hypotheses

Table 10 and Table 11 show respectively 9 and 5 linear regression models. The tables are formatted similar to Table 9. The goal is to get a glance of differences between speed pedelec and e-bike users. Within the following two tables models are a way to test for the hypothesis which were set up in chapter 3. Only important models have been shown.

Table 10: E-Bike results

		Predictors	**B**	**SE**	**Beta(β)**	**p-value**
Model 1	0.0%	Constant	19.653	0.935		p=0.000
		Age	-0.005	0.018	-0.014	p=0.787
Model 2	1.2%	Constant	18.271	0.558		p=0.000
		High road quality	0.016	0.008	0.109	p=0.040
Model 3	9.4%	Constant	19.973	0.170		p=0.000
		Traffic lights/km	-1.984	0.327	-0.307	p=0.000
Model 4	4.7%	Constant	17.032	0.579		p=0.000
		Smooth	0.035	0.008	0.217	p=0.000
Model 5	2.1%	Constant	19.955	0.255		p=0.000
		Not Smooth	-0.026	0.009	-0.144	p=0.007

Model 6	19.2%	Constant	21.741	0.289		p=0.000
		Perc. built-up	-0.044	0.005	-0.438	p=0.000
Model 7	12.7%	Constant	18.026	0.233		p=0.000
		Perc. agriculture	0.041	0.006	0.356	p=0.000
Model 8	6.4%	Constant	18.685	0.200		p=0.000
		Perc. forest	0.052	0.011	0.254	p=0.000
Model 9	9.9%	Constant	20.636	0.244		p=0.000
		Avg. DoA	-0.001	0.000	-0.315	p=0.000

Table 11: Speed pedelec results

		Predictors	**B (confidence interval)**	**Standard Error**	**Beta(β)**	**p-value**
Model 1	0.1%	Constant	21.942	0.962		p=0.000
		Age	-0.010	0.019	-0.029	p=0.606
Model 2	5.8%	Constant	21.997	0.194		p=0.000
		Traffic lights/km	-1.519	0.350	-0.249	p=0.000
Model 3	8.3%	Constant	23.371	0.393		p=0.000
		Perc. built-up	-0.034	0.006	-0.287	p=0.000
Model 4	6.0%	Constant	20.463	0.268		p=0.000
		Perc. agriculture	0.034	0.008	0.244	p=0.000
Model 5	2.7%	Constant	20.937	0.231		p=0.000
		Perc. forest	0.036	0.012	0.165	p=0.004
Model 6	7.6%	Constant	22.737	0.295		p=0.000
		Avg. DoA	-0.001	0.000	-0.275	p=0.000

A first conclusion that could be drawn after looking at differences between Table 10 and 11 is that these models explain average e-bike speed better than the speed pedelec speed. 8 out of 13 hypotheses of the e-bike user group have been tested truthfully. Interesting is the larger R^2 values of the linear models in comparison to analysis 1. By making a differentiation between e-bike users and speed pedelec users the linear models of e-bike users seem to explain cycling speed of that group better. The first hypothesis that stands out is hypothesis 1. The means of average cycling speed are similar for both sexes(t-test: there is no significant difference in means; p=.782). Since the speed pedelec group does have different means (p=.000; mean difference=1.38) between males (M=22.00) and females (M=20.62), an explanation could be found there.

E-bike results

Four infrastructure variables are affecting average cycling speed of e-bike users: 'traffic lights per km', 'smooth road surface' or 'not smooth road surface' and 'road quality'. There is expected that traffic lights have the biggest impact and the type of cycling facilities have the least effect. This appears to be true as there is no linear relationship found between type of road facility and the average cycling speed of e-bike users. Interesting within the other relationships is that roads that are considered 'comfortable' (road quality = high, road surface = smooth) are affecting cycling speed more than roads that qualified as 'less comfortable' (road quality = low, road surface = not smooth). The latter ones did not show a linear relationship. Other factors probably influence cycling speed more than the quality of roads. Another explanation could be that participants with lower average speeds more use of high quality roads than users with higher average speeds. When measuring the extent of influence of infrastructure it is good to focus on the objects that make cycling comfortable, although within this sample only the variable smooth would be considered to have a positively effect on the average cycling speed (of in this case e-bike users).

Nearly all built environment variables affect cycling speed of e-bike users. Although a model has been built for the forest variable assumptions have not been met. The heteroscedasticity of the forest variable could be explained by the skewness of the forest percentages. Most routes did not cycle through forest and only a few participants have higher percentages of riding through a forest. The main explanation of land use effects on cycling speed can be found in the difference between built-up area and other land uses. Built-up area has a much more complex structure in comparison to forest areas and agricultural areas. The high R^2 of the built-up area could probably be explained by the high number of complex effects exists within this type of land use. Built-up area has higher densities of road users, the ratio between road size and the volume of road user is higher and cyclists have more interactions with their direct environment (Klinger et al., 2013).

Speed pedelec results

The independent t-test result of hypothesis 1 states that there is a significant difference in average cycling speed of speed pedelec users between sexes in contrast to the e-bike group. An explanation for the differences between analysis 1, the e-bike group and the speed pedelec group could lie in the fact than women perceive a decreased safety on speed pedelecs in comparison to men (Haustein & Møller, 2016). Whereas e-bikes only support up to 25 km/h the speed pedelec supports up to 45 km/h. Lower safety feeling of women could affect the cycling speed, especially at higher speeds. Since e-bikes users cycle at lower speed this effect would be less significant. There could be other reasons for differences between gender. Further research would be necessary to conclude these assumptions.

