GIMA

Geographical Information Management and Applications

Behavioral cycling profiles and their potential as estimators for cycling motives

Aaron Korver - a.korver@students.uu.nl

06-05-2018

Abstract

Cycling behavioral research is increasingly conducted by means of GPS data. The presence of these data-sets allow for large-scale investigation of complicated travel behavioral aspects such as cycling motives and enables one to enrich raw GPS data-sets with these attributes based on contextual information. Currently, both the differences in cycling behavioral between cyclists with different motives and the extent up to which these differences can be used to estimate cycling motives for raw GPS tracks have received little attention. Even though more insights on these topics can provide useful insights for policymakers and can stimulate travel behavior research by enabling others to enhance their GPS tracks with more accurate cycling motive attribute data. This research tries to tackle both these problems by establishing cycling behavioral profiles based on trip, route and origin-destination behavioral characteristics and subsequently using the differences in these profiles to estimate cycling motives by means of machine learning. In addition to that, multiple machine learning algorithms are assessed to determine the most suitable The results show that there are significant differences in cycling behavioral profiles between motives. Trip, route and origin-destination behavioral characteristics all outperform a standard model for estimating cycling motives, with a combined model including all behavioral characteristics scoring highest (74.0% accuracy versus 51.4% standard model accuracy). Furthermore the results indicate that Random Forest and Gradient Boosting are among the most suitable algorithms for this purpose. Finally, recommendations and potential improvements are provided for future research on cycling behavior and motive estimation.

Acknowledgements

Quite some people were involved in my thesis in different kinds of ways. I would like to specifically mention and thank some of those that provided me with essential support and/or feedback during the process of conducting this research.

Simon Scheider *Utrecht University.* For his helpful and positive attitude as well as his extensive feedback and the time investments he made during the process.

Joost de Kruijf *NHTV / Utrecht University.* For all his feedback and help during the process, his facilitating role in providing me with the necessary data-sets as well as his pragmatic advise for solving some complex problems.

Dirk Bussche *NHTV.* For his feedback regarding technical aspects of this thesis, his facilitating role in providing me with the necessary data-sets, as well as his help with map-matching of GPS tracks.

Jaap Kamminga *Fietsersbond.* For providing me with the Fietsersbond network which yields some important variables that enabled me to conduct specific parts of my analysis.

Contents

1 Introduction

1.1 Context

Cycling as a mode of transport
Cycling is considered one of the most sustainable modes of transport (Klinkenberg and Bertolini, 2012). As cities get more urbanized and traffic congestion increases, cycling provides a green, environment-friendly, low carbon and physically healthy transport modality for short distances (Guo et al., 2013). The Netherlands is one of the leading countries in terms of utilizing cycling as a mode of transport (over 4 billion trips per year, 27 percent of total trip share) (Ministry of Transport, 2009). It is also one of the few countries in which non-motorized transport makes up a substantial part of the total amount of transport (Rietveld and Daniel, 2004). Because of the high amount of cycling trips, the Dutch cycling infrastructure is well developed and documented (Hellmann, 2016). The compactness of most Dutch cities facilitate cycling as an efficient mode of transport but are also a potential pitfall. As Dutch people are increasingly living in cities, the high cycling traffic volume in urban areas causes (bicycle) traffic jams and overcrowded networks, which deteriorates cyclists experiences (KiM, 2016).

The need for knowledge on motive dependent cycling behavior patterns
Just like for any other mode of transport, each cycling trip is initiated for a specific motive (i.e. leisure, sports, cycling from home to work and so on). The fact that an increasing amount of people, with different kinds of cycling motives all use the same cycling network infrastructure can lead to conflicts, accidents and irritations, which discourages people to choose cycling as their (sustainable) mode of transport (Hellmann, 2016). A study conducted by Hull and O'Holleran (2014) has shown that network infrastructure can influence cycling. This turned out to be especially true for the origin and destination of potential journeys. The authors state that: *The cycling infrastructure close to the origin of potential journeys and at the destination is a key facilitator or potential barrier to encouraging cycling* (Hull and O'Holleran, 2014, p. 370). The design and planning of cycling networks can make use of such insights to help stimulate cycling. People might for example be discouraged to use the bicycle to buy their daily groceries if the network between them and the shopping facilities is often crowded or packed with fast-paced groups of sport cyclists. Vice versa, a positive cycling experience encourages cycling. Contrary to car users, who tend to only take 'travel time' into account for their route choices regardless of the trip motive, cyclists tend to take many more aspects into account, all-together commonly grouped as 'route suitability'. (Khatri, 2015). The importance of each aspect that is take into account is also known to vary for different cycling motives (Mcneil, 2012; Guo et al., 2013).
There is a lot of existing research regarding travel behavior of cyclists in general. This data is typically used by policy-makers in order to enhance their infrastructure in data-driven way in order to promote cycling with a maximized return on investments. However, motives-based differences in cycling behavior are rarely included even though, within existing literature, 'cycling motives' of cyclists are mentioned as factors that should be taken into consideration for the development of cycling network plans. It is stated that different cyclists have different behavioral cycling characteristics, which results in different needs. Examples of this are leisure cyclists that tend to prefer a green, comfortable network infrastructure or commuters, who are mainly looking for convenient, unobstructed cycling space (Guo et al., 2013). Even though general differences between cycling behavioral patterns of groups of cyclists with different motives are known, detailed information regarding differences in specific behavioral aspects are missing due to a lack of research on this topic (Guo et al., 2013; Sun and Mobasheri, 2017; Ermagun et al., 2017). This research tries to obtain more knowledge on the

differences in cycling behavioral patterns between cyclists with different cycling motives. The results of this add to the current scientific knowledge on the topic of motive-based cycling behavior research, and could be used to enhance cycling network infrastructures for specific purposes in a data-driven way in order to facilitate their specific behavioral cycling patterns. The results could also be used to help facilitate cyclists with different cycling motives within the same cycling infrastructure, thereby counteracting the potential negative effects of interfering cycling motives.

The need for more knowledge on cycling motives
As stated before: one of the reasons why motive-based differences in cycling behavior are under-researched is the absence of suitable data-sets. There are several reasons for this. Travel motives are harder to distinguish than other travel aspects such as transport modality. Therefore, GPS data-sets generally do not contain motive-related attributes unless they are specifically gathered by for example follow-up interviews. Increased processing power, increased availability of (spatial) data and the rise of smartphones as GPS data collectors have recently enabled researched to derive estimated motives from raw GPS tracks based on quantitative aspects of the trip and other contextual information. However, up to now little of this type of research has been conducted and the accuracy of the derived motives is often far from perfect. There is also no consensus on which types of variables are most suitable for estimating cycling motives. Most studies tend to focus on destination-related properties and/or socio-economic data. The existing studies on this topic are discussed more extensively in Section 2.2 (Guo et al., 2013; Sun and Mobasheri, 2017; Ermagun et al., 2017). This research tries to add to the existing studies by establishing a machine learning framework in order to estimate cycling motives for raw GPS tracks, and is unique in its kind because it is based on the differences in cycling behavioral patterns between cyclists with different motives. The outcomes of this analysis will also provide more insights in cycling behavior, and can be used to enrich existing data-sets with this attribute data, so that more data-sets can be made suitable for motive-based cycling behavior studies.

Investigating cycling behavior for different motives
In order to be tackle the problems mentioned above, it is necessary to investigate the differences in cycling behavior for different cycling motives, using empirical data in which all groups of cyclists with different cycling motives are represented. For this purpose, GPS data will be used. The B-Riders data-set is a bicycle stimulation program performed in the Dutch province Noord-Brabant. This program[1] was originally initiated to generate interest for e-bikes, but contains a lot of valuable GPS-data. The data-set contains over 400.000 GPS tracks of over 700 people, which were all collected between September 2013 and September 2014. The B-Riders data-set is unique in its kind, because it does not only provide GPS tracks, but it also provides contextual information such as socioeconomic data, cycling motive data and bicycle type data (e-bike or regular bike). The fact that cycling motive data is present allows for the unique opportunity to quantitatively investigate differences in cycling behavior between cycling motives. Cycling behavior characteristics can be extracted from the GPS tracks (trip characteristics such as trip length and trip speed) and can be combined with existing data-sources regarding route characteristics, and types of origins and destinations. These sources can be combined into behavioral cycling profiles for each different cycling motive. Subsequently, all this available data can be combined into a model by using machine learning, which is able to learn directly from the input data in order to help estimate additional features such as cycling motives (Society, 2017a). By combining classical statistics with the different machine learning algorithms it can be assessed whether the differences in behavioral cycling profiles among cycling motives can

[1] `http://www.b-riders.nl`

be used to estimate cycling motives. The estimation impact of individual variables, as well as the performance of the entire model will also be assessed. This will give detailed information regarding the influence of each behavioral characteristic.

1.2 Problem Statement

Due to the fact that cycling behavior is generally more diverse than e.g. car-behavior, little is known about the characteristics that influence the behavior of cyclists (Khatri, 2015). More specifically, the differentiation in cycling behavior between cyclists with different types of cycling motives has received very little attention up to this point (Guo et al., 2013). Research on this matter could lead to more insights on cycling behavior of different groups of cyclists, and could therefore be used to enhance existing cycling network infrastructures in a positive way, thereby encouraging people to choose cycling as their way of transport for their specific cycling motive, rather than less sustainable transport options such as cars. Estimation of cycling motives based on cycling behavior characteristics that are derived from raw GPS data will potentially help others to also enrich their GPS data-sets with this attribute data ('cycling motive'), which can contribute towards more availability of data-sets that are suitable for research on the topic of motive-based cycling behavior.

1.3 Relevance

1.3.1 Societal Relevance

In the Netherlands, cycling networks are often heavily used by many different groups with different purposes as well as different cycling characteristics. Cycling characteristics like speed are often very diverse for different groups (i.e. average speed of sports-related cycling versus leisure cycling of elderly; or bike versus e-bike versus speed pedelec). On intensively used network segments, this leads to conflicts, accidents and/or irritations, which negatively influences cycling experiences and could therefore discourage people to cycle and instead switch to less sustainable modes of transport like cars (Hellmann, 2016). Vice versa, a cycling environment which leads to more positive cycling experiences is known to stimulate people to pick cycling as a mode of transport over less sustainable options such as cars. Therefore, from a societal point of view, more detailed insights into network usage for different cycling motives could be used by policymakers to enhance the existing cycling network infrastructure in a positive way in order to promote cycling (Wardman et al., 2007). The outcomes could for example be used to design networks with specific purposes (work, leisure etc.). This is beneficial from a society-perspective, because cycling is a space and energy-efficient as well as an affordable mode of transport for most of all households. It also provides additional health benefits. More cycling means less use of other transport modes such as cars, which have negative side-effects such as air and noise pollution (Vandenbulcke et al., 2009).

1.3.2 Scientific Relevance

From a scientific point of view, the main added value of this research compared to similar research is the use of GPS data rather than more traditional data-gathering methods such as surveys. GPS data is proven to be more accurate data for assessing cycling behavior than traditional data-gathering methods (Richardson et al., 2013; Bohte and Maat, 2009). GPS data also allows one to address large-scale data-samples, since the data is collected automatically and can be pre-processed towards the specific needs for an analysis (Bohte and Maat, 2009). This research can therefore provide new insights into cycling behavior in a data-driven way. There is relatively little scientific work carried

out regarding the differences in cycling behavior for different cycling motives (Guo et al., 2013). The ones that do address this topic mostly use data from countries with low bicycle-usage rates, which are known to produce different outcomes compared to research conducted in countries with high bicycle-usage rates such as the Netherlands (Harms et al., 2014). A larger data-sample can lead to more reliable outcomes compared to a smaller data-sample. Therefore, an extensive Dutch data-set will be used for this research. The outcomes of this research might also be useful in order to create more accurate cycling motive estimation algorithms. Cycling motive estimation has received little attention compared to the estimation of other behavioral characteristics like travel mode (Ermagun et al., 2017). Also, current models that are used to estimate cycling motives have plenty of room for improvements, because they are mostly based on start and end-points of a trip, which makes them inaccurate (see Section 2.2). The implementation of a broader definition of cycling behavior could provide additional context and could lead to more precise identification of cycling motives for GPS cycling tracks. Finally, the focus of this research -the differentiation in cycling behavior for different cycling motives- is unique in its kind and could lead to new knowledge on the topic of cycling behavior in transport, health and sports research.

1.4 Objectives

The main objectives of this research are to identify up to what extent cyclist behavior is differentiated for different cycling motives, and to test to what extent this behavior can be used to help estimate cycling motives for raw GPS tracks. This is achieved by combining existing GI-data-sources. In order to compare 'cyclist behavior' and 'cyclist motives', these concepts have to be defined upfront. Cyclist behavior is defined as the sum of three different cycling behavioral dimensions that can each be measured quantitatively: route, trip and origin-destination. Each dimension is composed of a number of characteristics (factors). Route factors are defined as all attributes of the cycling network that (may) stimulate the use of it (i.e. road traffic, road type, scenery etc.). Trip factors are trip-specific values such as trip length and travel time. Origin-destination characteristics are defined as characteristics of respectively the start and end-point of each trip. These dimensions, that altogether represent 'cyclist behavior', and the characteristics of which these dimensions consist are described more extensively in the theory section of this research (Section 2.1) The other main concept is the concept of 'cycling motives'. Cycling motives are defined as the primary reason to undertake a cycling trip (i.e. work, sport, leisure). Since these are known for the data-set that is used, the effectiveness of different machine learning algorithms can be tested by determining which algorithms yield the most optimal estimation performance. In order to achieve the goals that were stated, the following specific objectives are pursued:

- **To distinguish and identify profiles of cycling behavior for different cycling motives**

In order to assess how cycling behavior is differentiated for different cycling motives, it is important to investigate cycling as a mean to fulfill a specific purpose. This will be done by a study of existing literature. The insights of this literature study are used to determine relevant variables for each cycling behavioral dimension, which can subsequently be used for quantitative analysis. The quantitative analysis is conducted to determine if there are significant differences between different cycling motives for the characteristics that represent 'cycling behavior'. These outcomes are used to establish cycling behavioral profiles for each individual cycling motives. In order to test up to what extent the differences between these profiles can be used to estimate cycling motives, the next goal was formulated:

- **To test the potential of cycling behavior for estimating cycling motives by means of machine learning.**

The estimation power of the cycling behavioral profiles will be assessed by means of machine learning. The outcomes of this provide information regarding the impact of each sub-variable that is used, as well as the estimation power of cycling behavior as a whole, and can be used for future work in order to help estimate cycling motives in order to enrich GPS tracks for which this is unknown. In order to optimize estimation accuracy, the final goal of this research is formulated:

- **To identify which machine learning algorithm(s) is/are most suitable for estimating cycling motives from behavioral patterns**

The last goal of this research is to investigate the influence of using different machine learning algorithm approaches in order to classify the cycling motives. This will be done by identifying relevant machine learning algorithms and comparing the performance of these algorithms for the same task (classifying cycling motives). The outcomes of this assessment can be used to enrich GPS tracks with more accurately generated attribute data in future work.

1.5 Research questions

In order to meet the three goals that were established in Section 1.4, the following main research question has been identified.

"How are behavioral cycling profiles differentiated for different cycling motives, and up to what extent can these differences be used to help estimate cycling motives for GPS tracks by means of different machine learning algorithms?"

This main question can subsequently be divided into several sub-questions which, together, answer the main research question. The following sub-questions have been identified:

1. How are behavioral cycling profiles differentiated for different cycling motives?
2. How can the differences in behavioral cycling profiles be used to help estimate cycling motives for GPS tracks by means of machine learning algorithms?
3. Which machine learning algorithm(s) is/are most suitable for estimating cycling motives from behavioral cycling profiles?

1.6 Constraints

In order to define the scope of this research, the following constraints have been identified upfront.

1. Only cycling data will be taken into account
2. The motives that are provided in the data-set are themselves estimated with up-to-date techniques, but are not based on interviews or self-reports. The motives in the data-set might therefore yield errors themselves too.
3. This research will not contain specific policy-related guidance, but will contain behavioral cycling profiles for different cycling motives that can be applied to real-world problems in order to facilitate cycling for specific groups of cyclists.
4. The data that is provided is (partially) anonymous. Ex post evaluation of their input is therefore no option as participants' GPS tracks are non-retraceable.

5. The data-set that is used originates from a bicycle stimulation program for working people, which implies some constraints regarding representativeness. These are further discussed in Section 3.1.1.

2 Theory

This chapter contains the most relevant theoretical concepts in order to describe processes of current cycling behavior as a means to reach a goal, and to establish a framework of variables, as well as a conceptual model that will subsequently be used for the analysis part of this research)

2.1 Assessing purpose dependent cycling behavior

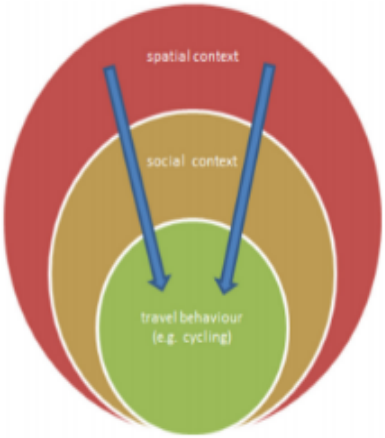
In scientific literature, cycling behavior is a sub-topic within the study field of 'travel behavior', which attempts to study what people do over space, and how they use transport modalities. Travel behavior research typically tries to answer questions like: 'why do people travel?', 'what is the destination of a trip?', 'what route choices do people make and how can travel behavior be changed?' (Wikipedia, 2018). Cycling behavior studies assess the same questions, but they are specifically focused on bicycle-related transport.