Interestingly, traffic lights influences the average speed of speed pedelec users less than the e-bike equivalent of the linear model. The variable explains 5.8% of the cycling speed of speed pedelec users. The lower coefficient and R^2 value is probably due to the fact that speed pedelecs accelerate faster than e-bikes. Other infrastructure variables do not affect cycling speed, which is unexpected. The study of Hölzel et al. (2012) concluded that there is a dependency of cycling speed when determining the comfortability and rolling resistance of a road surface type (for example cobblestones or asphalt). It would be logical that speed pedelecs cycling speeds are more affected by the road surface type and road quality than e-bike cycling speeds. The result does not conclude that e-bike speeds are more influenced by infrastructure variables than speed pedelecs, but whereas e-bikes average speeds have a linear relationship with infrastructure variable, speed pedelec average speeds do not have this relationship.

Hypotheses that were found to be true for the speed pedelec group have to be looked at with caution. As discussed with the case of the forest variable in the e-bike result section a couple of variables have a distribution of values with a high skewness. Although skewness of independent variables does not always result in violation of assumptions, this is the case for some speed pedelec hypotheses. Another reason could be that the assumption of 27 km/h as a maximum speed threshold is wrong. Speed pedelec cyclists how did not cycle faster than 27 km/h are not included in this analysis. Above that, speed pedelec users could behave differently in comparison to its e-bike equivalent because of the difference in amount of electrical support.

The last variable which shows an interesting result is the type of road variable. There is a positive relationship between the average cycling speed of speed pedelec users and the percentage on non-separated roads. This would mean that speed pedelecs cycle faster while cycling through traffic. However this relationship tested not significant (p=.105) using the rule of thumb (p > .05). Still, it is worth to mention for further studies.

Quality of variables when explaining cycling speed

Looking at hypotheses results, histograms, means and standard deviations some conclusions can be drawn about the quality of infrastructure variables. Firstly the DoA addresses variable influences cycling speed similar in all three analysis' so far. The coefficient -0.001 is for every analysis the same (Table 9, Table 10 and Table 11 show respectively 9 and 5 linear regression models. The tables are formatted similar to Table 9. The goal is to get a glance of differences between speed pedelec and e-bike users. Within the following two tables models are a way to test for the hypothesis which were set up in chapter 3. Only important models have been shown.

Table 10Although R^2 percentages range between 5.6% and 9.9% it is reliable to conclude that the DoA explains average cycling speed substantially. Furthermore could be concluded that infrastructural variables that measure 'comfortable' characteristics are better predictors than infrastructural variables which measure 'uncomfortable' characteristics. However, it cannot be concluded that the effect of low quality roads is less than the effect of high quality roads, since there is only tested whether a linear model explains relationships between the described variables. Analysing the dataset could clarify some results in this respect. Figure 16 shows the histograms, means and standard deviations of the road quality and road surface variables. In the top left corner and the top right corner histograms of road quality are shown. In the lower left corner and the lower right corner the road surface type histograms are shown. Histograms on the left are (heavily) positively skewed and histograms on the right a little negatively skewed. Especially the low road quality variable is skewed so much that it results in unreliable results within this study. A possible reason for the skewed data could be that respondents chose for routes that have high quality roads, or they try to avoid low quality roads. A more likely reason is the lack of low quality roads within the dataset of the Fietsersbond (cycling union). Only 1.3% of the road segments (not length!) have the value 'low quality'. Furthermore 44.4% of the road segments have missing values. A similar comparison can be made for the surface type variable. Although the values 'smooth' and 'not smooth' are better distributed amongst each other. Missing: 44.4%; Smooth: 24.0%; Not smooth: 31.6%. Although the percentage of missing values is quite high one should be aware that the percentages reflect segment counts. Segments that are part of intersections often have missing values. Those segments are very short. Comparing missing values to the actual length of the network 33% of the network has missing values.

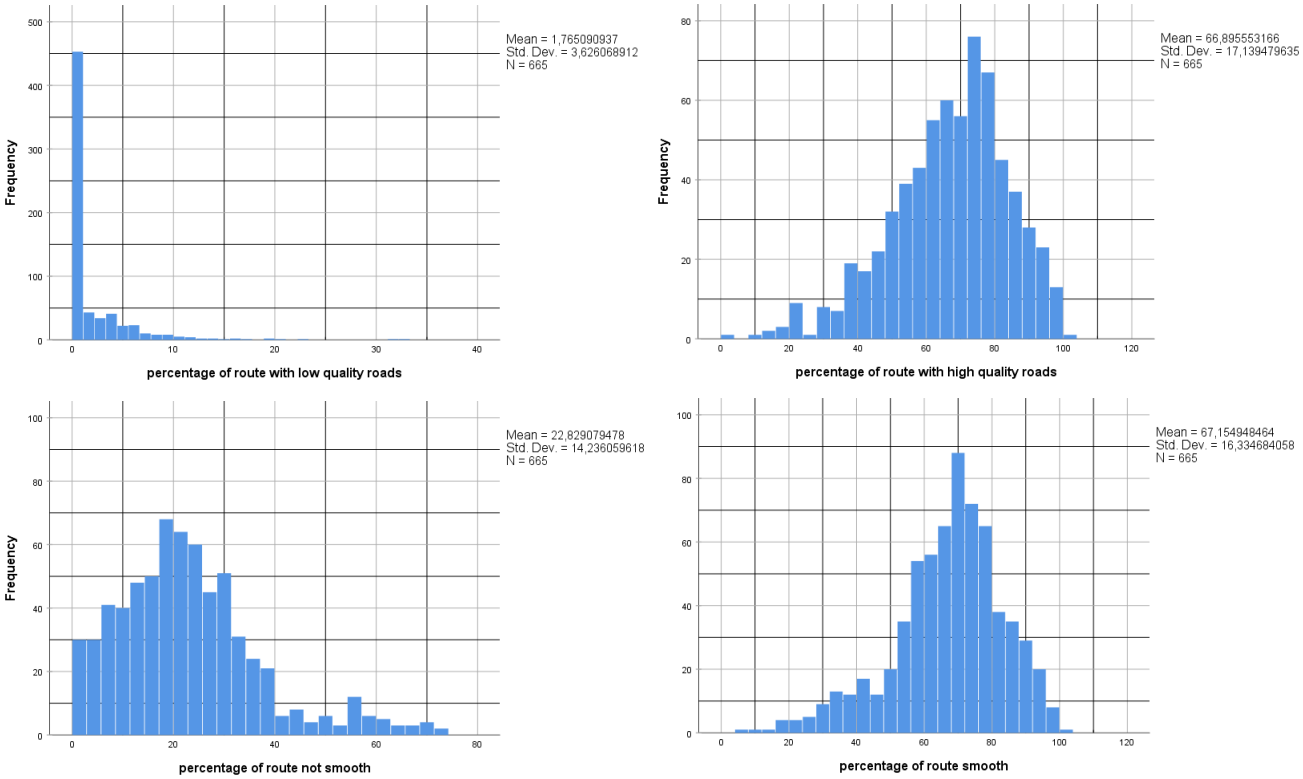

Figure 16: Distribution of frequencies (Histograms) of four variables: low & high quality roads and not smooth & smooth roads

On the contrary the quality of the built environment variables is higher. In all three analysis' built environment multiple models are significant explaining the average cycling speed. While road infrastructure datasets have a lot of missing values built environment variables hardly have any missing values. The land use variable has been split up in three values: built-up, forest and agriculture. Since there are less participants cycling through forests or only a short part through forests, the forest variable is heavily skewed when looking into its frequency histogram. Moreover there are less forest areas within the area of study than urban areas or agricultural areas. This resulted in all three analysis' in possible problems with assumptions of either residuals or homoscedasticity. Within this study the forest variable is considered too unreliable for using it in the multiple regression model. However this does not mean that there is no relationship between cycling speed and cycling through forest areas. The other two land use models perform much better. Especially the built-up variable explains the average cycling speed with a very high certainty.

Although the violations of assumptions of these models could be analysed better (important for improving models) it will not be discussed within this study any further. The main goal of this section is to provide a brief overview of the differences between the e-bike and speed pedelec group with respect to cycling speed. Additional analyses will be performed in order to build a multiple regression model. Within the scope of that model every variable will be discussed furthermore.

5.3 Multiple regression model

After analysing all variables in the first two analysis' this section will try to estimate a multiple regression model that explains the average cycling speed by infrastructural and built environment variables. After analysing the simple linear regression models a better understanding This model should answer the main research question, to what extent the built environment and infrastructure influence cycling speed of commuters.

Exclusion of duplicate variables

In previous analysis' variables have been examined account for the same variation as other variables in the dataset. The variables in question are: 'road quality' (2 sub variables), 'road surface type'(2 sub variables), 'type of cycling facility' (2 sub variables) and 'land use' (3 sub variables). For example percentages surface type = 'smooth' and road surface type = 'not smooth' have been calculated. Road surface type only has only one value per segment, so in fact both variables are measuring a similar effect. However, the result depends on whether cyclists are more influenced by smooth roads or by roads that are not smooth. In a statistical model only one predictor measuring a similar effect can be used. For that reason only four sub variables will be evaluated within the selection of predictors. Out of the 9 optional independent variables the following 5 have been selected for evaluation:

- Percentage 'Smooth'
- Percentage 'Built-up area'

Other variables that will be evaluated for the final model are:

- Average DoA
- Number of traffic lights per km
- Standard deviation of segment speeds within trips

Furthermore two dummy variables will be used within the model to control for influences on the cycling speed that are not lying within the original scope of this study. The first dummy variable is gender and the second one the type of bike that was used. Although in the first two analysis' age did not affect cycling speed it will be added to the model. Firstly, age is important to control for and secondly the simple linear regression cannot conclude that age will not affect cycling speed in the multiple regression model either.

Correlation matrix

The first step of building a model is to identify correlations of all possible predictors.

Table 12: Overview of correlations between predictors.

Variable	Average speed	DoA	Smooth percentage	Built-up percentage	Traffic light/km	Std. dev. of	Bike type
Avg. speed		-.236**	0.084*	-.327**	-.233**	-.047	.356**
DoA			-.129**	.727**	.604**	.230**	.114**
Smooth				-.247**	-.089*	-.042	-.051
Built-up					.572**	.148**	.064
Traffic lights						.289**	.071
Std. dev. speed							.331**
Bike type							

Significance values: Sample: N=665

** when $p \leq 0.01$

* when $p \leq 0.05$

The first important insight is that all independent variables show a significant correlation with average speed with exemption of the std. dev of speed. Furthermore it is important to look for high correlations amongst the independent variables. Values that stand out ($r > 0.5$) have been highlighted in Table 12. The threshold value for excluding variables from the multiple regression is $r > 0.9$, which is seen as a rule of thumb by Field (2013). None of the correlation values exceed the threshold value. Nevertheless, the correlations that stand out are a warning for collinearity. Especially the high correlation ($r = .727$) between DoA and the built-area predictor was expected. Correlating independent variables that stand out will be tested with the VIF.