Numerous studies have attempted to measure the influence of internal and external factors on travelling behavior of cyclists. Acker et al. (2010) grouped these factors into three different types of contexts: spatial context, socio-economic context and individual context. The spatial context refers to the quality of the built environment, such as infrastructure. The socio-economic addresses the 'quantitative' personal aspects of a person such as age, income, education level and ethnicity. The individual level refers to more subjective personal aspects of a person like his/her perception towards cycling. As figure 1 shows, these contexts do not act independently, but are rather intertwined. Even though socio-economic and individual context is shown to play a role in travel behavior, this research will primarily focus on the influence of the spatial context (the red circle in figure 1). The main reason for this is that while the social context of a person can hardly be influenced directly by policy, the spatial context can be. Also, the spatial dimension is relatively under-researched compared to the other two dimensions (Acker et al., 2010). The existing research that assesses the ability to stimulate cycling trough enhancing aspects of the spatial context is discussed below:

Hull and O'Holleran (2014) have investigated the relationship between network design and cycling levels by the means of interviews. They mention the use of some of the earlier mentioned route dimension characteristics such as wide cycling lanes, direct routes, segregation of traffic, clear signage, safe intersections, high quality surface material, speed barriers, adequate lightness, attractive settings, bicycle parking spaces and end-of-route facilities as main aspects that should be taken into account when enhancing cycling infrastructures in order to stimulate cycling. Their conclusion is that cycling infrastructure design can indeed encourage more cycling. The same conclusion was drawn by Rietveld and Daniel (2004) who assessed the cycling context of the Netherlands in a more quantitative way. As the degree of satisfaction with the cycling infrastructure increases, the share of bicycle use also increases. Segadilha and Sanches (2014) attempted to assess the relative importance of each different cycling factor, as that information was considered essential in order to estimate the value of trade-offs between these factors when designing cycling infrastructures. They categorized relevant factors into five categories: road, traffic, environment, trip and route characteristics. According to their research: motor vehicle speed, security and street lighting were considered the most important factors. However, the fact that this research was conducted in a medium-sized city in Brazil (which has a low share of cyclists), with only a small data sample, implies that the results could be completely different in for example the Netherlands, or for a larger data sample. Nevertheless, all these studies conclude that there is, regardless of the exact factors that are mentioned, influence of network design on cycling levels.

Because of the role of cycling as a sustainable form of transport, governments' attention towards the quality and capacity of the spatial context (e.g. their bicycle dedicated infrastructure) has increased recently. As stated in the introduction, more insights regarding the influence of the cycling behavioral aspects that are part of the spatial context can be used to identify which aspects should be prioritized by governments if they decide to invest money into the enhancement of their cycling infrastructure.

Figure 1: The influence of the spatial context on travel behavior (Acker et al., 2010)



2.1.1 Opportunities for new insights in cycling behavior

The rationality of the cycling network obviously plays a huge role in the mitigation of the contemporary cycling problems that were stated in section 1.1 such as traffic jams, overcrowded networks, and conflicts between cyclists with different motives. Cycling behavior studies try to generate knowledge that can be applied to these problems by using traditional methods and more technical methods such as GPS big data analysis.

Traditional cycling behavior research used to rely on methods such as activity diaries and surveys. For these methods, people are usually asked to reconstruct their travel behavior in hindsight. It has been widely shown that the travel patterns that are collected by this method are systematically different from the actual travel behavior. Respondents usually tend to under-report short trips and/or trips that do not start or end at home. Sample sizes are also typically a problem. The rise of GPS data, and the increasing amounts of available GPS data, is increasingly considered a more adequate replacement of traditional methods. Mainly because it better matches the present data requirements. Various studies confirm the improvements in accuracy of GPS data compared to traditional methods of collecting travel behavior data. It also allows assessment of larger data-samples, because the data is collected in an automated way and can be pre-processed to suit analysis needs (Bohte and Maat, 2009).

2.1.2 Cycling behavioral dimensions

The behavior of a cyclist can be captured by combining information regarding all kinds of properties of a trip which are part of the spatial context of travel behavior as defined by Acker et al. (2010). Examples of these are information regarding the cycling network on which a trip is carried out, contextual information regarding the spatial locations on which a trip is commenced and ended, and aspects of the trip itself such as its duration. The availability of large data-sets like GPS data-sets and land-use data-sets enables us to assess the behavior of cyclists in a data-driven way. According to Bohte and Maat (2009, p. 285), current travel behavior research attempts to capture travel behavior by: *combining data on the location of origins and destinations, trip purpose, trip length, trip duration, departure and arrival times and travel modes in its analyses.".* By using this definition: the aspects of cycling-specific behavior can be grouped into three main behavioral dimensions: trip, route and origin-destination. For each of these so-called behavioral cycling dimensions, a list of factors can be composed for which corresponding variables can be gathered from the GPS data itself, or from alternative data-sources. This subsection gives an overview of the existing work on these dimensions, and gives an overview of the variables that have been used as determinants for cycling behavior. The framework that is derived from this will be used as input for the analysis.

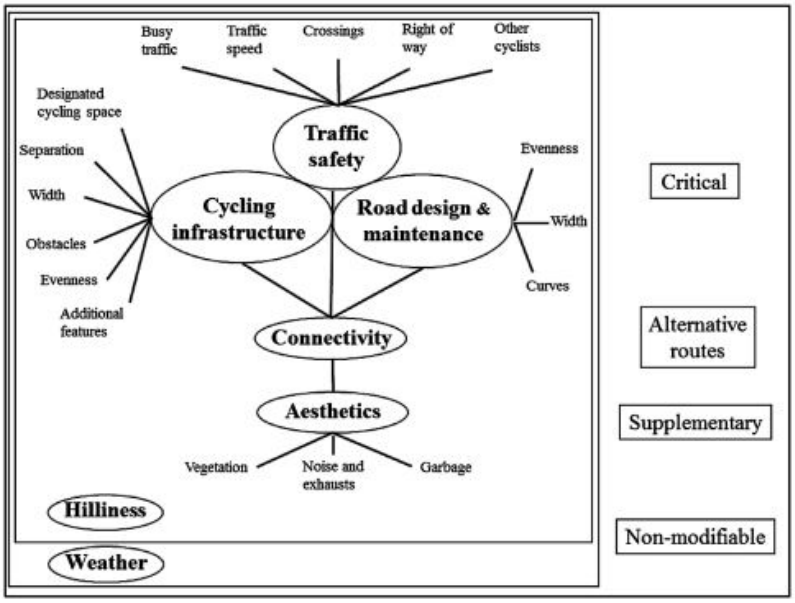
Trip dimension
Trip dimension characteristics are defined as trip-specific values that can be derived from the trip data itself such as trip length and travel time. Speed, and more specifically average trip speed, is a widely used variable in GPS-based cycling behavior studies (Axhausen et al., 2004; Bohte and Maat, 2009). Another variable that is often derived from the date-time information that is attached to the trip data is trip duration (Bao et al., 2017; Axhausen et al., 2004; Bohte and Maat, 2009; Schönfelder et al., 2002). Finally, even though less commonly used, trip length can serve as a characteristic of the trip dimension (Bohte and Maat, 2009; Khatri, 2015).

Route dimension
Cycling behavior is known to be influenced by the physical environment. The physical environment is defined as: *the objective and perceived attributes of the context in which people spend their time (e.g. home, neighborhood), including elements of urban design (e.g. presence of cycling paths), traffic density and speed, distance to and design of venues for physical activity (e.g. parks), crime, safety and weather conditions* (Cauwenberg et al., 2018, p. 38).
Route characteristics are main contributors to the objective and the perceived physical environment. Therefore, route dimension characteristics are defined as: all attributes of the cycling route that (may) stimulate the use of it. A first route characteristic is the presence of dangerous crossings. Cyclists are known to be willing to cycle longer distances if it allows them to avoid unsafe network segments such as crossings (Gaterslebena and Appleton, 2007).
Another factor is the presence of separated cycling lanes. Research conducted by Wardman et al. (2007) concluded that, for their data sample, travel time on separated cycling lines was valued only 14% of the travel time on a non-separated road. This translates to a willingness to travel over 3 more minutes on a 20 minute trip, if it means that the trip can be made on a separated cycling line instead of a non-separated road. This is in line with the more general observation that cyclists are risk-averse in their route selection (Gaterslebena and Appleton, 2007). The influence of the prominence of bike paths on the actual willingness to cycle has also been shown by Hunt and Abraham (2007) and Broach et al. (2012). Maximum speed on routes is also known as an attribute that influences cycling behavior. It is shown that cyclists prefer roads with lower speed, especially a lower speed for motorized vehicles. This relationship is more prominent for inexperienced cyclists,

but is also there for experienced cyclists (Broach et al., 2012).
Signalized intersections such as traffic lights and stopping signs also have a negative influence on cycling experience. This has to do with two factors: it generates delays and it requires people to stop and speed up again, which is energy-consuming compared to a steady pace (Broach et al., 2012). Other literature sources provide contradicting results. Khatri (2015) mentions the amount of signalized intersections as a positive factor in terms of route selection. In his point of view, signalized intersections allow for safer and more protected crossings of roadways with high traffic intensity.
Another factor that is mentioned in literature is the traffic volume on the network. The preference of cyclists for low traffic volume local roads and off-street roads over arterial and main roads has been shown by Winters et al. (2010). Respondents that were interviewed in the same study also mentioned the presence of appealing scenery as a positive aspect of cycling routes. The attractiveness of the route setting is also shown to positively enhance cycling experiences. Hull and O'Holleran (2014) mention greenery as one of the examples of attractive attributes of the environment. This factor is also mentioned by Cauwenberg et al. (2018), who also addresses two factors that contribute towards the attractiveness of settings in a negative way: noise and litter. Several environmental characteristics have also been shown to influence bicycle use. Vandenbulcke et al. (2009) mention gradient, weather, climatic conditions, urban spatial structure and infrastructure as main factors. In this context extreme temperatures, rain and high wind speeds decrease the willingness to cycle, whereas moderate temperature, dryness and low wind speed positively enhance the willingness to cycle.

Cauwenberg et al. (2018) made a more generic assessment of the influence of the route and its environment on cycling experience. In this study, which focused specifically on elderly cyclists, they grouped the resulting variables, which are all potential route dimension characteristics, into influential (i.e. traffic speed) and non-influential (i.e. weather) variables. The results of this are shown in figure 2.

Figure 2: Environmental factors influencing cycling experience (Cauwenberg et al., 2018)



Origin-destination dimension
The origin-destination dimension consists of attribute data regarding the origin of a trip, and the destination of a trip. This dimension is commonly used when trying to impute additional characteristics for a data-set because this dimension possesses a lot of contextual information. There is a wide variety of methods available to gather origin-destination characteristics for GPS data-sets. These methods all combine the start and/or end point of a trip with contextual data-sets such as land-use data-sets, parcel information databases, points of interest, travel regularities, social geo-data and follow-up questionnaires. Some researchers consider the start and end-point of a trip as point data, whereas other researchers generate a buffer around the start and end-point of a trip in order to measure the presence of for example points of interest near these points (Schönfelder et al., 2002; Hasan and Ukkusuri, 2014; Bohte and Maat, 2009; Bao et al., 2017; Axhausen et al., 2004).

2.2 Estimating transport motives

2.2.1 Estimation of motives based on GPS tracks

The estimation of transport motives from GPS tracks is a major research topic within the domain of activity-travel behavioral analysis (Feng and Timmermans, 2014b). This because it is generally accepted that post-processing of these data-sets can provide essential information for the research domain of 'travel behavior analysis', and can provide new insights in (social) processes that were poorly understood and/or under-sampled until recently (Axhausen et al., 2004; Romanillos et al., 2016). GPS tracks by itself are 'just' a chronological collection of coordinates but, when combined with other data sources, can be used to generate meaningful knowledge regarding travel behavior.
A known issue when estimating activity types is that there might be only weak relationships between activity types and activity locations. For example: people can use a shopping mall to buy

groceries (daily shopping), but also to go out for lunch (leisure). Most methods that are used to estimate transport motives only use one source of information in order to make estimations. Feng and Timmermans (2014b) propose the use of multiple different sources of information to make the calculations more precise. Personal information, socioeconomic information and aggregated trip patterns can be used to achieve this. However, this information is often either not present in data-sets or not suitable for use due to privacy issues. A type of information that is always present in a GPS data-set is the GPS data itself. If cycling behavior dimensions route, trip and/or origin-destination characteristics are shown to be related to cycling motives in a certain way, this information can be used in addition to the existing transport motive estimation methodology. The main advantages of using route characteristics as a data-source for estimating transport motives would be that it does not rely on potentially privacy-sensitive data, and that it is generally applicable to all GPS tracks for which route characteristics are available.

2.2.2 Existing work on GPS-based cycling motive estimation

Recently there has been quite some scientific work on imputation and validation on GPS-based motive estimations. Existing studies that focus on cycling motive detection generally use a rule-based, probabilistic or a machine-learning approach (Schönfelder et al., 2002; Hasan and Ukkusuri, 2014; Bohte and Maat, 2009; Bao et al., 2017; Axhausen et al., 2004). A comparison of existing studies showed that machine learning algorithms generally outperform rule-based approached in terms of accuracy (Xiao et al., 2016; Ermagun et al., 2017). Of course this does not mean that machine learning is objectively superior for this purpose since data-input, which plays a large role, is different for each of the studies that were compared. Even though some of these studies are focused on car travel behavior, their approached can still be relevant for this research. This subsection assesses these approaches and the corresponding results.

Axhausen et al. (2004) conducted a study on a data-set that consists of 240.000 GPS-tracks that were gathered by GPS-based car-units. They used a multi-stage approach in order to automatically impute cycling motives. They combined factors that could be derived directly from the GPS tracks (origin-destination location, travel duration, date-time information) with survey information (age, education, profession, working hours) and external data-sources like land-use maps and parcel information databases. They created clusters around points of interests like the home-location of a person, a restaurant or a gas station. Distance was used to estimate the likelihood of visiting specific locations. The likelihood of each of the potential purposes (work, pick-up, school, business, daily shopping, non-daily shopping, home, leisure, other) was used to assign purposes to trips. Unfortunately they did not validate their data. Still, they noticed that related motives (for example shopping and leisure when visiting a shopping mall) are harder to distinguish than relatively unrelated motives (for example work and leisure).

Feng and Timmermans (2014a) used GPS tracks in order to estimate cycling motives of e-bike users. They created a framework that combines information of the surrounding of origin and destination of a trip (based on OpenStreetMap features), trip characteristics like trip time and speed and frequency patterns (number of visits to a specific location) in order to assess the motive of a trip. They used travel diaries which were manually submitted by participants in the GPS study in order to validate their results. OpenStreetMap was used because it is the most accurate data-set that is available on a global scale according to the authors. The final input list for their model contained the following variables: percentage retail area, percentage commercial area, percentage industrial area, percentage residential area, percentage leisure area, duration of activity, travel time, activity start

time, the frequency of activity locations within zones, and the day of the week which is calculated as the ratio between the number of weekdays and weekends. (Feng and Timmermans, 2014a, p. 9). They also assessed model quality by measuring the accuracy of their Bayesian machine learning model using different groups of variables as input (temporal data, spatial data, frequency data or a combination of these data-types). Their combined model performed best (a mean accuracy of 67,4%), whereas frequency data turned out to be the worst estimator (a mean accuracy of 32,1%). Other machine learning algorithms were not assessed

Bao et al. (2017) used GPS data in order to determine cycling motives of users of bike-sharing systems. The data output of the about 5500 bike-sharing bikes, which are equipped with GPS sensors, were combined with points of interest (POI's) derived from Google Places, and data that was collected on a smart-card. This smart-card contained the trip characteristics (trip duration, start and end-time), individual characteristics (user type, gender, year of birth) and information related to the specific station where the bike was obtained. Home, eating, leisure, shopping, transport and education were derived as potential motives. Model fitness was tested for a model with and without POI data. The results of this indicated that POI's are a relevant variable for imputation of these specific cycling motives. The authors conclude that more efforts are needed to accurately impute cycling motives. They suggest the use of social media data, socio-demographic data and other additional data sources for future work.

Bohte and Maat (2009) combined GPS logs, GIS technology and a web-based validation application in order to derive and validate cycling motives and travel modes (car, bicycle or walking). In order to do so, they used the data of 1104 respondents that participated in a one-week study. GIS data-sets, characteristics of respondents and the GPS data itself were combined in order to estimate cycling motives and travel modes. These results were subsequently validated by participants in the web-based validation application. Travel modes could be estimated quite accurately (68%-75%). The accuracy of the cycling motives was mostly lower and ranged from 4% for 'study' up to 74% for estimating 'home' cycling motives. The researchers indicate that while possibilities for imputing characteristics like cycling motives are being explored, almost every existing study has opportunities for improvement. They propose to combine multiple aspects of travel behavior because, due to the increasingly varying travel patterns of individuals, single aspects are no longer sufficient.

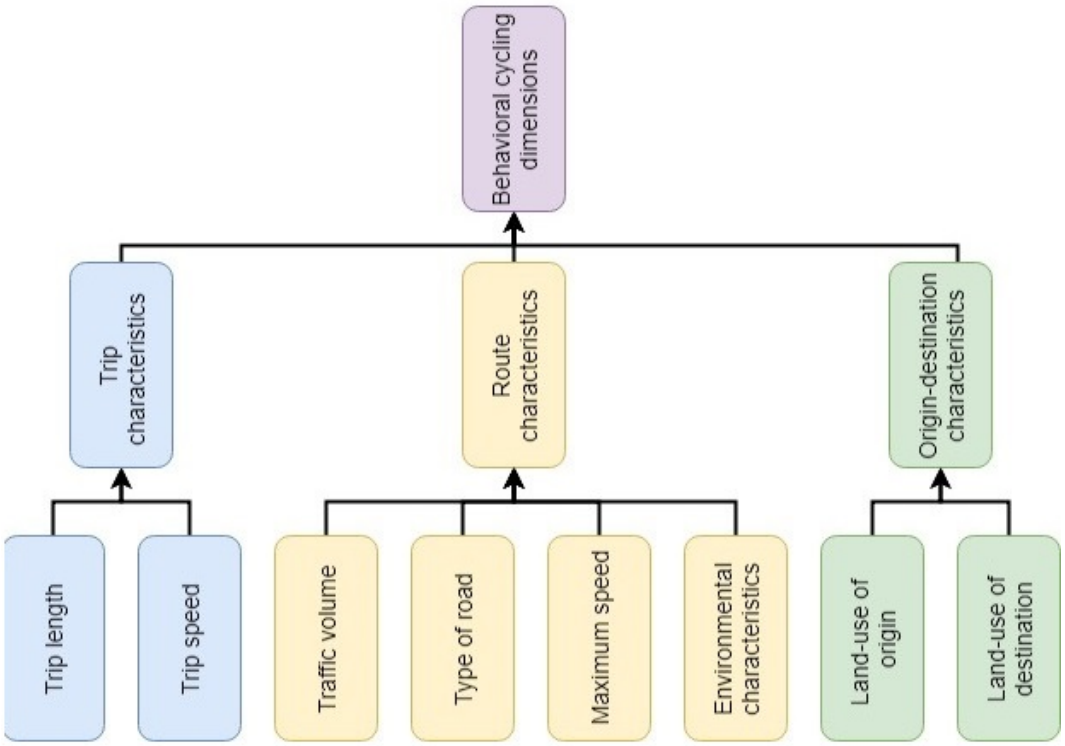
Hasan and Ukkusuri (2014) propose the use of topic modelling based on geo-location data from social media in order to: *"develop methodologies to understand individual activity patterns using large-scale location-based data"* (Hasan and Ukkusuri, 2014, p. 364). They tried to categorize trips as: home, work, eating, entertainment, recreation, shopping, education or social service. Even though their outcomes do not have sufficient explanatory power for the motives, they propose that as more geo-location data becomes available in the future, it might yield increasing potential for purpose imputation and/or to explain behavioral travel patterns of people.

Schönfelder et al. (2002) created a framework that could be used to impute cycling motive based on daily regularities of travelling, using a combination of trip-end with land-use data, as well as interviews. However, they did not manage to automatize this process (which is not that strange given the year of publication (2002)) and could therefore not provide concrete results due to the fact that their sample size was too large to manually process (over 200.000 trips). The challenges that they could not overcome are: the availability of land-use data (which was limited for their study area), low accuracy of the available GPS data, and the lack of the ability to follow-up on respondents with additional surveys/questions.

Altogether one can conclude that the most successful examples of motive imputation all contain a combination of different data-sources, including but not limited to route information, origin-destination information and additional route-related information. Social geo-location information is suggested as an opportunity, but within existing research successful applications of this data-source are scarce.

2.3 Conceptual Model

The conceptual model below shows how the variables that have been addressed in the context and the theory chapters relate to each other. It is expected that the values of the identified route, trip and origin-destination behavioral dimensions of a trip influence the motive of that trip.
The next chapter will provide a methodological framework for the empirical part of this research in order to be able to assess the assumed relationships.



3 Methodology

The methodology chapter is structured as a step-by-step plan to give a global overview of how the empirical part of this research is going to be carried out. Data selection, preparation, quality and the proposed data analysis methodology will be covered for each main type of data. Finally, the different types of analysis will be covered.

3.1 Data selection, preparation and quality

3.1.1 GPS Cycling tracks

The GPS cycling tracks will be extracted from the B-Riders data-set. The B-Riders data-set is a bicycle stimulation program in Brabant, a province within the Netherlands, that pays volunteers in order to cycle to work, rather than using their cars. The program is a community-focused initiative that is centered around reducing congestion by car. Participants received a small compensation fee per kilometer that was logged by their smart-phones, using a GPS application that was specifically developed for this program. Due to the fact that the program is centered around people who use their bicycle to go to work, certain socio-economic groups are over-represented (people of working age), whereas other groups are under-represented (elderly). However, since this research focuses on groups with different cycling motives, rather than groups with different socio-economic characteristics, this should not influence the outcomes. There are over 700 participants who have contributed to the data-set who, together, have generated over 6.1 million kilometres of GPS data tracks[2]. For this study a subset of the months January, July and September 2014 is used. Also, not every GPS track in this time-span is used for this analysis: only the trips of which at least 1 segment was carried out within the borders of the municipality of Tilburg are included. The reason for this is that the Fietsersbond cycling network data, which is used to enrich tracks with route characteristics, is specifically made available for this municipality by the owners of the data-set. The GPS points are collected by smartphones and are considered accuracy enough for travel behavior-related research (Feng and Timmermans, 2014a). The GPS tracks are enriched with information regarding the accuracy of each GPS measurement, which allows us to filter out inaccurate measurements. The GPS tracks can subsequently be map-matched to the cycling network that is used. The following sub-paragraph discusses the process of map-matching.

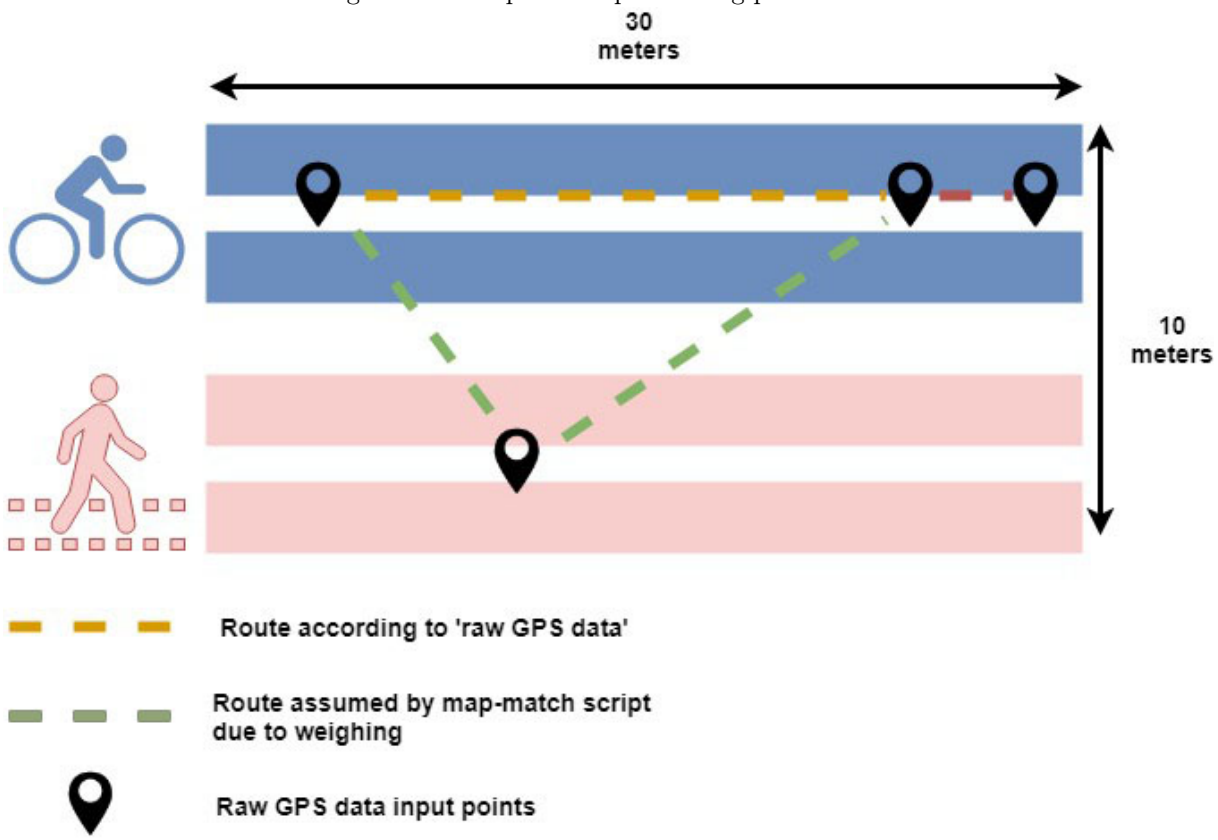
3.1.2 Map Matching

For the map-matching of the data set an external algorithm will be used (Bussche and van de Coevering, 2015). Because of the fact that not the original, but map-matched GPS tracks are used to derive route characteristics like 'type of road' as input variables for the analysis, it is important to describe the methodology that has been used to map-match the GPS tracks. As described before, GPS coordinates that were gathered 'in-field' always yield measurement errors. In a dense urban environment in which cycling lanes, footpaths and car lanes are often literally constructed next to each other, this can result in inconsistencies and errors while assigning GPS tracks to network segments. The map-matching script that was used tries to minimize this error margin by applying least cost path analysis. The cycling routes are map-matched one by one. The 'raw' GPS data of a cycling route consists of a list of GPS point data entries. Each link (road segment) in the network is assigned a resistance value which is higher as it is more distant from the GPS entry points. Subsequently, the route of least resistance is determined. Due to the nature of this method, the 'map-matched' route will always closely resemble the GPS route. However, the resistance method

[2]www.b-riders.nl

allows to subtract weight to segments that are more likely to facilitate cycling (such as cycling-only roads), to add weight to segments that are unattractive to cyclists (such as service roads), or to add an infinite weight to segments that do not facilitate cyclists at all (such as public highways). This results in a logical 'cycle-able' route, which still closely resembles the 'raw' input data. Figure 3 provides an example of the consequences of weighing the road segments.

Figure 3: Example of map-matching process



In this example the raw GPS input point data indicates that a (small) fraction of the route was carried out on a footpath (in pink). Creating a route line polygon based on just the input data would result in the route that is displayed in Figure 3 in green. However, because of the fact that the specific cycling road has a lower weighting factor (because it is more likely to be used by cyclists rather than footpaths) within the map matching process, and therefore causes less 'resistance' even though it (slightly) deviates from the original raw GPS entry points, the map-matching script will consider the route in orange as the more likely route and will therefore clip the entire length of this example route poly-line onto the cycling lane. This does not mean that the map-matching script will automatically clip routes onto cycling paths when they are nearby. If the large majority of GPS entry points is on a side-walk, the track will also be map-matched to the side-walk path. The weighing just ensures that in the cases where there is reason to doubt, the most logical outcome will be selected. This removes most of the inconsistencies in the output data, and it generates more accurate output data. Some routes simply cannot be map-matched to the infrastructural network

because the network data is missing (for example when parts of the route are carried out in another country), or when no single connecting route can be constructed based on the GPS entry points and the weighted network. In these cases (which have an occurrence rate of less than 1%) the routes are dismissed.

3.1.3 Cycling network

The cycling network that will be used is the Fietsersbond cycling network, which is based on voluntarily generated geographic information and is considered accurate, and comprehensive. Because it is updated by volunteers all over the Netherlands, it is also one of the most up-to-date data sources. In order to prepare the data, the GPS data of the B-Riders data-set will be map-matched to the Fietsersbond cycling network. The quality of this data-set is considered sufficient. The Fietsersbond cycling network was selected over alternatives like OpenStreetMap because of its additional attribute data. It yields attributes such as road type, maximum speed, traffic obstruction, lightning, attractiveness and maximum speed, of which some are used as route behavioral characteristics (Fietsersbond, 2017).

3.1.4 Cycling motives

The cycling motives are defined as a separate column within the GPS tracks data-set that is used. The motives are split into eight different categories: 'services', 'home', 'paid-work', 'non-daily-shopping', 'daily-shopping', 'social', 'leisure' and 'recreation'. These will also used in the analysis. The cycling motives are already prepared for in the data-set that is used, and do therefore not require further preparation. It is important to mention the fact that the motives themselves are based on an algorithm. This will yield an error component for the resulting model that will be composed to estimate cycling motives. However, this type of error is non-systematic and will therefore not influence the actual outcomes, but will only add additional noise to the resulting model.
The algorithms used to define the motive for each cycling trip are well-described by Feng and Timmermans (2014b) and are covered in this research in order to provide background information regarding the ways in which cycling motives are defined for the used data-set. The main differences between the ways these motives were assigned, and the ways in which this research attempts to assign motives are the differences of used data. Feng and Timmermans (2014b) have based their work mainly on start- and end points of trips, whereas this research attempts to include the differentiation in cycling behavior between cycling motives, which has not been done so far.
The motives are assessed on two levels: raw and adjusted. Raw motives are the 7 categories that were mentioned earlier, whereas adjusted motives represent grouped motives that represent 'utilitarian' trips or 'non-utilitarian' trips. Paid work, daily-shopping, non-daily shopping, services and study are considered utilitarian motives. The other raw motives are considered non-utilitarian. This is a common practice within travel behavior research, because larger differences should be expected for the latter categories due to their prevalent differences. Fernandez-Heredia (2014) even state that: "*When conducting a study about bicycles, it is necessary to distinguish forced mobility from mobility for sports or recreational and leisure pursuits, as the behavior and decisions made by cyclists differ completely depending on the purpose of their bicycle journeys*' (p. 3). Therefore, grouping motives based on these characteristics could result in more distinguishable cycling behavior profiles and might therefore make the estimation process more accurate (Wegener, 2013; Meloni et al., 2004).

3.1.5 Behavioral cycling dimensions

Route characteristics
The route characteristics that are going to be used for this research are extracted from theory (see: Section 2.1.2). Factors which are known to or expected to influence cycling behavior are included if the data turned out to be available and the quality of this data is considered sufficient. Table 1) gives an overview of the variables that were selected.
Because of the 'immeasurability' of emotions such as 'appreciation', the appreciation of the environment was measured using two variables: amount of green (objective), and hand-reported appreciation of the environment (subjective) The data preparation of the networks factors varies for each factor (this is described more extensively at the end of this sub-chapter).

Table 1: List of route characteristics

Factor	**Type of expected effect**	**Literature**
Type of road (bike lane vs non bike lane)	Positive	Broach et al. (2012)
Maximum speed	Negative	Heinen (2011)
Characteristics of the environment	Positive	Segadilha and Sanches (2014) Winters et al. (2010)
Appreciation of the environment	Positive	Segadilha and Sanches (2014) Winters et al. (2010)
Traffic volume	Negative	Segadilha and Sanches (2014)

Trip characteristics
The main trip characteristics that will be used are trip length and travel time. Trip length will be calculated by calculating the sum of the length of the segments of which a route is composed. Travel time will be calculated by calculating the difference in seconds between the first and the last date-time value of a trip.

Origin-destination factors
The origin, as well as the destination of each trip, is taken into account. These are extracted from the GPS tracks by selecting the first and last GPS measure point with an acceptable measurement error. In order to identify the type of origin or destination, the identified start or end-point of each track is intersected with the 'Bestand Bodem Gebruik' (BBG), which is the most comprehensive data-source regarding the function of land for the Netherlands (CBS, 2012). The main categories are: traffic related, built-environment, semi-built-environment, recreational, agricultural, forest, backwater, open water and foreign. These are further divided into sub-categories which are described extensively in the documentation of the data-set [3]. The resulting values for each trip are added to the data-set in two new columns ('startbbg' and 'endbbg').
Since GPS is known for its variety of accuracy (especially when devices are just switched on), only GPS measure points with a potential measurement error of less than 10 meters are taken into account for this process. This value is selected to minimize the chance of a falsely identified origin or destination types.

Pre-processing of behavioral cycling dimensions
The map-matched data-set was imported into PostgreSQL, after which the data of the individual

[3] https://www.cbs.nl/NR/rdonlyres/5B353A4E-6B56-4756-A19C-4BF6AA520C65/0/BestandBodemgebruikProductbeschrijving.pdf

tracks were converted into shapefiles by using the following query: The Fietsersbond network consists of a large sum of linked segments which, together, make up the cycling network. Each segment has its own attribute data which describes the type of road, By making selections of all rows (in which each row represents one segment of the cycling network) for each unique routeid, it is possible to reconstruct all individual routes. Subsequently, the resulting shape-files are enriched with the Fietsersbond cycling network attribute data. Also, the length of each segment is calculated and added to the attribute table, so that the 'averages' for each variable can be weighed based on segment lengths. The result is exported as a CSV file using the 'Export Feature to ASCII' tool. This file is imported into Python so that the attribute data, which consisted out of string data for the attribute fields can be prepared for the analysis. The variables are then re-coded into quantitative values using Python [4] and the re-code-chart that is displayed below (Table 2) .

[4]https://github.com/AKorver/ThesisGIMA

Table 2: Pre-processing of variables for quantitative analysis

Behavioral cycling dimension	**Variable**	**Raw data values**	**Pre-processed data values**
Trip	Trip duration	Timestamp for each GPS entry point	Time between first and Last GPS entry point
Trip	Trip length	Segment length for each segment	Sum of segment lengths
Route	Type of road	'normale weg' 'ventweg' 'bromfietspad (langs weg)' 'solitairbromfietspad' 'voetgangersdoorsteekje' 'veerpont' 'voetgangersgebied' 'weg met fiets(suggestie)strook' 'fietspad (langs weg)' 'solitair fietspad' 'fietsstraat'	1: Non-Cycling-specific: 'normale weg' 'ventweg' 'bromfietspad (langs weg)' 'solitair bromfietspad' 'voetgangersdoorsteekje' 'veerpont' 'voetgangersgebied' 2: Cycling-specific 'weg met fiets(suggestie)strook' 'fietspad (langs weg)' 'solitair fietspad' 'fietsstraat'
Route	Average max speed	Maximum speed per segment	Sum (Max_Speed * Segment Length) / Sum of Segment Lengths
Route	Average traffic volume	Intensity per segment	Sum (Intensity * Segment Length) / Sum of Segment Lengths
Route	Environmental score	'bos' 'natuur (behalve bos)' 'bebouwd (veel groen) 'landelijk of dorps' 'akkers/weilanden' 'bebouwd '(weinig of geen groen)'	3: 'bos' 'natuur (behalve bos)' 2: 'bebouwd (veel groen) 'landelijk of dorps' 'akkers/weilanden' 1: 'bebouwd(weinig of geen groen)'
Route	Appreciation of environment	'mooi' 'schilderachtig' 'neutraal' 'lelijk/saai' 'zeer lelijk'	4: 'mooi' 'schilderachtig' 3:'neutraal' 2: 'lelijk/saai' 1: 'zeer lelijk'\end{tabular}
Origin-destination	Origin	First raw GPS point of a track with accuracy 10m	Intersected BBG land-use Category (i.e. 'housing')
Origin-destination	Destination	Last raw GPS point of a track with accuracy 10m	Intersected BBG land-use Category (i.e. 'housing')

3.2 Statistical testing

3.2.1 ANOVA Analysis

The goal of this type of statistical testing is to try to test if the trip and route characteristics are significantly different for different cycling motives. In order to do so analysis of variance (ANOVA) is used (Field, 2009). For this purpose the SciPY package is used [5]. The trip and route characteristics that have been described before are calculated for each different cycling motive. Subsequently, ANOVA is applied to test whether there is significant difference between groups (based on cycling motive). In this analysis the variable cycling motive is used as grouping variable, whereas the cycling behavior dimensions function as testing variables. It is important to emphasize that the factors that are analyzed are only one aspect of a larger range of factors that explain why people cycle specific routes. This methodology assesses the influence of cycling behavioral characteristics on motives, independently from other influences. This is referred to as decomposition of variance. The outcomes of this can be used to (formula-wise) describe the differences of route characteristics for each of the cycling motives.