Multiple regression results

Firstly the result will be outlined, and thereafter the reliability of the model will be discussed. This includes discussion on indicators used the assess the fit of the model. As outlined in the methodology chapter hierarchical regression is used. In four steps, four models have been built. Firstly entered (enter method per step) are the control variables such as age and gender. Moreover in the first step the dummy variable type of bike was added. Those variables are expected to affect cycling speed, but are not the variables of interest. Secondly entered are infrastructural variables that have proven to be affect cycling speed. In the third block the built environment variables has been added: percentage of route through built-up area. The average density of addresses have been left out the model since it correlates highly with the built-up variable. Moreover when adding DoA to the model it does not add any explanation of cycling speed. A logical explanation would be that the built-up variable does explain for more underlying factors than the DoA does. Lastly the standard deviation of speeds during trips have been added. Table 13 shows the model summary of entering these predictors.

Table 13: Model summary

Model	Predictors	R^2	ΔR^2 – significance	ANOVA
1	Type of bike speed pedelec vs e-bike Gender male vs female Age	13.9%	13.9% **	.000**
2	Traffic lights (TL/km) Smooth percentage	21.2%	7.3% **	.000**
3	Built-up percentage	26.9%	5.7% **	.000**
4	Std. dev. speed	27.9%	1.0% **	.000**

Significance values: Sample: N=665
** when $p \leq 0.01$
* when $p \leq 0.05$

The summary provides first insights in how well the models predict the average cycling speed of e-bike users. The ANOVA column concludes that these four models better explain cycling speed that using the mean as a model. The ΔR^2 gives information about whether or not adding predictors was leading to an increased explanation of the dependent variable. A side note here is that with the increase of numbers predictors the R^2 always grows. In addition it is helpful that SPSS gives a significance value for the ΔR^2. Table 14 is a summarized version of the SPSS coefficient table. This table gives the final results of the multiple regression model.

Table 14: Linear models of predictors of average cycling speed of commuters.

	Predictors	B (confidence interval)	Standard Error	Beta(β)	p-value
Model 1	Constant	19.641	.679		
	Gender (d)	.692	.212	.119	.000
	Type of bike (d)	1.963	.211	.339	.001
	Age	-.011	.013	-.032	.000
Model 2	Constant	19.796	.778		.386
	Gender (d)	.623	.204	.107	.000
	Type of bike (d)	2.092	.203	.361	.002
	Age	-.020	.013	-.055	.000
	Smooth perc.	.012	.006	.069	.116
	TL/ km	-1.754	.238	-.258	.048
Model 3	Constant	21.907	.807		.000
	Gender (d)	.610	.197	.105	.000
	Type of bike (d)	2.120	.196	.366	.002
	Age	-.018	.012	-.050	.000
	Smooth perc.	.002	.006	.010	.141
	TL/km	-.630	.278	-.093	.774
	Built-up perc.	-.034	.005	-.299	.024
Model 4	Constant	22.610	.834		.000
	Gender (d)	.504	.198	.087	.000
	Type of bike (d)	2.337	.207	.403	.011
	Age	-.018	.012	-.051	.000
	Smooth	.002	.006	.010	.128
	TL /km	-.410	.286	-.060	.764
	Built-up perc	-.035	.005	-.304	.152
	Std. dev speed	-.178	.059	-.113	.000

Note: R^2 increase per step shown in Table 13. ; (d) = dummy variable

Table 14 outlines the model result in four modelling steps. This means that model 2, 3 and 4 can be separately used to explain cycling speed. Every model has advantages and disadvantages due to overlapping effects. In the column 'B' the coefficient of each predictor can be found as well as the constant of the model. The last column provides the result of a t-test which tests whether the B coefficient is significantly different than 0. The lower the p value the stronger this particular predictor affects the model. A non-significant p value does not conclude there is no relationship between the particular independent variable and avg. cycling speed. The first thing that stands out is the lack of significance of the traffic lights predictor in model 3 and 4. After adding the built-up variable the amount of influence of the TL/km variable decreased. As a lot of traffic lights are located nearly or within built-up area a logical explanation is that the built up variable explains a similar part of the variation. Furthermore the increase of only 5.6% R^2 in step 3 is interesting (Table 13**Fout! Verwijzingsbron niet gevonden.**), while a simple linear regression of Built-up area explained average cycling speed with a R^2 of 19.2% (Table 10).Table 10 and Table 11 show respectively 9 and 5 linear regression models. The tables are

formatted similar to Table 9. The goal is to get a glance of differences between speed pedelec and e-bike users. Within the following two tables models are a way to test for the hypothesis which were set up in chapter 3. Only important models have been shown.

Table 10This could be explained by the high correlation between the number of traffic lights/km (Table 12**Fout! Verwijzingsbron niet gevonden.**). Interesting is the influence of age within all four models. While simple linear models within analysis 1 and 2 did not show any relationship, in the multiple regression the contribution is significant and stable. The smooth variable does influence the model only marginal. Although the influence of the smooth variable is small it probably does explain a little cycling speed that is not already covered by other variables.