3.2.2 Chi square

Chi-square is used to test if the BBG categories that are assigned as origin and destination of a trip are significantly different for different cycling motives (Field, 2009).

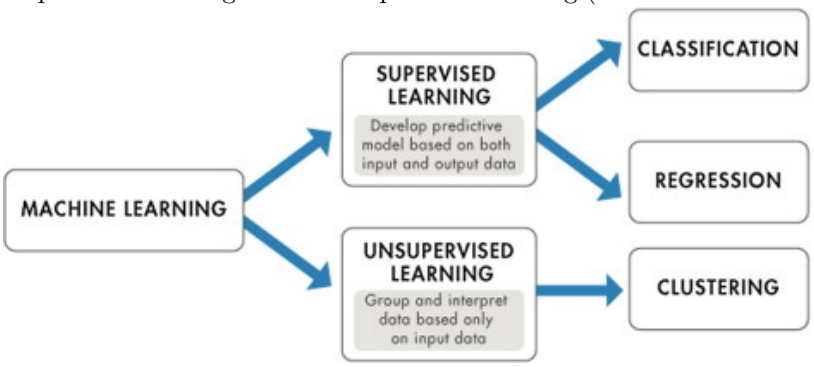
3.3 Estimation of cycling motives: machine learning

3.3.1 Machine learning as a method for estimating transport motives

Machine learning is defined as: "the technology that allows systems to learn directly from examples, data, and experience." (Society, 2017a). The main aspect of machine learning that distinguishes it from other kind of quantitative research methods is that is allows the system to directly learn from that data that is used as input, rather than using predefined rules.
Machine learning algorithms can be subdivided into two main branches: supervised learning and unsupervised learning. Supervised refers to machine learning systems which try to use existing data for which the outcome is known, in order to learn a function that can be applied to new 'test' data. Unsupervised refers to machine learning algorithms that use data without known outcomes. They try to detect data points with similar characteristics (i.e. clusters) and subsequently assign data to these clusters. (Society, 2017a).

[5] http://www.scipy-lectures.org/packages/statistics/index.htmlpost-hoc-hypothesis-testing-analysis-of-variance-anova

Figure 4: Supervised learning versus unsupervised learning (Bunker and Thabtah, 2017)



Since transport motives are given in the data, supervised machine learning is the most suitable method for estimating transport motives. However, within the group of supervised machine learning algorithms, there are still a large number of individual methods that each have their own distinguishing ways in which they try to tackle the same problem (finding statistical correlations in training data that can be used to help estimate features for test-data-sets). Because the nature of the methods is often diffuse, the results of different kind of methodologies can also yield diffuse results when given the same task. Different variables can have different relationships (i.e. linear, inverse linear, hyperbolic etc.). When different classification methods are tested in order to tackle a certain classification problem, one is able to identify which machine learning algorithm is most suited for the purpose. The most suitable methodologies for the purpose of this research (machine learning based classification) are covered briefly below.

Logistic Regression
Logistic regression is a simple classification method that is most suitable for binary classification problems (target variables with 2 possible classes). It is able to fit most linear relationships and is one of the most frequently used machine learning algorithms. Because of the fact that it also fits non straight linear relationships, it is more effective in assigning adequate weight to values which lie within the decision frontier compared to values that are more extreme, and is therefore also more suitable for estimation purposes compared to linear regression (SKLearn, 2018c).

Decision Tree Classifier
Decision trees are a non-parametric method that is used for regression and classification purposes. It attempts to establish a model which can estimate the target variables by decision rules which are inferred from the available data-input. It is easy to use and a well-performing method in many situations. However, one of the problems of this method is that it sometimes over-fits statistical relationships (SKLearn, 2018a).

Random Forest Classifier
Random Forest classifier is en ensemble method. Ensemble methods are machine learning algorithms that combine different algorithms in an attempt to outperform the score of a single method. The way random forest classifying does this is by generating multiple decision trees and uses the average scores of these trees in order to make decisions. This also leads to a reduced amount of variance (and a higher potential for bias) (SKLearn, 2018f).

Support Vector Machine
Support vector machine is a method that is used for classification and regression. It tries to maximize the chance of correct estimations for new data by maximizing the margins between input data and the 'classification boundary' for the training data-set. It is most accurate if training and test data are at least somewhat related (SKLearn, 2018g).

K-nearest neighbor
K-nearest neighbor is a method that uses the values of a predefined number of nearest neighbors in order to estimate the value of a test data entry. It is a non-parametric method and may therefore to able to be successful for specific classifying purposes where parametric classifying methods are non-applicable (SKLearn, 2018b).

Naïve Bayes Naive Bayes is a conditional probabilistic model. It relies on the assumption of independence of variables and will therefore yield incorrect results if (sub)-variables are interrelated. It only requires small samples for accurate results and is considered surprisingly effective given its simplicity, especially for tasks like document classification (SKLearn, 2018d).

Neural Network
Neural network machine learning algorithms like the Multi-layer perceptron algorithm are versatile methodologies that are able to capture non-linear relationships, which makes it a popular method for estimation purposes. It uses hidden layers (layers in-between input and output) that are used to distill the most important patterns from the input data (SKLearn, 2018e).

Gradient Boosting (GBM)
Gradient boosting is another ensemble method that combines multiple 'weaker' classifiers into one ensemble method.It supports both binary and higher level classifications and typically yields a high estimation power (SKLearn, 2018h).

3.3.2 Machine learning: methodology

If distinguishing trip, route and origin-destination characteristic are identified, they can be used to assign motives to GPS tracks more accurately. This will be done by machine learning. Machine learning is a type of methodology that allows computers to self-learn from examples. It uses large data samples to learn from patterns in order to for example make decisions or estimate activities (Society, 2017b). There is a wide range of different types of problems that can be solved by machine learning, which are grouped into different types of 'canonical problems'. As stated in Section 3.3.1, a 'classification problem' has be addressed. The type of machine learning that is used for this research,'supervised machine learning', has to be trained with labelled data. The structure of this labelled data is then used to estimate categories of new data. In order to achieve this, the data-set will be split up in training and testing data. This will be done iteratively to obtain accurate results and rule out the influence of test/training data-selection (Society, 2017b). In order to test the effectiveness of the different machine learning classifying methodologies that were discussed in the theory chapter (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neigbor, Naive Bayes, Neural Network) thoroughly, they will all be used for the analysis in order to determine the most accurate method for this purpose. In the analysis, the different cycling motives will serve as classification options, whereas the trip, route and origin-destination characteristics will

be used to identify similar patterns in order to classify each cycling track. The python package that will be used, SciKit learn, is open-source and contains all the required functionalities for this purpose and for the different types of machine learning algorithms that are used [6]. First, Patsy is used to create a pandas data matrix in which the cycling motive serves as y-value and the 'estimation' data (the network and origin-destination factors) serve as X-values. Subsequently, the cycling motive data is converted into a 1D numpy array which makes it suitable for machine learning. The accuracy, null-error and a 10-fold cross-validated mean accuracy, precision and recall are calculated in order to assess the consistency of the estimations, as well as its performance compared to the null-error. The reasons why 10-fold cross-validation is used and these values are calculated is discussed extensively in the sub-section below.

3.3.3 Over and under-fitting

Any machine learning algorithm is established in order to fit any relationships within training data-sets as well as possible. Because the algorithms are based on the training data, the errors of estimations will always be lower for training data than for 'new' data. As the complexity of a model increases (for example by adding new variables to the model) the degree up to which the model fits the training data also increases. This is a fundamental aspect of machine learning models. This holds true up to an extent where even a variable that yields a randomly generated numeric value will increase the degree up to which machine learning algorithms are able to fit the model. As more and more variables are added (increasing the complexity of the model) the fitting increases up to a point where the modelled relationship fits the 'data relationship' of the training data almost perfectly, but does not represent the actual relationships between variables anymore. This will decrease the error margin for the training data-set, but will usually *increase* the estimation errors for new test data. This relationship -that applies to both classification and regression- is illustrated in figure 5, in which a modeled relationship between wealth and happiness is used as an example.

[6] http://scikit-learn.org/stable/modules/linear$_m odel.html logistic - regression$

Figure 5: Illustration of the relationship between model complexity and model estimation error (Fortmann-Roe, 2018b)

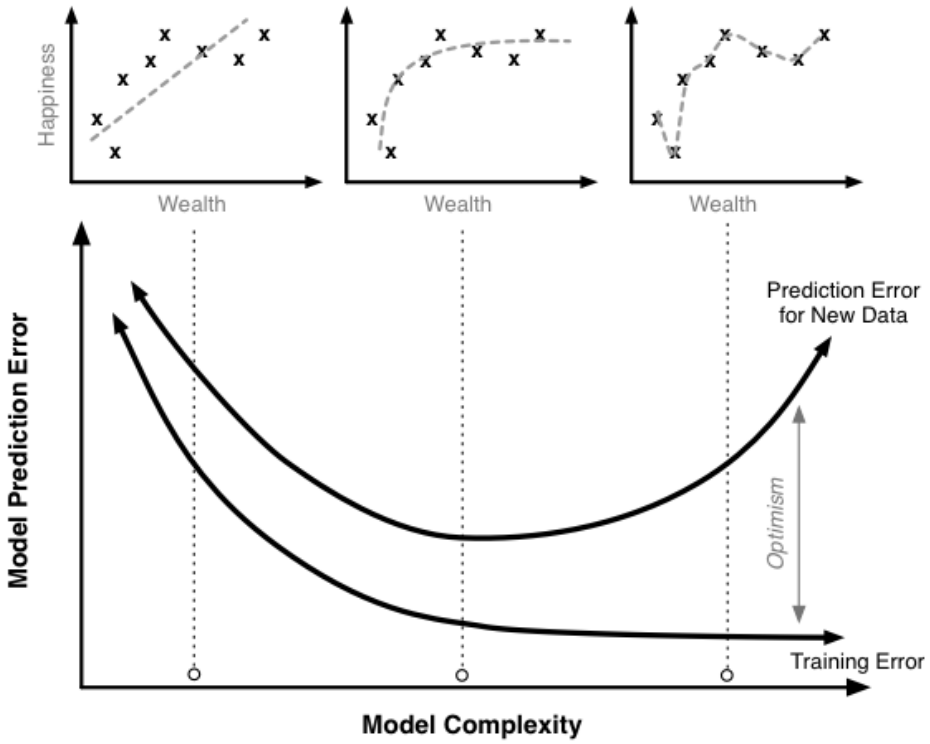


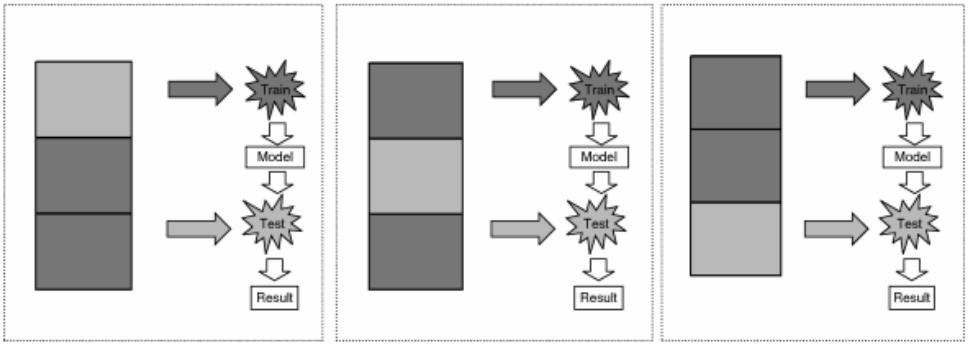
Figure 5 shows that as complexity increases (in this example as more 'wealth' related variables are added) the statistical trend-line also becomes more complex. A too simplistic model which might consist of only one or two variables will generate a model output which under-fits the relationship between wealth and happiness, and will therefore yield poor estimation performance. The other extreme, an overly complex model, will over-fit the data up to a degree where it becomes less accurate for estimation purposes. The optimum is a degree of model complexity that captures the statistical relationship between goal and target variables up to a degree where further increments of model complexity will compromise the estimation accuracy of the model.

3.3.4 Machine learning: validation

Variance between data entries can largely impact the results of machine learning based on which data entries are assigned the status of either training or test data. If estimations are made based on one specific training data-set, the machine learning model is simply trained to fit that specific training data-set in an optimal way. This brings the danger of unwanted variance in the outcomes due to differences between the training data-set and the data-set as a whole. In order to be able to make more generalized statements regarding the performance of variables within a machine learning model, one should assess all data and not just a specifically selected training data-set. In order to mitigate this unwanted potential form of variance, cross-validation can be used. Cross-validation is a statistical method that is used to evaluate machine learning algorithm performance by carrying out

successive iterations of model estimations based on randomly selected test and training data-sets. This way each data entry has an equal chance of being represented in either test or training data in one of the iterations. There is a variety of cross-validation methods, the most popular being: re-substitution, K-fold, repeated K-fold and leave-one-out. K-fold is considered most suitable for accurate performance estimations and will therefore be used for the analysis (Refaeilzadeh et al., 2009). In a K-fold cross-validation, the data-set as a whole is split into K-parts, after which an iteration will be carried out for each K-part, using that specific part as test data, and all other K-parts as training data. This is illustrated in figure 6. The mean of each K-iterations' performance estimate is used as final accuracy score. The result of this indicates up to which a machine learning algorithm is 'generalizable'. The higher the K-value (the more iterations), the more performance estimates scores. A higher K-value also leads to a larger training data-set (i.e. 50% for 2-fold and 95% for 20-fold), which trains the model to 'fit' to a larger part of the entire data-set when higher K-values are used, making it potentially even more generalizable.
Another potential that cross-validation offers is the opportunity to compare the outcomes of different kinds of machine learning algorithms. Due to the K-number of iterations, it decreases the influence of selected training/test-data on the outcomes. This leads to a situation where most of the differences in estimation accuracy can be attributed to the used machine learning algorithm, rather than the data (if the same data-set is used as input for both models). This allows one to assess which type of machine learning model is 'best' (most accurate) for a specific purpose or data-set (Refaeilzadeh et al., 2009).

Figure 6: Schematic overview of a three-fold cross-validation ((Refaeilzadeh et al., 2009))

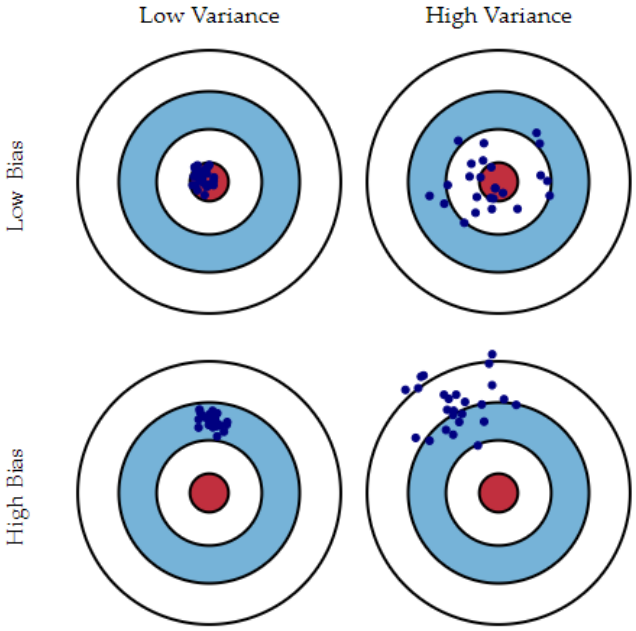


3.3.5 Error sources

Bias-Variance Trade-off
In order to be able to establish accurate models that are able to generate reliable estimations, it is important to assess different sources of potential errors and to discuss their (potential) influence on model outcomes. One of the main problems for machine learning algorithms is the so called 'bias-variance trade-off'. Errors due to bias are referred to as the difference between the expected outcome and the actual outcome. Bias measures the difference between these 2 values. Model bias can be attributed to all kinds of factors, model parameters, model selection or selected variables (a certain selection of variables might for example be an incorrect estimator for the value that one tries to estimate)
Variance refers to the variability in the outcome of the model. If one would repeat the entire modelling process multiple times, the difference between the outcomes of the individual model runs

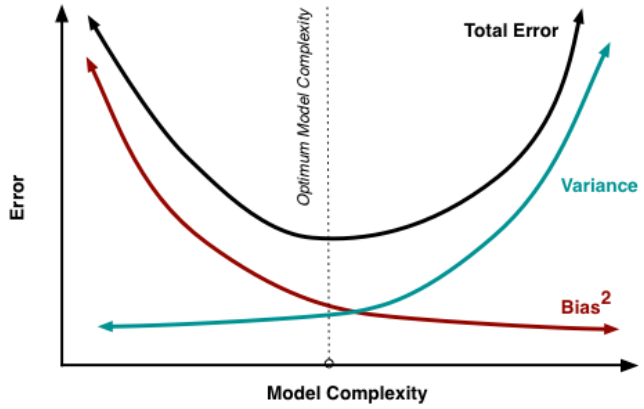
is referred to as variance. Typical sources of variance are sample sizes, model selection and model parameters. The differences between these two kinds of error sources are illustrated in figure 9

Figure 7: Illustration of the differences between bias and variance (Fortmann-Roe, 2018a)



In figure 9, the bulls-eye represents a perfectly accurate estimation. Each dot represents one run of a model. For some selections of training and test-data the results might be extremely accurate, whereas model estimations for selections where for example the training data-set contains a lot of outliers or extreme values might be less accurate, and therefore further off the bulls-eye. The problem of the bias-variance trade-off is that bias and variance are inversely related. Therefore, model enhancements that are implemented in order to decrease variance, often increase bias and vice-versa. The sweet spot for a model is the point at which at which the increase in bias equals the reduction in bias (or vice-versa) This inverse relationship is illustrated in figure 8.

Figure 8: Illustration of the inverse relationship between bias and variance (Fortmann-Roe, 2018a)



Assessing model quality

As the primary goal of the model that is developed is to accurately estimate new data based on the training data, one should also assess the error for estimating new data, rather than training data. Researchers often report on measurement errors for their training data-set, rather than their test data. While this might seem correct (and often results in lower estimation errors) it does not represent the *true* error for the estimation of 'new' data values. 'The true estimation error' is defined by (Fortmann-Roe, 2018b) as training error + training optimism, whereas training optimism refers to a measure that illustrates how much worse a model performs on new data compared to training data. In order to measure the true estimation error, the error should always be assessed for data values that are not in the training data-set.

For classification methods, the most used validation measures for assessing model quality are: accuracy, precision and recall. Accuracy is the most commonly used measurement that consists of the ratio of correctly classified data entries to the total amount of data entries. An accuracy score of 0,82 means that for 82% of the data entries, the model was able to estimate the right classification based on the training data. However, data entries can also be false-positively assigned to a class, or false-negatively to a class. This phenomenon, which is illustrated in table 3 leads to false-positives and false-negatives. In order to assess this, the other performance scores (precision, recall and F1) are used. Precision is the amount of correctly estimated positives, compared to the total amount of positive data entries. In other words: of all cycling trips that are labeled as utilitarian, how many were in fact utilitarian? Recall measures the opposite: 'of all cycling trips that are utilitarian, how many were correctly labeled as utilitarian?' Extreme values in either of the quality indicators could indicate out-performance or under-performance of a model.

Table 3: Assessment of model quality: false-positives and/or false-negatives

	Estimated Class		
Actual class		Utilitarian	Non-Utilitarian
	Utilitarian	True Positive	False Negative
	Non-Utilitarian	False positive	True Negative

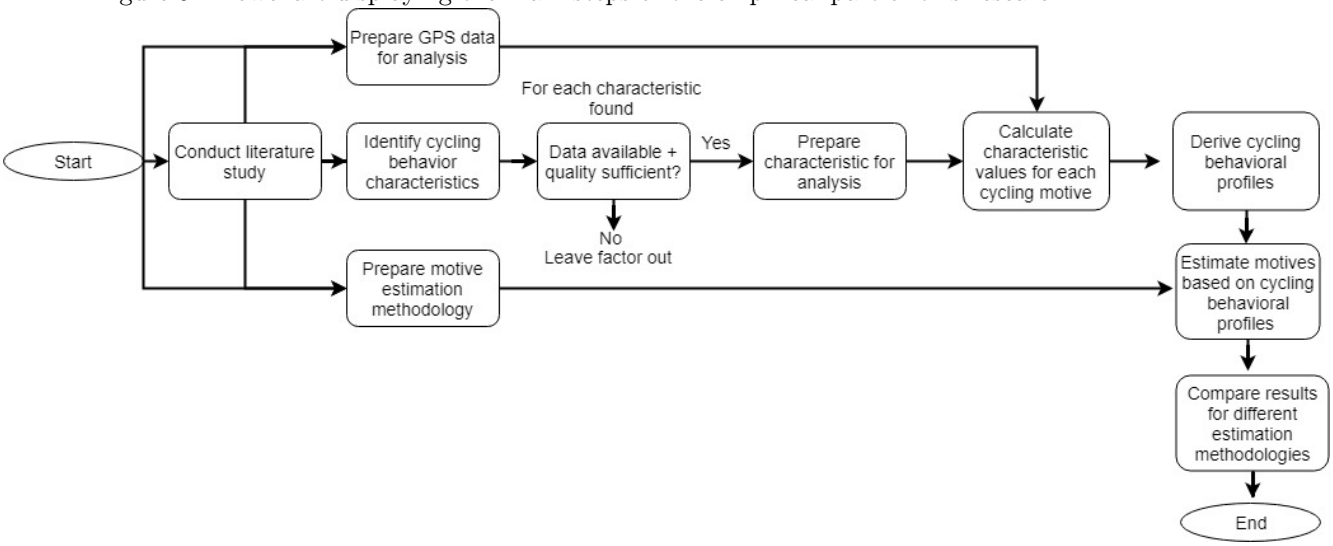
K-fold cross-validation is one of the methods that can be used to assess these model performance

indicators in an accurate simplistic way. For this purpose performance values for the test data are calculated for each K-iteration. Because only test-data is assessed, the averages of these estimations will provide a robust estimation of the *true* accuracy, precision, recall and error margin. Because of the earlier discussed bias-variance trade-off, it is important to select a value for K in which the bias-variance trade-off is optimal. Higher values will provide less biased results, but might also lead to increased variance. In practice: a K value of 10 is generally considered an effective fold size value (Fortmann-Roe, 2018b).

3.3.6 Flowchart of research

A flowchart is used to give an overview of the subsequent steps that are carried out to meet the objectives of this research. These steps are displayed in the flowchart below.

Figure 9: Flowchart displaying the main steps of the empirical part of this research



As stated in Figure 3.3.6 a literature study is conducted to identify cycling behavior characteristics and to identify existing methodologies for motive estimation and GPS data analysis. Subsequently, data of sufficient quality is gathered and prepared, after which behavioral cycling profiles are derived and their differences are assessed. These are then used to estimate cycling motives based on the estimation methodologies that are based on earlier similar studies. Finally, the results for the different estimation methodologies are compared.

3.4 Required additional material

Apart from the B-Riders data-set, the Fietersersbond cycling network and the external data-sets which contain additional data for the cycling behavioral characteristics, the ArcGIS platform, Python (more specifically: mainly the Pandas, Matplotlib, SciKit-learn and ArcPY packages) and literature studies are all resources required to carry out the (empirical part of the) research.

4 Analysis: behavioral cycling profiles

This chapter contains the results regarding the analysis of the behavioral cycling dimensions

4.1 Analysis of Variance: trip and route characteristics: raw motives

In order to determine the impact of the cycling motive on the trip and route characteristics that were extracted from theory, analysis of variance has been conducted for the 'raw' cycling motives that were provided for by the data-set.

4.1.1 Trip duration

Table 4 shows the average trip duration (in minutes) as well as the standard deviations. Trip duration is relatively low for study, daily shopping and non-daily shopping, and high for services, leisure and paid work.

Table 4: Trip duration for different purposes

	Average	Standard deviation
Leisure	24.1866279072	16.3545041211
Paid-work	30.6628184282	17.8990658221
Daily-Shopping	21.5115633074	18.7469787636
Services	25.5653846152	23.0326036662
Home	32.2714869808	20.4898845477
Non-Daily-Shopping	20.5372881355	17.3657937943
Recreational	25.2899807318	19.3782344124
Social	24.3461059192	19.0343650775
Study	18.6285714284	12.6870644491

A one-way ANOVA was conducted to compare the effect of different cycling motives on trip duration. The result of this test shows that there is a significant effect of cycling motive on trip duration on the $p<0,01$ level [$F=19.4$, $p<0,01$].

4.1.2 Average max speed

Table 5 shows the average max speed (in kilometres per hour) as well as the standard deviations of the average max speed on the segments of the cycling network that were used by the groups of cyclists for each different cycling motive. Average max speed is relatively low for paid work and relatively high for services.

Table 5: Average max speed for different purposes

	Average	Standard deviation
Leisure	17.3275634464	8.60997594443
Paid-work	14.879158729	8.36316614437
Daily-Shopping	16.8023981538	9.6729996027
Services	18.8651890862	10.5780524512
Home	15.8193639703	9.25636509967
Non-Daily-Shopping	16.5919682473	9.63169199113
Recreational	16.7142705026	9.58128433732
Social	17.1989185836	9.56946373889
Study	17.7890779934	9.16585712399

A one-way ANOVA was conducted to compare the effect of different cycling motives on average max speed of the cycling network. The result of this test shows that there is a significant effect of cycling motive on trip duration on the p<0,01 level [F=4.6, p<0,01].

4.1.3 Average trip speed

Table 6 shows the average trip speed (in kilometres per hour) as well as the standard deviations of the average trip speed for different cycling motives. There are no groups with extraordinary high or low average trip speeds.

Table 6: Average trip speed for different purposes

	Average	Standard deviation
Leisure	20.9438856794	1.06361111513
Paid-work	21.392858058	1.07245464268
Daily-Shopping	20.5634356322	1.52470300544
Services	21.1310525154	0.964558632872
Home	21.423678813	1.05117244977
Non-Daily-Shopping	20.874897065	1.29104592065
Recreational	21.0116763771	1.24469146977
Social	20.6949589096	1.40669518779
Study	20.7605556679	1.32726438366

A one-way ANOVA was conducted to compare the effect of different cycling motives on average trip speed. The result of this test shows that there is a significant effect of cycling motive on trip speed on the p<0,01 level [F=33.0, p<0,01].

4.1.4 Trip length

Table 7 shows the average trip length as well as the standard deviations of trip length for different cycling motives. Average trip length is relatively high for paid work and home, and relatively low for daily- and non-daily shopping, as well as study.

Table 7: Trip length for different purposes

	Average	Standard deviation
Leisure	11402.7386012	9080.15198921
Paid-work	15588.2157764	10395.3563945
Daily-Shopping	9294.80671127	9579.19247469
Services	12468.4248682	13342.2960742
Home	16103.7137955	11293.0541531
Non-Daily-Shopping	8973.72505413	9424.63635987
Recreational	11782.8096412	10889.7075174
Social	11140.6350877	10391.7705537
Study	9165.12427757	9366.08521188

A one-way ANOVA was conducted to compare the effect of different cycling motives on trip length. The result of this test shows that there is a significant effect of cycling motive on trip length on the $p<0,01$ level [$F=25.1$, $p<0,01$].

4.1.5 Type of road

Table 8 shows the average fraction of the type of road (cycling-specific or non-cycling-specific) for different cycling motives (where value '2' is cycling-specific only and '1' is non-cycling-specific only). The average fraction of cycling-specific roads is relatively high for study, and relatively low for recreational and home purposes.

Table 8: Type of road for different purposes

	Average	Standard deviation
Leisure	1.44621207443	0.270858276937
Paid-work	1.4146281961	0.285383054201
Daily-Shopping	1.45019669941	0.309520012149
Services	1.50866354162	0.343737225905
Home	1.3890645978	0.293587108232
Non-Daily-Shopping	1.46964281545	0.343675963473
Recreational	1.40463439973	0.320723301287
Social	1.43982615946	0.289429036262
Study	1.57853978614	0.261312120694

A one-way ANOVA was conducted to compare the effect of different cycling motives on the fraction of cycling-specific roads. The result of this test shows that there is a significant effect of cycling motive on the fraction of cycling-specific roads on the $p<0,01$ level [$F=3.6$, $p<0,01$].

4.1.6 Average Traffic Volume

Table 10 shows the average traffic volume for different cycling motives. There are two 'outlier' categories for this variable: the traffic volume is extraordinary high for 'services'-oriented trips, and exceptionally low for 'study'-purposes.

Table 9: Traffic volume for different purposes

	Average	Standard deviation
Leisure	1.43223985013	0.338542918977
Paid-work	1.52645705534	0.3162771317
Daily-Shopping	1.39799567035	0.364290411654
Services	1.460213466	0.340501679361
Home	1.58840931783	0.288517031557
Non-Daily-Shopping	1.38078772362	0.392843421999
Recreational	1.43494814091	0.386461958129
Social	1.41821175583	0.346926025382
Study	1.48733041757	0.358348377588

A one-way ANOVA was conducted to compare the effect of different cycling motives on the traffic volume. The result of this test shows that there is a significant effect of cycling motive on traffic volume on the $p<0,01$ level [$F=7.9$, $p<0,01$].

4.1.7 Appreciation of the environment

Amount of green
Table 10 shows the average amount of green for different cycling motives. There are two 'outlier' categories for this variable: the amount of green is extraordinary high for 'services'-oriented trips, and exceptionally low for 'study'-purposes.

Table 10: Trip length for different purposes

	Average	Standard deviation
Leisure	924.921845276	749.895035591
Paid-work	1090.46455828	826.388500294
Daily-Shopping	921.941360973	820.596299547
Services	1553.92838723	1357.48627852
Home	974.274488032	770.795410834
Non-Daily-Shopping	911.427892685	829.155553313
Recreational	946.309560912	744.337525844
Social	946.166894522	818.46974844
Study	615.965636857	378.073703999

A one-way ANOVA was conducted to compare the effect of different cycling motives on the amount of green. The result of this test shows that there is a significant effect of cycling motive on the amount of green on the $p<0,01$ level [$F=23.0$, $p<0,01$].

Appreciation of the environment
Table 11 shows the average appreciation of the environment for different cycling motives.

Table 11: Appreciation of the environment for different purposes

	Average	Standard deviation
Leisure	2.85046698427	0.337665880383
Paid-work	2.85403489387	0.359904077785
Daily-Shopping	2.86086277056	0.372159359537
Services	2.94889230115	0.361730105724
Home	2.88062519871	0.372900760353
Non-Daily-Shopping	2.82322167502	0.438449835004
Recreational	2.81137077295	0.509373278009
Social	2.82351163912	0.386307795693
Study	3.06869344971	0.365344288941

A one-way ANOVA was conducted to compare the effect of different cycling motives on environmental appreciation. The result of this test shows that there is no significant effect of cycling motive on appreciation on the $p>0.01$ level [$F=2.2$, $p<0,05$].

4.1.8 ANOVA Analysis for trip and route characteristics: conclusion

The ANOVA analysis shows that there are significant differences between groups for all variables that were assessed except 'appreciation of the environment' which was significant on the $p<0,05$ level but not on the $p<0,01$ level. Machine learning will be used to determine whether these differences actually make it possible to distinguish the different cycling motives using the trip and route characteristics that have been analyzed and were found to have significant differences.

4.2 Independent samples t-test for trip and route characteristics: adjusted motives

Theory states that most of differences in network usage between cyclists with different cycling motives do not originate from the specific motive (i.e. study), but rather on the fact whether the purpose of the trips are either utilitarian or non-utilitarian. In order to investigate this, cycling motives were grouped as either utilitarian or non-utilitarian based on the descriptions that have been discussed in Section 3.1.4. Subsequently, an independent samples t-test has been performed for each trip and/or route characteristic to determine whether there is a significant difference for these factors between these two groups of cyclists. The results of this are provided below:

4.2.1 Trip duration

Table 12 shows the average trip duration (in minutes) as well as the standard deviations. Trip duration is lower for trips with a utilitarian purpose.

Table 12: Trip duration for different purposes

	Average	Standard deviation
Utilitarian	28.9766162514	18.3406627644
Non-Utilitarian	30.1802738654	20.3207208223

An independent samples T-Test was conducted to compare the effect of the two different cycling motives for trip duration. It shows that there is a significant difference in the trip duration for utilitarian (M=28,98, SD=18,34) and non-utilitarian (M=30,18, SD=20,32) trips; t=-2,05, p=0,04. The effect size is small (d=-0,06).

4.2.2 Average max speed

Table 13 shows the average max speed (in km/h) as well as the standard deviations. Max speed is lower for trips with a utilitarian purpose.

Table 13: Average max speed for different purposes

	Average	Standard deviation
Utilitarian	15.2309668069	8.64764050971
Non-Utilitarian	16.1622292465	9.32143545596

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on max speed. It shows that there is a significant difference in the max speed for utilitarian (M=15.23, SD=8.64) and non-utilitarian (M=16.16, SD=9.32) trips; t=-3.42, p=<0,01. The effect size is small (d=-0,10).

4.2.3 Average trip speed

Table 14 shows the average trip speed (in km/h) as well as the standard deviations. Trip speed is almost equal for the two different kinds of cycling motives.

Table 14: Average trip speed for different purposes

	Average	Standard deviation
Utilitarian	21.2650261681	1.18076236297
Non-Utilitarian	21.2610385872	1.16186968144

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on average trip speed. It shows that there is no significant difference in the trip speed for utilitarian (M=21.27, SD=1.18) and non-utilitarian (M=21.26, SD=1.16) trips; t=0,11, p=0,9. The effect size is small (d=0,00).

4.2.4 Trip length

Table 15 shows the average trip length as well as the standard deviations. Trip length is lower for utilitarian trips.

Table 15: Trip length for different purposes

	Average	Standard deviation
Utilitarian	14460.6846346	10551.9888353
Non-Utilitarian	33729.7620375	872847.805328

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on trip length. Even though the averages suggest a large difference: the T-test shows that there is no significant difference in trip length for utilitarian (M=14460.68, SD=10551.99) and non-utilitarian (M=33729.76, SD=872847.80) trips; t=-1,02, p=0,3. The effect size is small (d=-0,03).

4.2.5 Type of road

Table 16 shows the average fraction of the type of road (cycling-specific or non-cycling-specific) for different cycling motives (where value '2' is cycling-specific only and '1' is non-cycling-specific only), as well as the standard deviations. Type of road is slightly more 'non-cycling-specific' for trips with a utilitarian purpose..

Table 16: Type of road

	Average	Standard deviation
Utilitarian	1.42316272488	0.292568510004
Non-Utilitarian	1.40026163941	0.295026172875

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on type of road. It shows that there is a significant difference in the type of road for utilitarian (M=1.42, SD=0.29) and non-utilitarian (M=1.40, SD=0.29) trips; t=2,58, p=0,01. The effect size is small (d=0,08).

4.2.6 Average traffic volume

Table 17 shows the average traffic volume, as well as the standard deviations. Traffic volume is higher for trips with a utilitarian purpose.