This section will discuss the final model (model 4). Although the Beta (fifth column Table 14) is not the best way of comparing different variables, it is the easiest way to have a glance of the size of effect of each predictor. The type of bike influences the model the most when looking at the Beta (.403). This is an obvious result as the amount of support increases a lot when cycling on a speed pedelec instead of an e-bike. As the built-up area predictor explains for the same variety as other predictor it's Beta is the second largest one (-0.304). The third significant contributor is the standard deviation of speed with a Beta of -0.113. Although its effect is small within the model the standard deviation is an interesting predictor to discuss. The standard deviation of speed implicates how much a cyclist differs in speeds. Although the variable does overlap in explanation of other variables that influence fluctuations of speed, it affects cycling speed that not has been explained by other variables. To clarify this, cyclists that have to slow down for corners more often are more likely to have bigger fluctuations in speed. While cyclists that have less corners to take on their route are likely to have a smaller standard deviation in speed. Although this predictor only added 1.0% of R^2 within the model it is important to mention.

Discussion

This chapter will discuss the research questions, limitations and recommendations for further studies.

Research questions

This study aimed to find the extent of infrastructural and built environmental influences on cycling speed. Although infrastructure and built environment variables have as well theoretical as well as practical overlap both explain cycling speed individually to some extent. This section will discuss the answers of the research questions. Each sub question will be discussed accordingly.

How are personal characteristics related to cycling speed?
This study performed three analysis': Testing of hypotheses, testing hypotheses while separating type of bike and a final multiple regression. The first conclusion is that within this sample is that age is only found significant in the multiple regression model. This means age is not a good predictor on its own. When controlling for gender, age and type of bike the interaction with the dependent variable changes. The first two analysis' did not give one clear outcome whether gender really did influence cycling speed within this sample. While studying the complete sample (no separation between e-bike and speed pedelec) there was a significant difference in means of average cycling speed. In addition a difference in means was found for the speed pedelec group as well. However, the e-bike group did not show a significant differentiation in means. It could be that female speed pedelec users are less willing to take risks that go hand in hand with higher speeds in comparison to males. An explanation could be that this is a less important factor on e-bikes due to the lower maximum and average speeds. An interesting result is thus that there is probably a difference in cycling behaviour between the sexes on speed pedelecs.

How do infrastructural variables, such as traffic lights, road quality and type of surface and type of facility, affect cycling speed?
Infrastructural variables that are considered important are, number of traffic lights, road quality and road surface type. Although traffic lights per km is an important variable it overlaps with built environment variables, which explain the average cycling speed for a similar part. Roads with a high level of comfortability appeared to have influence on the average cycling speed but this effect is marginal, especially in comparison to built environment variables. Low comfortability levels were more difficult to interpret because of the characteristics of the data. This study cannot conclude there is no relationship between less comfortable roads and average cycling speed, but it can conclude there was no relationship found. Two variables are considered part of the 'comfortability' measure: road quality and surface type. The latter one appeared to be the better predictor of the two. The last infrastructural variable, type of facility, did not show any relationship with the average cycling speed within this study, despite Namezi et al. (2018) concluded there was. Interesting is that within the analysis for speed pedelecs a small positive relationship between non-separated and average cycling speed could be detected. It was however not significant by the rules of thumb.

To what extent affects the built environment cycling speed?
The best predictors of cycling speed within this research were built environment variables. As well the average density of addresses as well as the land use variable have a clear relationship with the average cycling speed. Although the DoA was expected to be more influential than the built-up area variable (land use) the latter one showed larger explanation values. Both can be

used in estimating average cycling speed. Because the DoA variable has a high geographical density of information this variable can be used in further studies which try to explain cycling speed within built environment factors. Influence of the land use variable on cycling speed was evident. Most important differences could be found between urban areas and outside urban areas, which was expected. Due to dataset characteristics it is difficult to conclude anything on the forest land use. On the other hand, a clear negative linear relationship has been found between built-up area and average cycling speed and a clear positive one has been found between agricultural area and the average cycling speed. The last variable which is considered part of the built environment for now (could also be included in personal characteristics) is the standard deviation of speed. Although on the one hand the standard deviation of speeds could be part of personal behaviour it is probably also an explanation for built environment characteristics in the sense of the amount and angle of corners. To what extent it is really a good predictor within the built environment the variable should be studied further.

In what way do infrastructure and the built environment influence commuting cycling speed, controlled for personal characteristics?

In general could be said that built environment variables in general are better predictors of cycling speed. When modelling cycling speed it is important to take the built environment into consideration. Although infrastructure variables do explain cycling speed of electrical supported bikes the effect is little. However it is important to keep in mind that infrastructure variables have been measured at higher detail, while built environment variables measure at more generalizable scale. The scope of one's study is an important determinant while choosing variables that affect the cycling speed. Personal characteristics seem to be less important when predicting cycling speed of electric supported bikes.

Limitations and recommendations

Lack of research within the subject of built environment and cycling speed is a problem when validating the results of this study. Despite this lack, the exploratory results could be a first step in this field of study. The growing amount of data could help in setting up similar studies. Reviewing the results of this research there a couple of points of discussion. In order to elaborate on the built environment the built environment elements should be operationalised better. This study only used a simple variables to study the built environment influences. Reclassification of land uses was not a complex process and the DoA has only been used on neighbourhood level (Dutch). Although the first results are promising, it is not yet clear how separate elements of the built environment affect the cycling speed. In this regard, average cycling speeds only are not providing enough evidence to conclude the relationship between elements of the built environment and cycling speed. At higher scale, speeds at specific locations as well as relative speeds (ratio of difference between actual speed and desired speed) are elements that have to be included in further studies. It would be interesting to discover more about the factors that lie within the scope of these two variables.

Infrastructure variables measure effects at this higher scale (more detailed). However, infrastructural influences that were found, are affecting cycling speed only marginal. As the variables become more detailed the size of explanation becomes lower. Large samples are necessary to find relationships that are not that strong. Although the availability of data was not a problem, the amount of time was limited. For that reason more analysis' could have been done, having a more profound understanding of relationships thereafter. Furthermore the infrastructure dataset could have been improved. Some variables had up to 30% of missing

values. These values could have been studied by using deep learning or machine learning techniques (e.g. traffic lights).

Cycling speed studies used different methodologies to calculate the average speed, as well survey based (Schantz, 2017; Stigell & Schantz, 2011) as well as GPS based (Broach et al., 2012; Larsen & El-Geneidy, 2011; Plazier et al., 2017). This study used GPS-based methodologies which create opportunities to build new speed derived variables such as the standard deviation of speed and the type of bike. The type of bike is a crucial factor in the final model. However it is important to consider the ambiguity in the methodology used to create the variable. In order to simplify the model a threshold of 27 km/h is used as only measure. This means that the groups within this sample could have been classified wrongly. Speed pedelec users who did not cycle fast, were wrongly classified as e-bike user. Nevertheless analysis concludes that the separation between e-bikes and speed pedelecs could be quite accurate. The extent of this accuracy and a study about differences between those groups could be part of future research.

This research presents some interesting results, especially within the exploratory part (built environment) of this thesis. The extent of generalizability is however limited. Firstly the results only make sense when considering the type of participants. Firstly the dataset is biased due to the fact that there was a motivational factor to participate in the B-Riders project. Secondly participants are mainly middle-aged and working. However, the area of research is representative for all countries that have higher mode shares of bicycle use (e.g. Denmark). It is important to consider that results are only valuable when infrastructure and built environment is similarly. Different built environment and infrastructural characteristics (e.g. U.S.) make the results less interpretable.

Further research could focus on differences between types of participants, day times, seasons and road segments. The rise of use in GPS-methodologies is evident. In this respect these themes could be used to get more insight in explanation of cycling speed.

Conclusion

Cycling behavioural studies are currently lacking within the subject of cycling speed. The emergence of electrical supported bikes introduces a new spectrum of cyclists riding with higher speeds than before. This study aimed to find what infrastructural and built environmental factors do influence cycling speed. As infrastructure and the built environment have partly overlapping definitions, both have been discussed within this research.

Ewing & Cervero (2010) and Saelens & Handy (2008) have theoretically placed infrastructure within the built environment. The built environment has a low geographical scale and affects cycling speed strongly. Whether the DoA variable was used or the land use variable its effect on the cycling speed was evident. The built-up variable accounts for more explanation than the DoA variable. A logical reason is that the built-up variable more strongly accounts for infrastructure variables such as road structure (not part of study), but also traffic lights. The DoA variable explains cycling speed more genuine while it less correlates with infrastructure variables.

Infrastructure variables in itself explain average cycling speed only marginal. This effect is not only reduced by averaging speed for the whole route, but thus also because of the higher geographical scale. Furthermore, variables that measured 'comfortable' parts of the network had more influence than variables that measured parts of the network that were less 'comfortable'. It is interesting to research these effects by using road segments as unit of measurement. Still this study can conclude that there within this sample differences in cycling speed behaviour between e-bike users and speed pedelec users. Whether this conclusion has been drawn truthfully should be part of further research.

Another interesting result is the difference between sexes with respect to the type of bike. Prior research found that males cycle at higher speeds than women (Schantz, 2017; Shafizadeh & Niemeier, 1997). A logical explanation is the difference in physical strength. However electrical bikes would reduce that effect. This study concluded that there is no difference between sexes, while comparing mean speeds of e-bike users. There is concluded that there is a difference between sexes when looking at speed pedelecs on the contrary. It strongly suggests that males are willing to take more risks.

No clear results were found when looking into separated cycling facilities. In the Netherlands cycling is seen as rather save. However, a positive effect was expected between separated cycling facilities and cycling speed. Although this variable could affect cycling speed locally, the average cycling speed was not significantly affected.

To conclude, the statistical model can be of aid in a first step towards more detailed studies that try to explain cycling speed. A broader understanding of cycling speed can be important for analysis in travel behaviour. Furthermore it is important to be aware of the increased use of electrical supported bikes and its consequences for society.

References

Braun, M. T., & Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, *43*(2), 331–339. https://doi.org/10.3758/s13428-010-0046-8

Broach, J., Dill, J., & 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. https://doi.org/10.1016/J.TRA.2012.07.005

Caulfield, B., Brick, E., & McCarthy, O. T. (2012). Determining bicycle infrastructure preferences – A case study of Dublin. *Transportation Research Part D: Transport and Environment*, *17*(5), 413–417. https://doi.org/10.1016/J.TRD.2012.04.001

CBS. (2012). Bestand Bodemgebruik 2012. Retrieved from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische data/natuur en milieu/bestand-bodemgebruik

CBS. (2014). *Regionale kerncijfers Nederland*. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1551187383852

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

de Vos, C. (2018). *Investigating the influence of safety on cyclists route choice: a revealed preference study*.