Table 17: Type of road

	Average	Standard deviation
Utilitarian	1061.45064966	831.957063386
Non-Utilitarian	965.486705702	775.310334505

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on average traffic volume. It shows that there is a significant difference in traffic volume for utilitarian (M=1061.45, SD=831.96) and non-utilitarian (M=965.49, SD=775.31) trips; t=3.95, p=<0,01. The effect size is small (d=0,12).

4.2.7 Appreciation of the environment

Because of the 'immeasurability' of emotions such as 'appreciation', the appreciation of the environment was measured using two variables: amount of green (objective), and hand-reported appreciation of the environment (subjective).

Amount of green Table 18 shows the average amount of green, as well as the standard deviations. Amount of green is higher for trips with a utilitarian purpose.

Table 18: Average environment value

	Average	Standard deviation
Utilitarian	1.50344061492	0.33077656486
Non-Utilitarian	1.5439870477	0.317383086798

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on average environment value. It shows that there is a significant difference in environment value for utilitarian (M=1.50, SD=0.33) and non-utilitarian (M=1.54, SD=0.32) trips; t=-4,14, p=<0,01. The effect size is small (d=-0,12).

Appreciation of the environment Table 19 shows the average appreciation of the environment, as well as the standard deviations. Appreciation of the environment is almost equal for the two kinds of cycling motives.

Table 19: Appreciation of the environment

	Average	Standard deviation
Utilitarian	2.8550864946	0.366438887401
Non-Utilitarian	2.86512322048	0.387271677873

An independent samples T-Test was conducted to compare the effect of the two different cycling motives on the appreciation of the environment. It shows that there is no significant difference in appreciation of the environment for utilitarian (M=2.85, SD=0.37) and non-utilitarian (M=2.87, SD=0.39) trips; t=-0.88, p=0,37. The effect size is small (d=0,02).

4.2.8 Independent samples T-test for trip and route characteristics: conclusion

The T-test shows that there are significant differences between the two groups for all variables except appreciation of the environment, trip length and average trip speed. However, the impact of each factor is generally fairly small. Machine learning will be used to determine whether these differences actually make it possible to distinguish the different cycling motives using the trip and route characteristics that have been analyzed and were found to have significant differences.

4.3 Chi-Square: origin-destination factors (raw motives)

In order to determine whether the origin and destination are significantly different for different cycling motives, a chi-square has been conducted. The extracted BBG values for 'origin' and 'destination' were used as input.
Scipy was used in order to generate cross-tabulations (Table 20 and Table 21) which contain the frequency data of the origins and destinations for different cycling purposes. This data was used to calculate a chi-square to determine whether origins are equally distributed among the different cycling motive. This turns out not to be the case for both origin (X2 (152, N = 4345) = 642,72, p <.01) and destination (X2 (128, N = 4345) = 1167,60, p <.01).

Table 20: Cross-tabulation of cycling motives and origins

Origins vs Cycling motives	paid work	daily-groceries	services	home	non-daily-groceries	recreation	social	study	vrijetijd
Business Districts	80	40	1	229	16	18	47	1	14
Forest	62	7	1	36	3	7	11	0	4
Building site	12	5	0	46	3	0	3	1	4
(Semi) Agricultural	327	9	1	152	5	11	15	1	4
Parks and gardens	68	10	0	59	5	5	7	1	3
Track field	14	4	1	36	1	8	11	0	4
Sports ground	32	4	1	38	4	2	7	0	2
Retail and catering	48	12	1	100	8	9	18	1	5
Public services	5	1	0	9	0	1	1	0	1
Socio-cultural	37	27	1	194	13	13	34	1	8
Road terrain	74	16	0	93	4	9	11	0	3
Housing	1062	116	6	521	51	89	152	8	32

Table 21: Cross-tabulation of cycling motives and destinations

Destinations vs Cycling motives	paid work	daily-groce-ries	serv-ices	home	non-daily-groce-ries	recre-ation	social	study	vrije-tijd
Business Districts	378	12	0	66	9	11	20	2	9
Forest	8	5	1	5	1	0	4	0	1
Building site	69	7	0	6	1	5	3	0	0
(Semi) Agricultural	64	8	1	158	5	6	12	0	3
Parks and gardens	75	8	0	39	6	12	11	1	2
Track field	21	7	1	19	1	3	7	0	1
Sports ground	66	1	0	21	0	5	8	0	3
Retail and catering	124	26	1	26	5	11	18	5	4
Public facilities	14	2	0	0	0	1	1	0	1
Socio-cultural	412	14	2	58	9	28	28	1	7
Road terrain	78	18	0	71	6	9	11	1	2
Housing	504	139	7	1052	70	82	196	4	51

4.4 Chi-Square: origin-destination factors (Adjusted motives)

The significance of the difference between groups was also tested for the adjusted motives. Table 22 contains the cross-tabulation of the origins for adjusted motives. Table 23 displays the cross-tabulation of the destinations for adjusted motives.

Table 22: Cross-tabulation of cycling motives and origins for adjusted motives

Origins vs Cycling motives	Utilitarian	Non-Utilitarian
Business Districts	138	308
Forest	73	58
Building site	21	53
(Semi) Agricultural	343	182
Parks and gardens	84	74
Track field	41	59
Sports ground	41	49
Retail and catering	70	132
Public facilities	6	12
Socio-cultural	79	249
Road terrain	94	116
Housing	1243	794

Table 23: Cross-tabulation of cycling motives and destinations for adjusted motives

Destinations vs Cycling motives	Utilitarian	Non-Utilitarian
Business Districts	401	106
Forest	15	10
Building site	77	14
(Semi) Agricultural	78	179
Parks and gardens	90	64
Track field	30	30
Sports ground	67	37
Retail and catering	161	59
Public facilities	16	3
Socio-cultural	438	121
Road terrain	103	93
Housing	794	1381

The results of the chi-square test indicate that there are significant differences between utilitarian and non-utilitarian cycling motives for both origins (X2 (19, N = 4319) = 367,30, p <.01.) and destinations (X2 (19, N = 4319) = 720,14, p <.01.).

4.4.1 Chi Square: Conclusion

Given the outcomes of the Chi-Square tests, it can be concluded that origins and destinations are not equally distributed among the different cycling motives for both raw and adjusted motives. The next sub-chapter will serve to test whether the differences in trip and route characteristics, as well as the differences in origin and destination can be used to help identify cycling motives for GPS cycling tracks, by means of machine learning.

4.5 Cycling behavioral profiles

This sub-chapter combines the different behavioral characteristics that were discussed in the previous sub-sections in order to compose cycling behavior profiles for each of the raw and adjusted motives based on these characteristics.

4.5.1 Raw motives

Table 24: Cycling behavioral profiles for raw motives: trip and route

Variable	**Full sample score**	**Paid-work**	**Daily-shop-ping-**	**Serv-ices**	**Home**	**Non-dai-ly-shop-ping**	**Recr-eation**	**Social**	**Study**	**Leis-ure**
Average trip speed	21.26	21.39	20.57	21.13	21.42	20.88	21.01	20.70	20.76	20.98
Trip length	23857	15518	9354	12468	42295	9033	11782	11168	9165	11521
Average max speed	15.72	14.91	16.78	18.86	15.87	16.72	16.71	17.15	17.79	17.41
Average traffic volume	1019	1095	923	1554	979	911	946	949	616	932
Type of road	1.41	1.42	1.45	1.51	1.39	1.48	1.40	1.44	1.58	1.45
Environ-ment value	1.52	1.52	1.40	1.46	1.58	1.39	1.43	1.42	1.49	1.44
Appre-ciation of envir-onment	2.86	2.85	2.86	2.95	2.88	2.84	2.81	2.82	3.07	2.85

The most important differences between the different cycling behavioral profiles for the trip and route dimensions are underlined in Table 24. Green cell colors represent values that are above the average sample score for the entire data-set, whereas red colors represent values which are lower than the data-set average. Grey cells represent values that are equal to the data-set average. What strikes most are the differences in trip length (trips towards home being a lot longer than other motives), differences in average max speed (paid work being the only motive scoring lower than average), traffic volume (paid-work scoring slightly higher than average and services scoring a lot higher, study scoring low).

Table 25: Origins behavioral profile: raw motives

Variable	Full sample score	Paid-work	Daily-Shop-ping	Serv-ices	Home	Non dai-ly-shop-ping	Recr-eation	Social	Study	Leis-ure
Business Districts	10.42%	4.39%	16.19%	7.69%	15.14%	14.29%	10.98%	14.87%	7.14%	17.28%
Forest -	3.06%	3.4%	2.83%	7.69%	2.38%	2.68%	4.27%	3.48%	0%	4.94%
Building site	1.73%	0.66%	2.02%	0%	3.04%	2.68%	0%	0.95%	7.14%	4.94%
(Semi) Agricul-tural	12.26%	17.96%	3.64%	7.69%	10.05%	4.46%	6.71%	4.75%	7.14%	4.94%
Parks and gardens	3.69%	3.73%	4.05%	0%	3.9%	4.46%	3.05%	2.22%	7.14%	3.7%
Track field	1.64%	0.77%	1.62%	7.69%	2.38%	0%	0%	3.48%	0%	4.94%
Sports ground	2.01%	1.76%	0%	7.69%	2.51%	3.57%	1.22%	2.22%	0%	2.47%
Retail & Catering	4.72%	2.64%	4.86%	7.69%	6.61%	7.14%	5.49%	5.7%	7.14%	6.17%
Public Facilities	0.4%	0.27%	0.4%	0%	0.59%	0%	0.61%	0%	0%	1.23%
Socio-cultural	7.59%	2.03%	10.93%	7.69%	12.82%	11.61%	7.93%	10.76%	7.14%	6.17%
Road Terrain	4.91%	4.06%	6.48%	0%	6.15%	3.57%	5.49%	3.48%	0%	3.7%
Housing -	47.58%	58.32%	46.96%	46.15%	34.43%	45.54%	54.27%	48.1%	57.14%	39.51%

Table 26: Destinations behavioral profile: raw motives

Variable	Full sample score	Paid-work	Daily-Shop-ping	Serv-ices	Home	Non dai-ly-shop-ping	Recr-eation	Social	Study	Leis-ure
Business Districts	11.82%	20.95%	4.86%	0%	4.34%	7.96%	6.36%	6.27%	14.29%	10.71%
Forest -	0.58%	0.44%	2.02%	7.69%	0.33%	0.88%	0%	1.25%	0%	1.19%
Building site	1.91%	3.33%	2.83%	0%	0.39%	0.88%	2.89%	0.94%	0%	0%
(Semi) Agricul-tural	5.99%	3.55%	3.24%	7.69%	10.39%	4.42%	3.47%	3.76%	0%	3.57%
Parks and gardens	3.59%	4.16%	3.24%	0%	2.56%	5.31%	6.94%	3.45%	7.14%	2.38%
Track field	1.4%	1.16%	2.83%	7.69%	1.25%	0.88%	1.73%	2.19%	0%	1.19%
Sports ground	2.43%	3.66%	0.4%	0%	1.38%	0%	2.89%	2.51%	0%	3.57%
Retail & Catering	5.13%	6.87%	10.53%	7.69%	1.71%	4.42%	6.36%	5.64%	35.71%	4.76%
Public Facilities	0.44%	0.78%	0.81%	0%	0%	0%	0.58%	0.31%	0%	1.19%
Socio-cultural	13.04%	22.84%	5.67%	15.38%	3.81%	7.96%	16.18%	8.78%	7.14%	8.33%
Road Terrain	4.57%	4.32%	7.29%	0%	4.67%	5.31%	5.2%	3.45%	7.14%	2.38%
Housing -	49.09%	27.94%	56.28%	53.85%	69.17%	61.95%	47.4%	61.44%	28.57%	60.71%

The origin-destination behavioral profiles show some large distinguishing features between the different cycling motives. For example paid work, recreation, social and study are more likely to originate from housing location. For people who have 'home' as a cycling motive, housing obviously is the least likely point of origin. This is in line with what should be expected. The opposite is true for destination. The paid work and study motives are least likely to have housing as their destination. Paid work has a large fraction of business districts and socio-cultural destinations, whereas study has retail catering and business districts as a destination relatively often. Other results worth mentioning are the fact that most of the categories that can be associated with recreation (parks gardens, track field, sports ground, retail catering, public facilities, socio-cultural) are over-represented for the recreation motive. Business districts on the other hand is underrepresented (almost half the mean fraction for the entire data-set). Finally, retail catering scores high for daily-shopping, which was also to be expected. The difference however between daily and non-daily shopping is large for this category which is surprising (10.53% versus 4.42%). The next sub-chapter will assess the behavioral cycling profiles for adjusted motives (utilitarian versus non-utilitarian).

4.5.2 Adjusted motives

Table 27: Cycling behavioral profiles for adjusted motives: trip and route

Variable	Full sample score	Utilitarian score	Non-utilitarian score
Average trip speed	21.26	21.26	21.26
Trip length	23857.14	14414.09	33841.21
Average max speed	15.72	15.26	16.20
Average traffic volume	1019.05	1065.29	970.15
Type of road	1.41	1.42	1.40
Environment value	1.52	1.50	1.55
Appreciation of environment	2.86	2.86	2.86

The behavioral profiles for the trip and route dimensions show some important differences. Mean trip length is much higher for non-utilitarian trips, whereas average traffic volume is higher for utilitarian trips. This is to be expected since trips carried out for utilitarian purposes -like paid work- usually have a high need for efficiency (to get from point A to point B as quickly as possible), whereas this is not necessarily true for non-utilitarian cycling motives like leisure. Average traffic volume is higher for utilitarian trips. This can also be attributed to the need for efficiency. Cyclists are known to prefer low traffic volume segments (Winters et al., 2010). People that have non-utilitarian trips might avoid route segments with a lot of other traffic because they have the time 'available' to do so, whereas people with utilitarian motives do not because efficiency is more important to them than their preference for low traffic volume segments. Utilitarian cyclists' trips are also more often carried out on cycling-specific roads (even though the differences are small). This might also be due to the fact that cycling-specific roads allow one to travel at a higher pace in a safe way compared to non-cycling specific roads. It might also be related to the fact that utilitarian trips are -on average- carried out on route segments with higher traffic volume, which are typical places for cycling lanes to be built. Finally, the environmental score is higher (on average) for non-utilitarian trips.

Table 28: Cycling behavioral profiles for adjusted motives: origin

Origins vs Cycling motives	Full Sample	Utilitarian	Non-Utilitarian
Business Districts	10,33%	6.19%	13.8%
Forest	3,04%	3.27%	2.6%
Building site	1,72%	0.95%	2.38%
(Semi) Agricultural	12,16%	15.37%	8.16%
Parks and gardens	3,66%	3.77%	3.32%
Track field	2,32%	1.84%	2.65%
Sports ground	2,09%	1.84%	2.2%
Retail and catering	4,68%	3.14%	5.92%
Public facilities	0,42%	0.27%	0.54%
Socio-cultural	7,6%	3.54%	11.16%
Road terrain	4.87%	4.21%	5.2%
Housing	47,17%	55.67%	35.56%

Table 29: Cycling behavioral profiles for adjusted motives: destination

Destinations vs Cycling motives	**Full Sample**	**Utilitarian**	**Non-Utilitarian**
Business Districts	11.61%	17.67%	4.67%
Forest	0.58%	0.67%	0.45%
Building site	2.09%	3.4%	0.62%
(Semi) Agricultural	5.89%	3.44%	7.89%
Parks and gardens	3.53%	3.97%	2.82%
Track field	1.38%	1.33%	1.33%
Sports ground	2.39%	2.96%	1.63%
Retail and catering	5.04%	7.1%	2.6%
Public facilities	0.44%	0.71%	0.14%
Socio-cultural	12.81%	19.3%	5.34%
Road terrain	4.49%	4.54%	4.1%
Housing	49.81%	34.98%	60.84%

The behavioral profiles for the origin-destination characteristics also have some unique differences. Most of the utilitarian trips originate from home, which is logical. Non-utilitarian trips are relatively more often originated from business districts and socio-cultural area's. The other categories are (at least somewhat) equally distributed.
As to be expected 'business districts' make up a way larger percentage of the total amount of *destinations* for utilitarian (17,67%) compared to non-utilitarian (4,67%) trips. For non-utilitarian there are more trips destined to area's classified as agricultural and home, whereas utilitarian trips have relatively more trips that end in social-cultural, retail and public-service area's.

The next chapter will assess up to what extent the differences in cycling behavioral profiles can be used to help estimate cycling motives. Also, the effectiveness of the different machine learning algorithms that were discussed in Section 3.3 will be assessed.

5 Analysis: Estimation of cycling motives

This chapter contains the results regarding the estimation of cycling motives and the effectiveness of the different machine learning algorithms.

As was shown in the previous sub-chapters, there are significant differences in trip and route characteristics and origin-destination factor values for the different cycling motives. Whether these differences are large enough in order to be able to distinguish different cycling motives solely based on these data-values is assessed by means of machine learning. As stated in section 3.1.4 motives are commonly grouped into adjusted motives for travel behavior analysis purposes. Therefore, for estimation purposes, motives were grouped into utilitarian and non-utilitarian (adjusted motives), which is also more in line with the existing theory regarding differences in network use among cyclists with different cycling motives (Plazier et al., 2017).
Four simulations are used to test the estimation capacities of the behavioral cycling dimensions that are assessed in this research: 1) Trip factors 2) Origin-destination factors 3) Route factors 4) A combination of one two and tree. Their scores are compared to a standard model, which simply assumes that all data entries belong to the largest group (utilitarian, which leads to an accuracy of 51.3%). In addition to that, a leave-one-out analysis is conducted to test the influence of each sub-variable on the outcome of the model. Subsequently, the influence of sample size and chosen variables is discussed. Finally, model learning curves are assessed for each machine learning algorithm in order to derive results regarding validity and generalizability of each model.