El-Geneidy, A., Krizek, K. J., & Iacono, M. (2007). Predicting bicycle travel speeds along different facilities using GPS data: a proof of concept model. *Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, DC, USA*, 21–25.

Ewing, R., & Cervero, R. (2001). Travel and the Built Environment: A Synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, *1780*, 87–114. https://doi.org/10.3141/1780-10

Ewing, R., & Cervero, R. (2010). Travel and the Built Environment. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics*. Sage. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=c0Wk9IuBmAoC&oi=fnd&pg=PP2&dq=field+spss+statistics+4th+&ots=LbEhOJ_sZC&sig=GWmFWqOgmKs2LTk6IKshp348CWQ#v=onepage&q=field spss statistics 4th&f=false

Fishman, E., Washington, S., & Haworth, N. L. (2012). *Journal of the Australasian College of Road Safety.* ACRS. Retrieved from http://eprints.qut.edu.au/53981/

Handy, S. L., Boarnet, M. G., Ewing, R., & Killingsworth, R. E. (2002). How the built environment affects physical activity: Views from urban planning. *American Journal of Preventive Medicine*, *23*(2), 64–73. https://doi.org/10.1016/S0749-3797(02)00475-0

Harms, L., & Kansen, M. (2016). *Fietsfeiten || CROW Fietsberaad*. Retrieved from http://fietsberaad.nl/?lang=nl&repository=Fietsfeiten

Haustein, S., & Møller, M. (2016). E-bike safety: Individual-level factors and incident characteristics. *Journal of Transport & Health*, *3*(3), 386–394. https://doi.org/10.1016/J.JTH.2016.07.001

Heinen, E., van Wee, B., & Maat, K. (2010). Commuting by Bicycle: An Overview of the Literature. *Transport Reviews*, *30*(1), 59–96. https://doi.org/10.1080/01441640903187001

Hölzel, C., Höchtl, F., & Senner, V. (2012). Cycling comfort on different road surfaces. *Procedia Engineering*, *34*, 479–484. https://doi.org/10.1016/J.PROENG.2012.04.082

Hull, A., & O'Holleran, C. (2014). Bicycle infrastructure: can good design encourage cycling? *Urban, Planning and Transport Research*, *2*(1), 369–406.

https://doi.org/10.1080/21650020.2014.955210
Jensen, P., Rouquier, J.-B., Ovtracht, N., & Robardet, C. (2010). Characterizing the speed and paths of shared bicycle use in Lyon. *Transportation Research Part D: Transport and Environment*, *15*(8), 522–524. https://doi.org/10.1016/J.TRD.2010.07.002
Klinger, T., Kenworthy, J. R., & Lanzendorf, M. (2013). Dimensions of urban mobility cultures – a comparison of German cities. *Journal of Transport Geography*, *31*, 18–29. https://doi.org/10.1016/J.JTRANGEO.2013.05.002
Krizek, K. J., El-Geneidy, A., & Thompson, K. (2007). A detailed analysis of how an urban trail system affects cyclists' travel. *Transportation*, *34*(5), 611–624. https://doi.org/10.1007/s11116-007-9130-z
Larsen, J., & El-Geneidy, A. (2011). A travel behavior analysis of urban cycling facilities in Montréal, Canada. *Transportation Research Part D: Transport and Environment*, *16*(2), 172–177. https://doi.org/10.1016/J.TRD.2010.07.011
Marchetti, C. (1994). Anthropological Invariants in Travel Behavior. Retrieved from http://pure.iiasa.ac.at/id/eprint/4071/
Metz, D. (2008). The Myth of Travel Time Saving. *Transport Reviews*, *28*(3), 321–336. https://doi.org/10.1080/01441640701642348
Mobiliteitsbeeld 2017 | Kennisinstituut voor Mobiliteitsbeleid. (n.d.). Retrieved from https://www.kimnet.nl/mobiliteitsbeeld#toc
Moudon, A. V., Lee, C., Cheadle, A. D., Collier, C. W., Johnson, D., Schmid, T. L., & Weather, R. D. (2005). Cycling and the built environment, a US perspective. *Transportation Research Part D: Transport and Environment*, *10*(3), 245–261. https://doi.org/10.1016/J.TRD.2005.04.001
Nazemi, Mohsen; van Eggermond, Michael A.B.; Erath, Alexander; Axhausen, K. W. (2018). Studying cyclists' behavior in a non-naturalistic experiment utilizing cycling simulator with immersive virtual reality. *Arbeitsberichte Verkehrs-Und Raumplanung*, *1383*. https://doi.org/https://doi.org/10.3929/ethz-a-010180262 Rights
Plazier, P. A., Weitkamp, G., & van den Berg, A. E. (2017). "Cycling was never so easy!" An analysis of e-bike commuters' motives, travel behaviour and experiences using GPS-tracking and interviews. *Journal of Transport Geography*, *65*, 25–34. https://doi.org/10.1016/J.JTRANGEO.2017.09.017
Saelens, B. E., & Handy, S. L. (2008). Built environment correlates of walking: a review. *Medicine and Science in Sports and Exercise*, *40*(7 Suppl), S550-66. https://doi.org/10.1249/MSS.0b013e31817c67a4
Schantz, P. (2017). Distance, Duration, and Velocity in Cycle Commuting: Analyses of Relations and Determinants of Velocity. *International Journal of Environmental Research and Public Health*, *14*(10), 1166. https://doi.org/10.3390/ijerph14101166
Schepers, P., Twisk, D., Fishman, E., Fyhri, A., & Jensen, A. (2017). The Dutch road to a high level of cycling safety. *Safety Science*, *92*, 264–273. https://doi.org/10.1016/J.SSCI.2015.06.005
Schleinitz, K., Petzoldt, T., Franke-Bartholdt, L., Krems, J., & Gehlert, T. (2017). The German Naturalistic Cycling Study – Comparing cycling speed of riders of different e-bikes and conventional bicycles. *Safety Science*, *92*, 290–297. https://doi.org/10.1016/J.SSCI.2015.07.027
Shafizadeh, K., & Niemeier, D. (1997). Bicycle Journey-to-Work: Travel Behavior Characteristics and Spatial Attributes. *Transportation Research Record: Journal of the Transportation Research Board*, *1578*, 84–90. https://doi.org/10.3141/1578-11
Stigell, E., & Schantz, P. (2011). Methods for determining route distances in active commuting – Their validity and reproducibility. *Journal of Transport Geography*, *19*(4), 563–574. https://doi.org/10.1016/J.JTRANGEO.2010.06.006