5.1 Estimating based on route characteristics

Table 30 shows the cross-validated accuracy, as well as the performance versus the standard model for all different machine learning algorithms that were used, using the route characteristics (average max speed, type of road, traffic volume, environment value, appreciation of environment) that were prepared as model input. Compared to the standard model, all methods except the Neural Network (-4,9%) method outperform the standard model. However, most of these methods (Logistic Regression, Support Vector Machine, Naive Bayes) only slightly outperform the standard model. The three methods that stand out are Gradient Boosting (+8,0%), Decision Tree Classifying (+12,3%) and Random Forest Classifying (+13,6%). Of these method Random Forest scored highest on both accuracy as well as precision. All-together these results indicate that route characteristics *could* be used in order to help estimate cycling motives more accurately. However, in order for it to be effective, the right machine learning algorithm (i.e. Random Forest) should be used, and even then the increase in estimation accuracy is only fairly small.

Table 30: Cross-validated accuracy and performance vs standard model for different ML methods

Machine learning algorithm	**10-fold cross validated accuracy**	**Performance vs standard model**	**Precision**	**Recall**
Standard model	51.4%	0.0%	51.4%	100%
Logistic Regression	51.6%	+0.4%	57.3%	64.3%
Decision Tree Classifier	57.7%	+12.3%	60.6%	66.1%
Random Forest Classifier	58.4%	+13.6%	73.9%	62.4%
Support Vector Machine	53.6%	+4.3%	65.7%	64.2%
K-nearest neighbor	54.5%	+6.0%	59.6%	59.6%
Naïve Bayes	51.9%	+1.0%	56.2%	56.8%
Neural Network	48.9%	-4.9%	53.0%	20.3%
Gradient Boosting (GBM)	55.5%	+8.0%	64.1%	63.8%

5.2 Estimating based on origin-destination factors

Table 31 shows the cross-validated accuracy, as well as the performance versus the standard model for all different machine learning algorithms that were used, using origin and destination factors as model input. Compared to the standard model, all methods strongly outperform the standard model. The K-nearest neighbor and naive bayes methodologies outperform the standard model, but are less accurate then the other machine learning algorithms. All models have a high precision. Naive Bayes is the only model with a low recall score, indicating that it is relatively weak at identifying utilitarian trips. All-together these results indicate that origin and destination factors are good estimators of cycling motives and, since they are easy to derive from raw GPS tracks, suitable for enrichment of existing GPS tracks.

Table 31: Cross-validated accuracy and performance vs standard model for different ML methods

Machine learning algorithm	**10-fold cross validated accuracy**	**Performance vs standard model**	**Precision**	**Recall**
Standard model	51.4%	0.0%	51,4%	100%
Logistic Regression	69.3%	+34.8%	73.3%	60.7%
Decision Tree Classifier	68.3%	+32.9%	71.4%	58.3%
Random Forest Classifier	68.3%	+32.9%	72.0%	59.4%
Support Vector Machine	68.8%	+33.9%	71.0%	64.1%
K-nearest neighbor	66.4%	+29.2%	67.1%	58.5%
Naïve Bayes	59.6%	+16.0%	70.8%	39.6%
Neural Network	69.3%	+34.8%	72.5%	60.9%
Gradient Boosting (GBM)	69.3%	+34.8%	73.0%	60.1%

5.3 Estimating based on trip characteristics

Table 32 shows the cross-validated accuracy, as well as the performance versus the standard model for all different machine learning algorithms that were used, using trip characteristics (length and travel time) as input. Motive estimations based on trip characteristics tend to outperform the standard model (except for the neural network method, which slightly under-performs (-1,2%) and also has an extremely low recall (14,%). The Random Forest Classifying (35,2% out-performance) and the Decision Tree Classifying method (27,0% out-performance) seem particularly accurate for

this purpose. The accuracy, precision and recall scores for Logistic Regression and Naive Bayes indicate that they act (almost) identical to a standard model, making them unsuitable for estimation purposes.

Table 32: Cross-validated accuracy and performance vs standard model for different ML methods

Machine learning algorithm	**10-fold cross validated accuracy**	**Performance vs standard model**	**Precision**	**Recall**
Standard model	51.4%	0.0%	51,4%	100%
Logistic Regression	53.2%	+3.5%	51,7%	98,1%
Decision Tree Classifier	65.3%	+27.0%	56,0%	59,0%
Random Forest Classifier	69.5%	+35.2%	61,0%	55,5%
Support Vector Machine	59.1%	+15.0%	59,4%	77,3%
K-nearest neighbor	58.4%	+13.6%	56,2%	56,2%
Naïve Bayes	53.8%	+4.7%	51,8%	96,8%
Neural Network	50.8%	-1.2%	51,7%	14,5%
Gradient Boosting (GBM)	60.7%	+18.1%	58,1%	60,1%

5.4 Estimating based on a combination of network, trip and origin-destination factors

The combined model is the most comprehensive model that is tested. It consists of all available prepared data (origin, destination, trip characteristics and network characteristics). The cross-validated accuracy, as well as the performance versus the standard model are shown in Table 33 for the different machine learning algorithms. What strikes is that the most accurate estimation (Random Forest Classifier: 74%) is the most accurate estimation of all models, and provides a high out-performance (44%) compared to the standard model. It also yields a high precision (80,6%) and an acceptable recall (69.4%)Other machine learning algorithms that perform well for the entire data-set are the Decision Tree Classifier (+36,6%) and Gradient Boosting (+37,7%) methods.

Table 33: Cross-validated accuracy and performance vs standard model for different ML methods

Machine learning algorithm	**10-fold cross validated accuracy**	**Performance vs standard model**	**Precision**	**Recall**
Standard model	51.4%	0.0%	51.4%	100%
Logistic Regression	53.8%	+4.7%	56.9%	66.5%
Decision Tree Classifier	70.2%	+36.6%	65.3%	70.7%
Random Forest Classifier	**74.0%**	**+44.0%**	**80.6%**	**69.4%**
Support Vector Machine	54.6%	+6.2%	63.3%	95.0%
K-nearest neighbor	57.5%	+11.9%	59.6%	61.1%
Naïve Bayes	53.6%	+4.3%	56.1%	93.6%
Neural Network	51.2%	-0.4%	53.8%	68.4%
Gradient Boosting (GBM)	70.8%	+37.7%	76.2%	70.2%

5.5 Assessing the effectiveness of individual factors: leaving one out

The previous sub-chapters assessed the effectiveness of the three main 'groups' of variables (origin-destination, trip characteristics and route characteristics). This sub-chapter assesses the impact of

individual factors. This is done by leaving one factor out at a time. Subsequently, the score of this 'incomplete' model is related to the score of the 'complete' model (including all available variables) in order to assess the impact of the factor that is being left out. The Random Forest classifying method is used for this purpose since it was found to be most all-round well-scoring classification method for the models that were assessed (in terms of accuracy). The accuracy of the Random Forest classifying method for the whole data-set was 74,0%. This value therefore serves as 'standard model' for this analysis. Because of the marginal differences that might occur, 10-fold cross-validated accuracy is used to ensure that even small differences can be detected accurately.
Table 34 shows the results of the analysis. The results indicate that the destination type of a trip has the largest impact on the accuracy of the model. Trip speed, max speed and road type also have a fair impact on the accuracy. The variables related to the environment, and the trip length of trips have only little impact on the models' accuracy. Leaving the origin of trips out has no measurable impact on the accuracy of the model. Average traffic volume is the only variable that actually has a negative impact on the accuracy of the model. In other words: the model as a whole is more accurate (0,5%) without the inclusion of this factor. In conclusion: all variables that were assessed can, with the exception of traffic volume, be used to help estimate cycling motives.

Table 34: Cross-validated accuracy and performance vs complete model for different factors left out

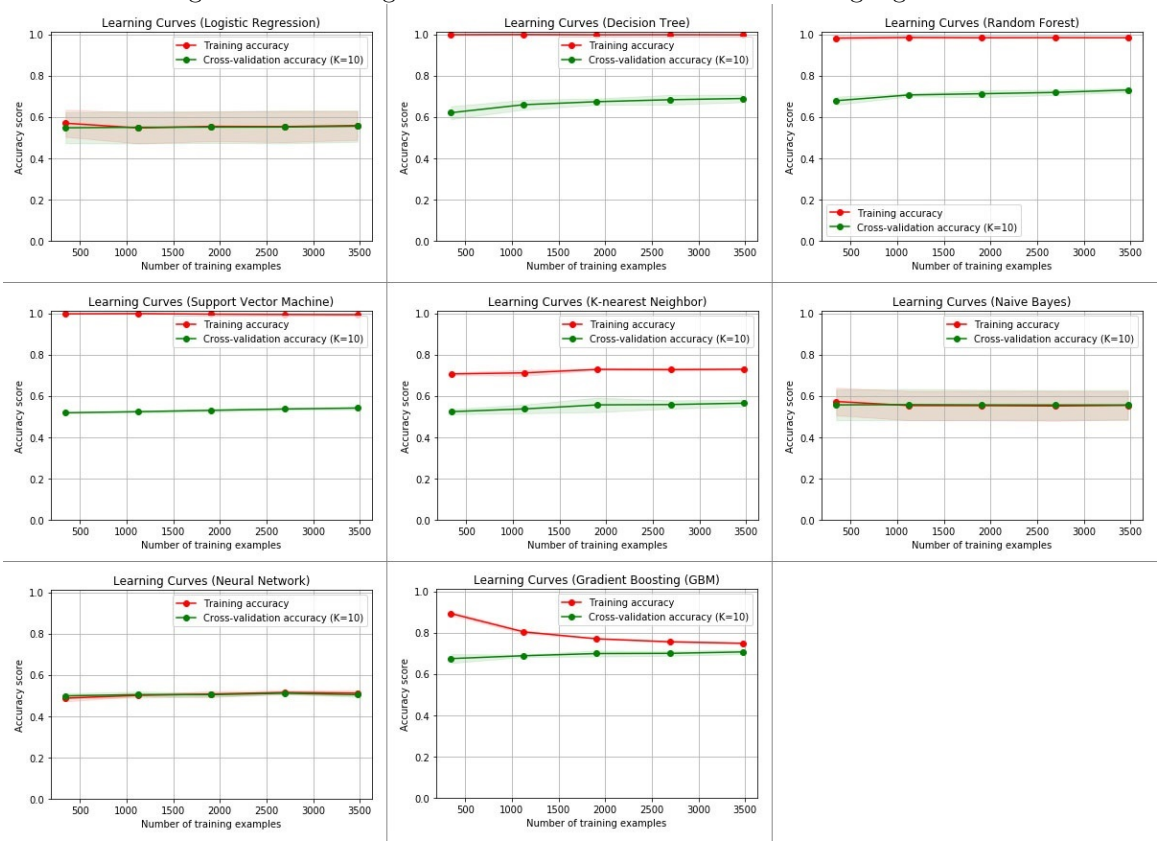
Factor left out	**10-fold cross validated accuracy**	**Performance vs standard model**
No factor left out	74,0%	0.0%
Origin	74,0%	0.0%
Destination	71,8%	-3,1%
Average trip speed	73,3%	-0,95%
Trip length	73,9%	-0,1%
Average max speed	73,1%	-1,2%
Average traffic volume	74,4%	+0,5%
Type of road	73,4%	-0,8%
Environment value	73,8%	-0,3%
Appreciation of environment	73,8%	-0,3%

5.6 The role of machine learning algorithms, the influence of chosen variables and sample size

One of the goals of this research is to assess the effectiveness of different machine learning algorithms to carry out the cycling motive estimation task. The influence of bias, variance and the relationship between models and their potentials for over-fitting have been discussed extensively. As stated multiple times in this research, the chosen variables have an extreme influence on the outcome of machine learning analysis. A model is only as strong as the input data that it is provided with. The variables that were chosen to represent route, trip and origin-destination were carefully selected based on theoretical justification and availability of data. However, that does not make the data framework on which the analysis is based perfect. In fact, the variables that were used only make up a small part of a much larger range of factors that *could* serve as estimators for estimating cycling purposes of which most have been discussed in Section 2.2.2. However, for most other factors that were theoretically justified, data was unavailable, of insufficient quality, or in need of such an amount of pre-processing that it would impose too large time constraints. Other factors for which data was available but were not in line with the scope of this research (e.g. socio-economic variables) are excluded. Factors within the scope of this research for which data was available, but could not be adequately theoretically justified were also excluded. One of the reasons for this is that, as discussed extensively in Section 3.3, adding more variables does not necessarily lead to more accurate and more valid results. In fact, adding more 'noise' facilitates over-fitting which yields inaccurate estimation results for new data-sets.

As stated before, the lowest estimation error is achieved at a level of model complexity where an increase in complexity leads to an increase of the estimation error for new data. One way to assess this is by generating and analyzing the learning curves for different machine learning algorithms. A learning curve shows the performance of a specific model for training and test data for a varying amount of data-samples. The resulting graph gives an indication of two things: the degree up to which a specific machine learning algorithm is able to improve as it is provided more data samples, and the degree up to which the captured relationship can be generalized. The The first thing is done by analyzing the curvature of the (cross-validated) score of the testing data-sets. As long as it increases, the machine learning model is still able to improve by providing it with more data-samples. Eventually, the accuracy score will plateau, indicating that maximum accuracy that a specific method can provide with the data that was given as input has been reached.The generalizability of a model is assessed by evaluating the accuracy score for the test-data, and by analyzing the gap between training and test-data accuracy, as well as the way in which this improves/worsens as sample size increases. Figure 10 shows the learning curve of the different kinds of machine learning algorithms that were used. The entire data-set (route, trip and origin-destination characteristics) were used as input variables for this purpose.

Figure 10: Learning curve for the different machine learning algorithms



There are a number of conclusions that can be drawn from this model. The Neural network, Naive Bayes, and Logistic Regression methods are unable to accurately fit the data at all. For both training and test data-sets they score low on accuracy. The accuracy for test-data also does not impressively improve as the number of training samples increases. The Decision Tree, Random Forest and Support Vector machine have an extremely high training accuracy (between 95% and 100%). However, the gap between training accuracy and cross-validation test-data accuracy is significant (especially for the Support Vector Machine method). This gap indicates over-fitting and/or high variance. The test-data accuracy does however increase significantly for the Decision Tree and Random Forest method as training sample is increased, decreasing the gap between test and training accuracy. This might indicate that the gap between training and testing accuracy could decrease even further if the model is provided more data samples. Also, even despite the potential of over-fitting, the cross-validated accuracy is still high, indicating that the model is nevertheless generalizable. The only method that has not been discussed yet -Gradient Boosting Method (GBM)- scores high on accuracy, and the gap between training and test data accuracy decreases as sample size increases. However, the test-data accuracy seems to stall after around 3000 samples, indicating that maximum accuracy has (almost) been reached for this method. In conclusion: for the machine learning algorithms that were assessed, a larger sample size might lead to more accurate estimations on test-data, but only slightly.

6 Discussion

This chapter contains the main conclusions of this research, and connects the described theory to the practical results. It also assesses some of the potentials in order to improve the models that were established in this research and/or to make them more generalizable. This chapter also serves to put the results of this research into the perspective of contemporary cycling literature. In other words: what new knowledge is developed in this research, and how can this knowledge aid in further innovation on the topic of cycling behavior literature and transport, health and sports literature in general.

6.1 Research results in perspective of objectives and contemporary cycling literature

This research has focused on the differences in cycling behavior (more specifically: route, trip and origin-destination behavioral dimensions) for different cycling motives, as well as the predictive power that these characteristics yield for estimating cycling motives. The section discusses in which ways the objectives of this research were met. Also, there are several key points for which the outcomes of this research add new knowledge on the topic of cycling behavior to the existing transport, health and sports literature. Those are also discussed.

New insights on cycling behavior for different cycling motives
There is currently little to no theory on the differences in cycling behavior for groups of cyclists with different purposes. This research provides insights into some quantitative aspects of cycling behavior. Based on the literature, it was to be expected that the cycling behavioral characteristics would be different for groups of cyclists with different cycling motives due to their needs, but the specific differences were not yet studied extensively. In order to meet the first objective of this research: '*to distinguish and identify profiles of cycling behavior for different cycling motives*' the characteristics of the three cycling behavioral dimensions that were established in the theory section were statistically tested for differences between groups based on cycling motives. Both raw cycling motives and adjusted cycling motives were assessed. The analysis of variance showed that there is a significant difference between all groups, but as shown in Section 4.5 the differences in cycling behavioral profiles are small for most characteristics. The main characteristics in which utilitarian and non-utilitarian trips differ are trip length (non-utilitarian trips being over twice as long on average), average traffic volume (which is significantly lower for non-utilitarian trips), and environmental value (which is significantly higher for non-utilitarian trips). There can be several explanations for these differences. It could be that non-utilitarian cyclists take detours to avoid dangerous/more crowded road segments (increasing travel time and decreasing traffic intensity). Another option is that non-utilitarian destinations are simply more scattered across space, increasing the average trip length. Destinations of trips -which also turned out to be the largest influencing factor in terms of estimation accuracy- are also quite different for both groups. Utilitarian trips are on average almost four times as likely to have business districts as their destination, whereas non-utilitarian trips are overrepresented in living areas and semi-agricultural areas. It is quite logical that infrastructural facilities should specifically be improved near business districts to facilitate utilitarian cycling (i.e. using a bicycle for commuting), but this study is among the first that provides quantitative justification for this claim. Since utilitarian trips most often originate from housing area's (56%),

bottlenecks in-between housing area's and business districts should be a top priority when trying to enhance the existing cycling infrastructure in order to better serve the needs of utilitarian cyclists. In regard of the raw cycling motives, that were also assessed, the behavioral profiles are in line with the results of the adjusted motives. Average traffic volume is high for paid work and services and lower than average for other motives. The differences between daily-shopping and non-daily shopping are almost negligible for all variables, which is interesting because earlier research by Feng and Timmermans (2014a) also indicated that these motives are hard to separate (even though they used entirely different data as input). It could mean that the cycling behavior of these two motives is roughly the same, or that differentiating variables do exist but are not included as data-input in either of the two studies.