Transportation research board institute of Medicine. (2005). *Does the Built Environment Influence Physical Activity?* Washington D.C. https://doi.org/0-309-09498-4

Van Genugten, W., & Van Overdijk, R. (2016). *Onderzoek Royal HaskoningDHV en TU/e toont gedrag van fietsers*. Retrieved from https://www.royalhaskoningdhv.com/nl-nl/nederland/nieuws/nieuwsberichten/royal-haskoningdhv-en-tu-eindhoven-tonen-gedrag-van-fietsers-aan/5682

Winters, M., Teschke, K., Grant, M., Setton, E. M., & Brauer, M. (2010). How Far Out of the Way Will We Travel? *Transportation Research Record: Journal of the Transportation Research Board*, *2190*(1), 1–10. https://doi.org/10.3141/2190-01

Appendix 1 SPSS Syntax

2.1 Hypotheses 1 to 12 in analysis 1

```
DATASET ACTIVATE DataSet1.
COMPUTE stop_km=stoplicht / Total_dist.
EXECUTE.

DATASET ACTIVATE DataSet1.
T-TEST GROUPS=gender('m' 'f')
  /MISSING=ANALYSIS
  /VARIABLES=gem_snel_realistic
  /CRITERIA=CI(.95).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER stop_km
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).



REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
```

```
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER DoA
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER perc_BU
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER perc_AG
  /PARTIALPLOT ALL
```

```
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER perc_FO
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER kwal_goed_perc
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).

REGRESSION
```

```
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER kwal_slecht_perc
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER smooth_perc
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
```

```
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER not_smooth_perc
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER std_dev
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER onRoad_perc
  /PARTIALPLOT ALL
```

```
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER offRoad_perc
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER leeftijd
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID, *ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2).
```

2.2 Hypothesis 1 to 12 in analysis 2 (separation e-bike and speed pedelec)

```
* Encoding: UTF-8.


DATASET ACTIVATE DataSet1.

SORT CASES  BY Fietstype_27.

SPLIT FILE LAYERED BY Fietstype_27.
```

Same syntax as 2.1, however the split file function is used to split the file in two parts.

2.3 Multiple regression model (e-bike)

```
* Encoding: UTF-8.


DATASET ACTIVATE DataSet1.
RECODE gender ('m'=1) (ELSE=0) INTO gender_d.
EXECUTE.

DATASET ACTIVATE DataSet1.
RECODE Fietstype_27 ('s'=1) (ELSE=0) INTO fietstype_d.
EXECUTE.


CORRELATIONS
  /VARIABLES=gem_snel_realistic DoA smooth_perc perc_BU perc_AG kwal_goed_perc
stop_km std_dev
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=perc_BU WITH gem_snel_realistic
  /MISSING=LISTWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=kwal_goed_perc WITH gem_snel_realistic
  /MISSING=LISTWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=smooth_perc WITH gem_snel_realistic
  /MISSING=LISTWISE.
```

```
GRAPH
  /SCATTERPLOT(BIVAR)=stop_km WITH gem_snel_realistic
  /MISSING=LISTWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=DoA WITH gem_snel_realistic
  /MISSING=LISTWISE.


GRAPH
  /SCATTERPLOT(BIVAR)=std_dev WITH gem_snel_realistic
  /MISSING=LISTWISE.


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT gem_snel_realistic
  /METHOD=ENTER gender_d fietstype_d leeftijd
  /METHOD=ENTER smooth_perc stop_km
  /METHOD=ENTER perc_BU
  /METHOD=ENTER std_dev
  /PARTIALPLOT ALL
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS DURBIN NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(2.5)
  /SAVE PRED ZPRED ADJPRED MAHAL COOK LEVER ZRESID DRESID SDRESID
SDBETA SDFIT COVRATIO.
```

Appendix 2: explanation python scripts and datasets USB

The usb-stick contains the following:

Final datasets:

- Final dataset in excel format
- Final dataset in .sav fomat (SPSS)
- Thesis in pdf format
- Syntaxes
- Python scripts used (further explanation can be provided on request)