New insights regarding suitability of route, trip and origin-destination behavioral dimensions as estimators of cycling motives

To meet the second goal of this research: '*to test the potential of cycling behavior for estimating cycling motives by means of machine learning*', the differences in cycling behavior profiles for different motives were used to estimate cycling motives by means of machine learning. As stated in Section 2.2, cycling motive estimation is under-researched compared to other aspects of travel behavior, and the ones that do assess cycling motive estimation are mainly focused on origin-destination data and/or socio-economic characteristics as input variables. This study adds to the existing work on motive estimation because it includes more behavioral dimensions like route and trip characteristics. Route characteristics turned out to be relatively poor estimators for the adjusted cycling motives when applying different kinds of machine learning algorithms. The use of the network data that was selected for this analysis only allowed for slight out-performance of the standard model (51,4% accuracy versus 58,4% accuracy (10x cross-validated)). It can therefore be concluded that, within this research setting, (average) route characteristics provide little predictive power for estimating cycling purposes. The trip characteristics turned out to poses more estimation power. The highest scoring machine learning algorithm (Random Forests) was able to estimate the cycling motive of a trip with an accuracy of 69,% (vs 51,4% standard model score), using trip characteristics as input data. Since average trip speed is exactly equal for both categories, trip length (which was over twice as high for non-utilitarian trips) is probably the main characteristic responsible for the out-performance of the standard model. The other input data that was prepared, origin-destination variables, was supposed to be a suitable variable for this estimating purpose according to the existing studies that have been discussed in Section 2.2. These assumptions turned out to be true in practice: the highest scoring models (Logistic Regression, Neural Network and Gradient Boosting all scored 69,3%) that were based on origin-destination data as input turned out to significantly outperform the standard model (69,3% accuracy versus 52,9% accuracy). Based on these analysises is can be concluded that origin-destination and trip characteristics are relevant groups of variables when trying to estimate cycling purposes. It can also be concluded that the BBG intersections for origin and destination are suitable proxies for these variables.

The combined model of all three cycling behavioral dimensions, including all available variables, turned out to poses the highest estimation power (74,0% for the Random Forests method versus 52,9% accuracy for the standard model). This relates to one of the main objectives of this research: *to test the potential of cycling behavior for estimating cycling motives by means of machine learning.*. One can conclude that the concept which states that 'cycling behavior is differentiated for groups of cyclists with different cycling motives' is theoretically sound, but -for the variables that were used as a proxy for 'route characteristics' in this research- does only translate into data with relatively small estimation power for cycling motives (compared to trip and origin-destination be-

havioral characteristics). Origin-destination and trip characteristics on the other hand have a rather large estimation power for cycling motives and also turned out to be helpful indicators in order to estimate cycling motives for GPS tracks for which the cycling motive is unknown. Compared to existing motive estimation research that has been discussed in Section 2.2, the accuracy scores (slightly) outperform most existing research, but it has to be mentioned that the adjusted motives that were estimated in this research might be easier to statistically distinguish than the raw motives.

In order to address the influence of each individual variable that was used as a behavioral characteristic, individual factor assessment was done by means of a leave-one-out-analysis. The cross-validated results of this analysis indicate that leaving the variable 'destination' out has the highest negative effect on the performance of the complete model. Leaving average max speed, average trip speed, and type of road out also decreases model accuracy but by smaller amounts. Origin, trip length and environmental values have little to no influence on the accuracy of the model, and average traffic volume is the only factor that actually turns out to *decrease* the accuracy of the model.

New insights regarding the suitability of (different kinds of) machine learning algorithms for the estimation of cycling motives
Finally, the performance of each machine learning algorithm was assessed in order to meet the last objective of this research: '*to identify which machine learning algorithm(s) is/are most suitable for estimating cycling motives from behavioral patterns*'. The work on motive estimation that has been carried out so far (discussed in Section 2.2) only focused on one method (either probabilistic, rule-based or by using one machine learning algorithm). This research adds to the current studies by not only addressing input variables (the behavioral cycling characteristics) but by also testing out different kinds of machine learning algorithms to assess the influence of each method on the accuracy of the estimations. In order to do so, literature regarding model performance and validity was assessed upfront, which resulted in a list of performance indicators, and model validity criteria. The performance was subsequently assessed by comparing accuracy, precision and recall of each model. The individual learning curves, model fits, bias-variance trade-offs were also discussed in order to derive conclusions regarding the validity of each model. In terms of model performance, ensemble methods like Random Forest and GBM turned out to score highest. Decision Tree also scored high on performance. With current sample sizes and variables, GBM turned out to yield the highest validity of these three methods. Therefore, within the context of this research, Gradient Boosting (GBM) and Random Forest are proposed as most suitable for estimating cycling motives. This can serve as important knowledge for future work on motive estimation or estimation of other types of behavioral transport characteristics.

All-together one can state that cycling behavioral profiles are significantly different among cycling motives and that these differences can be used to help estimate cycling motives. However, the outcomes of the analysis also led to new questions and remarks regarding the analysis results and the setup of this research, these are covered in the next sub-chapter.

6.2 Future work

Even though the outcomes indicate that some of the models that were assessed already hugely outperform the standard model, some enhancements could be made in order to make the approach that is used in this research even more accurate and/or generalizable. These enhancements could be taken into account for future research on motive-based cycling behavior and/or motive estimation. The most important ones are discussed below:

Hyperparameter optimization
During this research the machine learning algorithms were compared using their most default or most commonly used hyperparameters. Hyperparameter optimization could be done in order to increase the effectiveness of some machine learning algorithms for the specific estimation purposes of this research.

Increased sample size
As shown in Figure 10, the sample size does not seem to be the main concern for most machine learning algorithms that were assessed. However, for some (Random Forest method, Decision Tree) an increased sample size *could* lead to a higher test-data accuracy (see Section 5.6).

Combining the most effective machine learning algorithms by using voting classifiers
The two ensemble methods (which yield a voting system) that were tested in this research (Random Forest and GBM), turned out to be the the most effective methods for estimation purposes. A voting system can also be used for a group of machine learning classifiers. In this case, the classification outcome of each machine learning algorithm represents one 'vote'. Majority vote (referred to as hard voting) or an average estimated probability (referred to as soft voting) can be used to classify data entries based on the outcomes of the different methods. Using a combination of the best performing machine learning algorithms for the purpose of cycling motive estimation could result in a model which balances out the individual weaknesses of each method, resulting in a superior 'combined' model.

Better proxies for variables
The factors that were used during this research were justified by theory, but sometimes 'proxied' by other variables due to the absence of data. Origin and destination for example were derived from BBG land-use classes. If more precise data was available on the type of origin and the type of destination of a trip, the variables might be better representations of the factors 'origin' and 'destination'. In order to improve the quality of these factors, more variables could be added such as points of interest (POI) and/or additional land-use data from other sources than the BBG.

Using validated cycling motives instead of modeled cycling motives
The motives that were used for this research were themselves modeled rather than verified by follow-up research or surveys. This introduces a margin of error that is non-systematic and will therefore not influence the actual outcomes, but will only add additional noise to the resulting model (which rather decreases than increases prediction accuracy). The use of a data-set of which the cycling motives were validated by follow-up research like surveys of questionnaires will decrease this error margin, potentially increasing the prediction accuracy even further.

A more complete analysis framework
The complete model that was used for estimating purposes consists of 10 variables, which turned out

to be sufficient to significantly outperform a standard model, but is still a small amount compared to some other models used to estimate cycling motives that were discussed in Section 2.2.2. Factors like 'type of road' could provide more relevant information if they were represented by more than one variable (the data from the Fietsersbond network). This could lead to a more all-round representation of cycling behavior as a concept. Also, more factors could be added such as socioeconomic statistics, straightness of roads etc. Other options would be to include additional data-sources like points of interest and social geo-data.

7 References

Acker, V., Van Wee, B., and Witlox, F. When Transport Geography Meets Social Psychology: Toward a Conceptual Model of Travel Behaviour. *Transport Reviews*, 30(2):219–240, 2010.

Axhausen, K., SchöUnfelder S., Wolf, J., Oliveira, M., and Samaga, U. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. *Transportation Research Record*, 1870(August):46–54, 2004. ISSN 0361-1981. doi: 10.3141/1870-06.

Bao, J., Xu, C., Liu, P., and Wang, W. Exploring Bikesharing Travel Patterns and Trip Purposes Using Smart Card Data and Online Point of Interests. *Networks and Spatial Economics*, 17(4): 1231–1253, 2017. ISSN 15729427. doi: 10.1007/s11067-017-9366-x.

Bohte, W. and Maat, K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297, 2009. ISSN 0968090X. doi: 10.1016/j.trc.2008.11.004. URL `http://dx.doi.org/10.1016/j.trc.2008.11.004`.

Broach, J., Dill, J., and Gliebe, J. 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, 2012. ISSN 09658564. doi: 10.1016/j.tra.2012.07.005. URL `http://dx.doi.org/10.1016/j.tra.2012.07.005`.

Bunker, R. P. and Thabtah, F. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 2017. ISSN 22108327. doi: 10.1016/j.aci.2017.09.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S2210832717301485`.

Bussche, D. and van de Coevering, P. Bikeprint – in depth analysis of cycling behavior and cycle network performance using extensive GPS-track data. 2015.

Cauwenberg, J. V., Clarys, P., Bourdeaudhuij, I. D., and Ghekiere, A. Landscape and Urban Planning Environmental in fl uences on older adults ' transportation cycling experiences : A study using bike-along interviews. *Landscape and Urban Planning*, 169:37–46, 2018. URL `http://dx.doi.org/10.1016/j.landurbplan.2017.08.003`.

CBS. Bestand Bodemgebruik: Productbeschrijving. 2012. URL `https://www.cbs.nl/-/media/imported/onzediensten/methoden/classificaties/documents/2009/04/bestandbodemgebruikproductbeschrijving.pdf?la=nl-nl`.

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., and Das, K. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, 77:96–112, 2017. ISSN 0968090X. doi: 10.1016/j.trc.2017.01.020. URL `http://dx.doi.org/10.1016/j.trc.2017.01.020`.

Feng, T. and Timmermans, H. Enhanced Imputation of GPS Traces Forcing Full or Partial Consistency in Activity Travel Sequences. *Transportation Research Record: Journal of the Transportation Research Board*, 2430(2430):20–27, 2014a. ISSN 0361-1981. doi: 10.3141/2430-03. URL `http://trrjournalonline.trb.org/doi/10.3141/2430-03`.

Feng, T. and Timmermans, H. J. P. Using Recurrent Spatio-Temporal Profiles in GPS Panel Data for Enhancing Imputation of Activity Type. pages 1–12, 2014b.

Fernandez-Heredia, A. Cyclists' travel behavior, from theory to reality. (June):1–17, 2014.

Field, A. *Discovering statistics using SPSS*. Sage publications, 2009.

Fietsersbond. Metagegevens database fietsrouteplanner + fietsknooppunten + POI ' s Database Knopen (nodes) kenmerken en waarden. 2017. URL `http://docplayer.nl/7295641-Metagegevens-database-fietsrouteplanner-fietsknooppunten-poi-s.html`.

Fortmann-Roe, S. Understanding the bias-variance tradeoff, 2018a.

Fortmann-Roe, S. Accurately measuring model prediction error, 2018b.

Gaterslebena, B. and Appleton, K. Contemplating cycling to work: Attitudes and perceptions in different stages of change. *Transportation Research Part A: Policy and Practice*, 41(4):302–312, 2007. ISSN 09658564. doi: 10.1016/j.tra.2006.09.002.

Guo, C.-l., Li, C.-y., and Zhu, T. Travel Purpose Oriented Urban Cycling Network Planning. *Procedia - Social and Behavioral Sciences*, 96(Cictp):2443–2452, 2013. ISSN 18770428. doi: 10.1016/j.sbspro.2013.08.273. URL `http://linkinghub.elsevier.com/retrieve/pii/S1877042813023999`.

Harms, L., Bertolini, L., and te Brömmelstroet, M. Spatial and social variations in cycling patterns in a mature cycling country exploring differences and trends. *Journal of Transport and Health*, 1 (4):232–242, 2014.

Hasan, S. and Ukkusuri, S. V. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44:363–381, 2014. ISSN 0968090X. doi: 10.1016/j.trc.2014.04.003. URL `http://dx.doi.org/10.1016/j.trc.2014.04.003`.

Heinen, E. *Bicycle commuting*. 2011.

Hellmann, H. Opportunities for Dutch cycling en- terprises in Germany How can the Netherlands assist in promoting cycling in Germany ? 31(0):0–23, 2016.

Hull, A. and O'Holleran, C. Bicycle infrastructure: can good design encourage cycling? 2(1):369–406, 2014. ISSN 2165-0020. doi: 10.1080/21650020.2014.955210. URL `http://www.tandfonline.com/doi/abs/10.1080/21650020.2014.955210`.

Hunt, J. D. and Abraham, J. E. Influences on bicycle use. *Transportation*, 34(4):453–470, 2007. ISSN 00494488. doi: 10.1007/s11116-006-9109-1.

Khatri, R. Modeling Route Choice of Utilitarian Bikeshare Users from GPS Data. 2015.

KiM. Cycling and walking : the grease in our mobility chain. 2016.

Klinkenberg, J. and Bertolini, L. There are still opportunities for Dutch cycling. pages 1–9, 2012.

Mcneil, N. Working Paper Four Types of Cyclists ? 2005(701):1–20, 2012.

Meloni, I., Guala, L., and Loddo, A. Time allocation to discretionary in-home, out-of-home activities and to trips. *Transportation*, 31(1):69–96, 2004. ISSN 00494488. doi: 10.1023/B:PORT.0000007228.44861.ae.

Ministry of Transport. Cycling in the Netherlands. page 77, 2009.

Plazier, P. A., Weitkamp, G., and Berg, A. E. V. D. " 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(September):25–34, 2017. ISSN 0966-6923. doi: 10.1016/j.jtrangeo.2017.09.017. URL http://dx.doi.org/10.1016/j.jtrangeo.2017.09.017.

Refaeilzadeh, P., Tang, L., and Liu, H. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL https://doi.org/10.1007/978-0-387-39940-9{_}565.

Richardson, D., Volkow, N., Kwan, M., Kaplan, R., Goodchild, M., and Croyle, R. Spatial turn in health reesarch. *Science*, 339:1390–1392, 2013.

Rietveld, P. and Daniel, V. Determinants of bicycle use: Do municipal policies matter? *Transportation Research Part A: Policy and Practice*, 38(7):531–550, 2004. ISSN 09658564. doi: 10.1016/j.tra.2004.05.003.

Romanillos, G., Zaltz Austwick, M., Ettema, D., and De Kruijf, J. Big Data and Cycling. *Transport Reviews*, 36(1):114–133, 2016. ISSN 14645327. doi: 10.1080/01441647.2015.1084067.

Schönfelder, S., Axhausen, K. W., Antille, N., and Bierlaire, M. Exploring the potentials of automatically collected GPS data for travel behavior analysis: a Swedish data source. *GI-Technologien für Verkehr und Logistik, Institut für Geoinformatik, Universität Münster*, (13):155–179, 2002. doi: 10.3929/ethz-a-004403386.

Segadilha, A. B. P. and Sanches, S. d. P. Identification of Factors that Influence Cyclistś Route Choice. *Procedia - Social and Behavioral Sciences*, 160(Cit):372–380, 2014. ISSN 18770428. doi: 10.1016/j.sbspro.2014.12.149. URL http://linkinghub.elsevier.com/retrieve/pii/S1877042814062508.

SKLearn. 1.10. decision trees — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/tree.html, 2018a. (Accessed on 04/21/2018).

SKLearn. sklearn.neighbors.kneighborsclassifier — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html, 2018b. (Accessed on 04/21/2018).

SKLearn. sklearn.linear_model.logisticregression — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, 2018c. (Accessed on 04/21/2018).

SKLearn. sklearn.naive_bayes.gaussiannb — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html, 2018d. (Accessed on 04/21/2018).

SKLearn. 1.17. neural network models (supervised) — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/neural_networks_supervised.html, 2018e. (Accessed on 04/21/2018).

SKLearn. 3.2.4.3.1. sklearn.ensemble.randomforestclassifier — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, 2018f. (Accessed on 04/21/2018).

SKLearn. 1.4. support vector machines — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/svm.html, 2018g. (Accessed on 04/21/2018).

SKLearn. 1.11. ensemble methods — scikit-learn 0.19.1 documentation. http://scikit-learn.org/stable/modules/ensemble.html, 2018h. (Accessed on 04/21/2018).

Society, T. R. *Machine learning : the power and promise of computers that learn by example*. 2017a. ISBN 9781782522591. doi: 10.1126/scitranslmed.3002564. URL https://royalsociety.org/{~}/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

Society, T. R. *Machine learning : the power and promise of computers that learn by example*. 2017b. ISBN 9781782522591. doi: 10.1126/scitranslmed.3002564. URL https://royalsociety.org/{~}/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

Sun, Y. and Mobasheri, A. Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data. *International Journal of Environmental Research and Public Health*, 14(3), 2017. ISSN 16604601. doi: 10.3390/ijerph14030274.

Vandenbulcke, G., Claire, D., Thomas, I., de Geus, B., Degraeuwe, B., Meeusen, R., and Int Panis, L. Cycle commuting in belgium : Spatial determinants and 're-cycling' strategies. 2009.

Wardman, M., Tight, M., and Page, M. Factors influencing the propensity to cycle to work. *Transportation Research Part A: Policy and Practice*, 41(4):339–350, 2007. ISSN 09658564. doi: 10.1016/j.tra.2006.09.011.

Wegener, M. The future of mobility in cities: Challenges for urban modelling. *Transport Policy*, 29: 275–282, 2013. ISSN 0967070X. doi: 10.1016/j.tranpol.2012.07.004. URL http://dx.doi.org/10.1016/j.tranpol.2012.07.004.

Wikipedia. Travel behavior. https://en.wikipedia.org/wiki/Travel_behavior, 2018. (Accessed on 05/09/2018).

Winters, M., Teschke, K., Grant, M., Setton, E. M., and Brauer, M. How far out of the way will we travel? Built environment influences on route selection for bicycle and car travel. *Transportation Research Record*, (March):1–18, 2010. ISSN 9780309160643. doi: 10.3141/2190-01.

Xiao, G., Juan, Z., and Zhang, C. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, 71:447–463, 2016. ISSN 0968090X. doi: 10.1016/j.trc.2016.08.008. URL http://dx.doi.org/10.1016/j.trc.2016.08.008.